ZACHODNIOPOMORSKI UNIWERSYTET
TECHNOLOGICZNY W SZCZECINIE
Wydział Informatyki

ROZPRAWA DOKTORSKA

mgr inż. Artur Karczmarczyk

Heterogeniczne oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych

Promotor
dr hab. inż. Jarosław Jankowski, prof. ZUT

Promotor Pomocniczy
dr hab. inż. Jarosław Wątróbski, prof. US

Szczecin 2021

Na podstawie art. 187 ust. 3 Ustawy z dnia 20 lipca 2018 roku - Prawo o szkolnictwie wyższym i nauce (Dz. U. 2018 poz. 1668 z późn. zm.) przedkładam rozprawę doktorską w formie zbioru powiązanych tematycznie artykułów opublikowanych w czasopismach naukowych i recenzowanych materiałach konferencyjnych, które stanowią oryginalne rozwiązanie problemu naukowego. Tytuł prezentowanej rozprawy to: *„**Heterogeniczne oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych**”*. W skład rozprawy wchodzi cykl 10 publikacji naukowych z lat 2018-2021.

W dalszej części znajduje się syntetyczny opis uzyskanych wyników w postaci streszczenia rozprawy doktorskiej, a w szczególności omówienie:

— problemu badawczego,

— głównego celu rozprawy,

— cyklu publikacji,

— heterogenicznego oddziaływania na procesy rozprzestrzeniania informacji w sieciach społecznych,

— otwartego zorientowanego obiektowo środowiska symulacyjnego do badania procesu dyfuzji informacji w sieciach złożonych,

— dorobku akademickiego kandydata do stopnia doktora.

Następnie zostały zamieszczone pełne teksty opublikowanych artykułów naukowych wchodzących w skład rozprawy, jako załączniki numerowane od A1 do A10:

A1. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2018). Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one, 13(12), e0209372.

Liczba cytowań: 34
Impact Factor: 2.776
Liczba punktów ministerialnych: 100
Udział w artykule: 60%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A2. Wątróbski, J., Jankowski, J., Ziemba, P., **Karczmarczyk, A.**, Zioło, M. (2019). Generalised framework for multi-criteria method selection. Omega, 86, 107-124.

Liczba cytowań: 170
Impact Factor: 5.341
Liczba punktów ministerialnych: 140
Udział w artykule: 5%
Wkład: Opracowanie tekstu powiązane z analizą i porównaniem metod MCDA, udział w implementacji bazy reguł i systemu ekspertowego.

A3. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2019, September). Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 663-673). IEEE.

Liczba cytowań: 1
Indeksacja w WoS, Scopus
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A4. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2019). Parametrization of spreading processes within complex networks with the use of knowledge acquired from network samples. Procedia Computer Science, 159, 2279-2293.

Liczba cytowań: 1
Liczba punktów ministerialnych: 70
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A5. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2019). Multi-Criteria Approach to Planning of Information Spreading Processes Focused on Their Initialization With the Use of Sequential Seeding. In Information Technology for Management: Current Research and Future Directions (pp. 116-134). Springer, Cham.

Liczba cytowań: 1
Indeksacja w WoS, Scopus; rozdział w monografii
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A6. **Karczmarczyk, A.**, Bortko, K., Bartków, P., Pazura, P., Jankowski, J. (2018, August). Influencing information spreading processes in complex networks with probability spraying. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1038-1046). IEEE.

Liczba cytowań: 1
Liczba punktów ministerialnych: 15
Udział w artykule: 50%
Wkład: Opracowanie koncepcji i założeń, opracowanie algorytmów, przeprowadzenie badań, opracowanie tekstu.

A7. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeting Multi-Attribute Nodes in Complex Networks. Symmetry, 13(4), 731.

Impact Factor za rok 2020: 2.645
Liczba punktów ministerialnych: 70
Udział w artykule: 65%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A8. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeted Influence Maximization within Social Networks – in proceedings of International Conference on Computational Science: ICCS 2021

Publikacja zaakceptowana, w druku
Liczba punktów ministerialnych: 140
Udział w artykule: 65%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A9. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2021). Seeding for Complementary Campaign Objectives in Social Networks - in proceedings of The Americas Conference on Information Systems: AMCIS 2021

Publikacja zaakceptowana, w druku
Liczba punktów ministerialnych: 140
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A10. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). OONIS—Object-Oriented Network Infection Simulator. SoftwareX, 14, 100675.

Liczba punktów ministerialnych: 200
Udział w artykule: 80%
Wkład: Opracowanie koncepcji i założeń, projektowanie i implementacja, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

Suma punktów ministerialnych cyklu wynosi **875**, suma ważona cyklu wynosi **518**.

Sumaryczny Impact Factor: **19.663**, Impact Factor w cyklu: **10.762**.

Liczba cytowań wg WoS: **200**, bez cytowań własnych: **194**.

Liczba cytowań wg Scopus: **429** z **257** dokumentów. Według Google Scholar: **545**.

H-indeks wg WoS: **9**, H-indeks wg Scopus: **12**, H-indeks wg Google Scholar: **13**.

STRESZCZENIE ROZPRAWY DOKTORSKIEJ

mgr inż. Artur Karczmarczyk

Heterogeniczne oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych

Promotor
dr hab. inż. Jarosław Jankowski, prof. ZUT

Promotor Pomocniczy
dr hab. inż. Jarosław Wątróbski, prof. US

Spis treści

1. Problem badawczy

Reprezentacja sieci społecznych w systemach elektronicznych ewoluowała od wczesnych systemów technicznych do zaawansowanych mediów społecznościowych integrujących mechanizmami komunikacji i interakcji podobne do tych znanych z realnego świata. Rozwój platform społecznościowych wpłynał na potrzebę zrozumienia zachowań, wzorców i predyspozycji milionów użytkowników online i ich powiązania z zachowaniami w świecie rzeczywistym [1].

Procesy rozprzestrzeniania informacji, obok relacji społecznych i aktywności online, należą do zjawisk absorbujących uwagę zarówno badaczy, jak i praktyków. W wielu przypadkach komunikacja elektroniczna, oparta na bazie procesów rozprzestrzeniania informacji w sieciach społecznych, daje wyniki lepsze niż tradycyjne kampanie reklamowe [2]. W związku z tym, rośnie liczba firm wykorzystujących te mechanizmy by dotrzeć do potencjalnych odbiorców. Fakt silnych więzi pomiędzy znajomymi, komunikacja w sieciach społecznych charakteryzuje się zwiększoną wiarygodnością komunikacji. To sprawia, że rekomendacje zorientowane społecznościowo mają większy wpływ na docelowych odbiorców niż tradycyjny przekaz [3].

Badania związane z dyfuzją treści cyfrowych zorientowane są wokół czynników wpływających na sukces kampanii [4, 5], czynników wpywających na uczestnictwo użytkowników w procesie rozprzestrzeniania informacji [6], czy też wyboru użytkowników w sieci do inicjalizacji kampanii [7, 8]. Ponadto, badany jest wpływ roli różnych miar centralności podczas selekcji początkowych influencerów [9], role treści i struktur w sieciach [10], motywacja użytkowników do przekazywania treści [5], jak również rola emocji [11, 12] i innych czynników [13] w procesie rozprzestrzeniania informacji.

Wiele wcześniejszych badań koncentrowało się na podejściach teoretycznych i empirycznych do maksymalizacji zasięgu, czyli zwiększaniu liczby węzłów, do których udało się dotrzeć w sieci [14]. Badania te mogą opierać się na modelach analitycznych stosowanych w epidemiologii [15] lub bardziej skupiać się na strukturach i cechach sieci [16]. Inną możliwością jest wykorzystanie teorii i modeli związanych z dyfuzją innowacji [17].

Chociaż zasięg, czyli liczba zainfekowanych węzłów w sieci, jest ważną miarą sukcesu kampanii, z praktycznego punktu widzenia kampanie rozprzestrzeniania informacji w sieciach społecznych mogą mieć różne cele i specyfikę [18]. Inna strategia może być wykorzystana w celu pozyskania dużej liczby potencjalnych odbiorców w bardzo krótkim czasie niż w przypadku potrzeby osiągnięcia organicznego wzrostu bazy odbiorców o zadanych cechach. Budżet kampanii wpływa na liczbę i cechy demograficzne początkowo infekowanych węzłów zasiewowych (ang. seeds). Jakość początkowych węzłów i ich liczba mogą być kluczowym czynnikiem wpływającym na zasięg kampanii i jej ogólne wyniki. Jednakże dodatkowa alokacja budżetu może posłużyć do zwiększenia dynamiki lub czasu trwania kampanii. Z drugiej strony, decydentowi może zależeć na maksymalizacji zasięgu przy ograniczonych kosztach inicjalizacji procesu, jednak bez nacisku na szybkość dotarcia do

użytkowników w kolejnych iteracjach. Aby uwzględnić różne cele, można wykorzystać wielokryterialną ocenę procesu i dobrać parametry oraz cele zgodnie z preferencjami i priorytetami.

Do niedawna, większość badań nad procesami rozprzestrzeniania informacji w sieciach społecznych zakładało jednorodność wszystkich węzłów (użytkowników) w sieci. Oznacza to, że dotarcie do każdego użytkownika sieci było dla decydenta punktowane tak samo, jak dotarcie do dowolnego innego. Nieliczne spośród najnowszych badań zaczynają koncentrować się na kampaniach celowanych [19, 20]. W procesach tych, spośród wszystkich węzłów sieci wybiera się zbiór węzłów, do których inicjator procesu chce dotrzeć. W kampaniach takich celem jest maksymalizacja nie tyle globalnego zasięgu w sieci, co zasięgu w grupie docelowej. Pozwala to skoncentrować wysiłki na dotarciu do faktycznej grupy docelowej, a z drugiej strony na ograniczeniu niechcianej korespondencji – co pozwala uniknąć negatywnych efektów agresywnych kampanii.

Jak zostało to przedstawione powyżej, problem rozprzestrzeniania informacji w sieciach społecznych jest zagadnieniem złożonym, łączącym w sobie różnorodne założenia i wyzwania. Ta różnorodność stanowiła motywację do podjęcia badań nad heterogenicznym oddziaływaniem na procesy rozprzestrzeniania informacji w sieciach społecznych, których wyniki przedstawione są w tej rozprawie. Heterogeniczność w rozprawie oznacza oddziaływanie na proces rozprzestrzeniania informacji różnymi podejściami, a nie ograniczając się do pojedynczych miar i sposobów, co charakteryzowałoby podejście homogeniczne. Oddziaływanie na proces rozprzestrzeniania informacji natomiast oznacza sposoby wpływania na zasięg, dynamikę i inne charakterystyki procesu rozprzestrzeniania informacji, między innymi poprzez dobór węzłów zasiewowych (ang. seeds), nierównomierny rozrzut prawdopodobieństwa propagacji, czy zróżnicowane sekwencje inicjalizacji procesu. W przypadku kampanii rozprzestrzeniania informacji w sieciach społecznych, inicjator procesu może być zainteresowany nie tylko maksymalizacją zasięgu kampanii, lecz również oddziaływaniem na jej dynamikę czy też ograniczeniem wymaganego budżetu. W sekcji 4.2 zaprezentowano wielokryterialne podejście do planowania i ewaluacji procesów rozprzestrzeniania informacji w sieciach społecznych.

Badania nad procesami rozprzestrzeniania informacji w sieciach złożonych opierają się często na modelach sieci rzeczywistych, które mogą zostać pobrane z licznych repozytoriów sieciowych. Badania często opierają się na modelach teoretycznych do budowy sieci syntetycznych. Sieci takie są sparametryzowane, co pozwala skoncentrować wysiłki badawcze na konkretnych cechach sieci (zob. sekcja 4.2). Kolejnym problemem wykorzystania sieci rzeczywistych jest ich wielkość. Przeprowadzanie symulacji na modelach opartych na dużej liczbie węzłów jest czasochłonne i zasobochłonne. Ten problem zaadresowany został w sekcji 4.3.

Wiele dotychczasowych badań w zakresie rozprzestrzeniania informacji w sieciach społecznych opiera się na jak najlepszej inicjalizacji kampanii. Wysiłki w takich kampaniach koncentrują się wobec tego na znalezieniu grupy użytkowników w sieci, którym przekazanie informacji skutkować będzie osiągnięciem jak największego zasięgu informacji w sieci. Działania takie ograniczają się na ogół do wybrania użytkowników i uruchomienia kampanii poprzez pojedynczy wysiew informacji (ang. seeding). W sekcji 4.4 przedstawiono oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych poprzez zróżnicowane sekwencje inicjalizacji kampanii.

Poza wyborem inicjalnych węzłów do uruchomienia kampanii, w celu zwiększenia zasięgu kampanii mogą być prowadzone akcje pomocnicze, polegające na bezpośrednim oddziaływaniu na prawdopodobieństwo przekazywania informacji w sieci przy stałej liczbie węzłów inicjalnych. W sekcji 4.5 przedstawiono badania nad oddziaływaniem na procesy rozprzestrzeniania informacji w sieciach społecznych poprzez nierównomierny rozrzut prawdopodobieństwa propagacji informacji. Zbadano, jak zmieniać się będzie zasięg informacji w sieci wraz ze wzrostem prawdopodobieństwa propagacji węzłów o średnich lub niskich wartościach miar centralności (potencjalnie łatwiejszych do przekonania), zamiast celować w węzły o wysokich rankingach (potencjalnie trudniejszych i droższych do przekonania).

W sekcji 4.6 przedstawiono badania nad oddziaływaniem na procesy rozprzestrzeniania informacji w sieciach społecznych poprzez kierowanie kampanii w określone grupy użytkowników w sieci. W szczególności, skoncentrowano się na sieciach o węzłach opisanych wieloma atrybutami.

2. Główny cel rozprawy

Głównym celem prezentowanej rozprawy doktorskiej jest opracowanie i weryfikacja algorytmów heterogenicznego oddziaływania na procesy propagacji informacji w sieciach złożonych z udziałem złożonych rankingów dynamicznych, mechanizmów topologicznych, z uwzględnieniem wielokryterialnej oceny efektywności.

Teza rozprawy:

Oddziaływanie na procesy rozprzestrzeniania informacji w sieciach złożonych zróżnicowanymi metodami umożliwi zwiększenie zasięgu procesu i jego dynamiki oraz innych charakterystyk procesu zgodnych z preferencjami decydenta.

3. Cykl publikacji wchodzących w skład rozprawy

Jako osiągnięcie naukowe w dyscyplinie Informatyka techniczna i telekomunikacja wskazuję cykl dziesięciu powiązanych tematycznie publikacji pt. **Heterogeniczne oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych**. Cykl ten obejmuje cztery artykuły opublikowane w czasopismach z otwartym dostępem, pięć artykułów wydanych w recenzowanych materiałach konferencyjnych oraz jeden rozdział monografii. Powiązania pomiędzy publikacjami zostały przedstawione na rysunku 3.1.

W skład cyklu publikacji wchodzą następujące prace:[1]

A1. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2018). Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one, 13(12), e0209372.

Liczba cytowań: 34
Impact Factor: 2.776
Liczba punktów ministerialnych: 100
Udział w artykule: 60%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

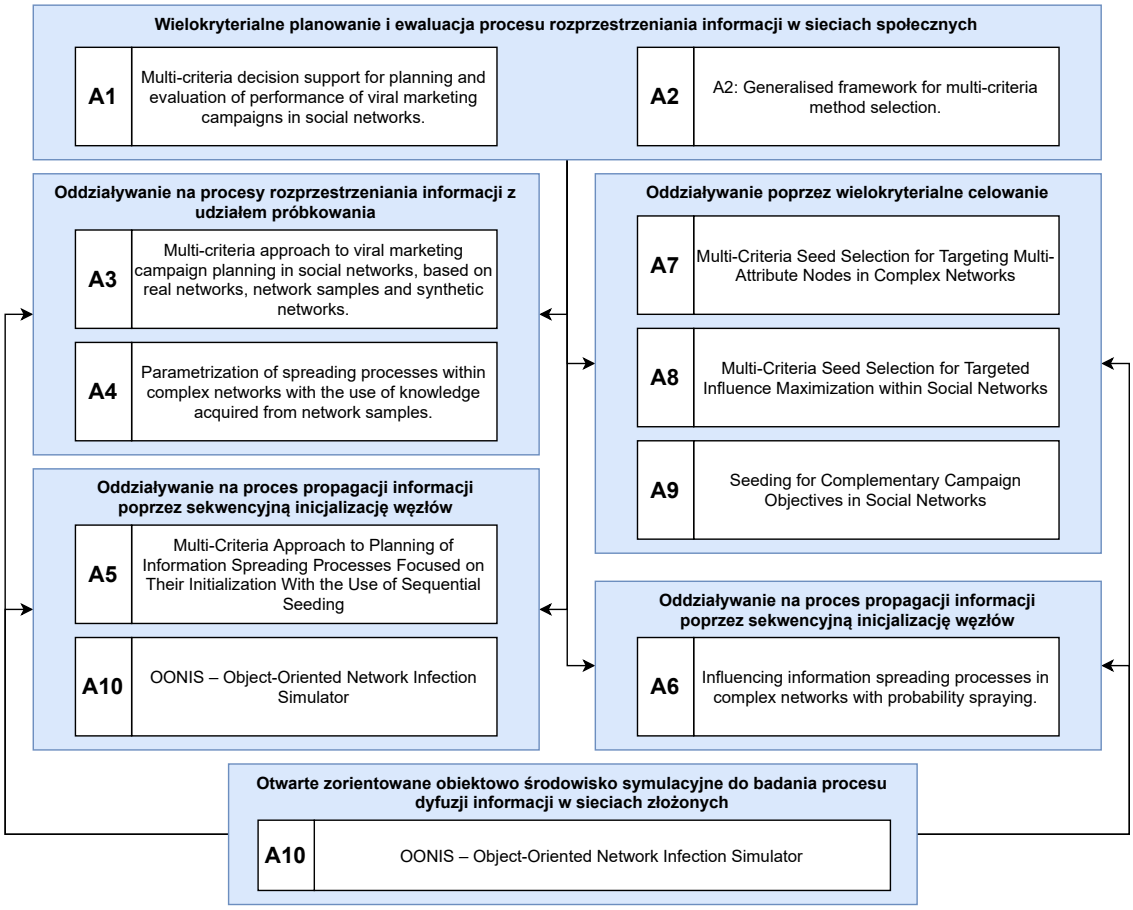
A2. Wątróbski, J., Jankowski, J., Ziemba, P., **Karczmarczyk, A.**, Zioło, M. (2019). Generalised framework for multi-criteria method selection. Omega, 86, 107-124.

Liczba cytowań: 170
Impact Factor: 5.341
Liczba punktów ministerialnych: 140
Udział w artykule: 5%
Wkład: Opracowanie tekstu powiązane z analizą i porównaniem metod MCDA, udział w implementacji bazy reguł i systemu ekspertowego.

A3. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2019, September). Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 663-673). IEEE.

Liczba cytowań: 1
Indeksacja w WoS, Scopus
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

[1]Informacje o cytowaniach pochodzą z Google Scholar, stan na dzień 4.05.2021 r.



Rysunek 3.1. Wizualizacja powiązań pomiędzy poszczególnymi publikacjami A1–A10

A4. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2019). Parametrization of spreading processes within complex networks with the use of knowledge acquired from network samples. Procedia Computer Science, 159, 2279-2293.

Liczba cytowań: 1
Liczba punktów ministerialnych: 70
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A5. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2019). Multi-Criteria Approach to Planning of Information Spreading Processes Focused on Their Initialization With the Use of Sequential Seeding. In Information Technology for Management: Current Research and Future Directions (pp. 116-134). Springer, Cham.

Liczba cytowań: 1
Indeksacja w WoS, Scopus; rozdział w monografii
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A6. **Karczmarczyk, A.**, Bortko, K., Bartków, P., Pazura, P., Jankowski, J. (2018,

August). Influencing information spreading processes in complex networks with probability spraying. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1038-1046). IEEE.

Liczba cytowań: 1
Liczba punktów ministerialnych: 15
Udział w artykule: 50%
Wkład: Opracowanie koncepcji i założeń, opracowanie algorytmów, przeprowadzenie badań, opracowanie tekstu.

A7. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeting Multi-Attribute Nodes in Complex Networks. Symmetry, 13(4), 731.

Impact Factor za rok 2020: 2.645
Liczba punktów ministerialnych: 70
Udział w artykule: 65%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A8. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeted Influence Maximization within Social Networks – in proceedings of International Conference on Computational Science: ICCS 2021

Publikacja zaakceptowana, w druku
Liczba punktów ministerialnych: 140
Udział w artykule: 65%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A9. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2021). Seeding for Complementary Campaign Objectives in Social Networks - in proceedings of The Americas Conference on Information Systems: AMCIS 2021

Publikacja zaakceptowana, w druku
Liczba punktów ministerialnych: 140
Udział w artykule: 70%
Wkład: Opracowanie koncepcji i założeń, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

A10. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). OONIS—Object-Oriented Network Infection Simulator. SoftwareX, 14, 100675.

Liczba punktów ministerialnych: 200
Udział w artykule: 80%
Wkład: Opracowanie koncepcji i założeń, projektowanie i implementacja, przeprowadzenie badań i opracowanie wyników, wizualizacja, opracowanie tekstu.

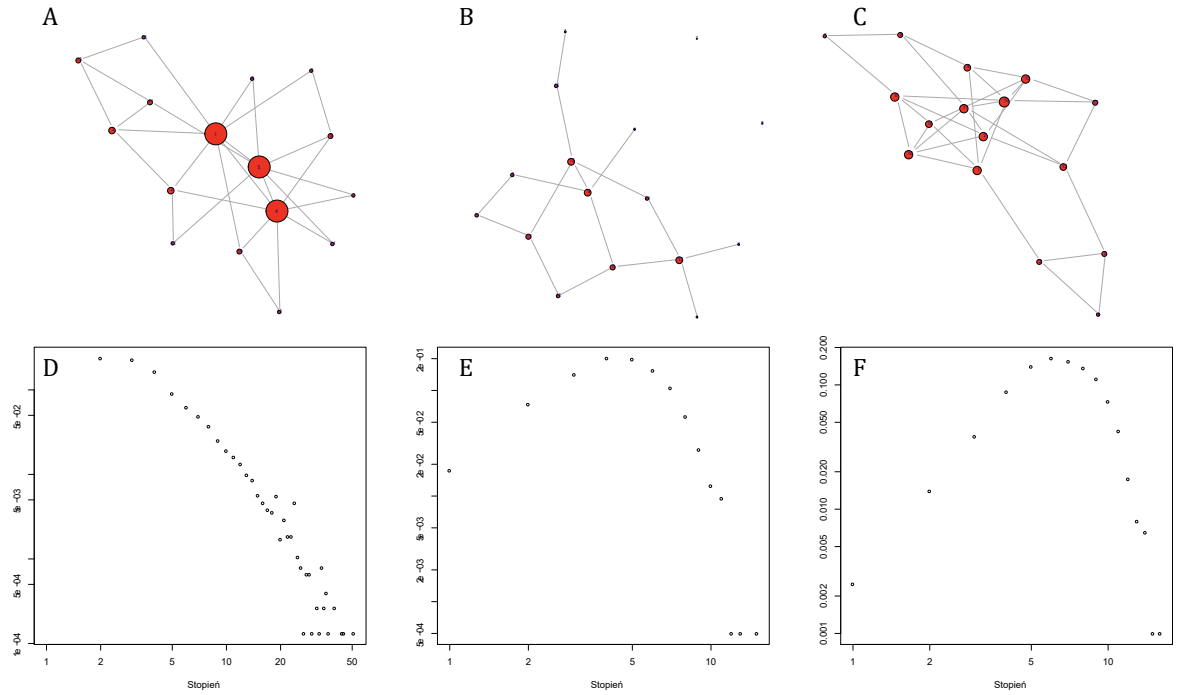
4. Heterogeniczne oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych

4.1. Podstawowe definicje

Działania prowadzone w sieciach społecznych mają często na celu motywowanie użytkowników do przekazywania informacji oraz treści cyfrowych znajomym oraz innym kontaktom w strukturach sieciowych. W związku z interdyscyplinarnością tego podejścia, badania prowadzone w tej dziedzinie angażują socjologów, fizyków, informatyków czy marketerów i obejmują szeroki wachlarz podejść i celów badawczych [3, 7]. Metodologiczne podstawy struktur sieciowych ewoluowały jednocześnie, ale oddzielnie, w różnych dyscyplinach [21].

Sieć społeczną G można zdefiniować jako zbiór węzłów (ang. nodes, vertices) $V(G)$ połączonych ze sobą za pośrednictwem zbioru krawędzi (ang. edges) $E(G)$. Sieć taką można opisać za pomocą notacji matematycznej: $G(V, E)$. Ścieżka (ang. path) w grafie G jest zbiorem krawędzi $\{\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_{n-1}, v_n\}\}$, gdzie koniec krawędzi $\{v_i, v_{i+1}\}$ stanowi początek krawędzi $\{v_{i+1}, v_{i+2}\}$ dla każdego $i = 0, \ldots, n-2$, gdzie każdy węzeł i krawędź są unikalne. Długość ścieżki to liczba krawędzi wchodzących w skład tej ścieżki. Odległość $d(i, j)$ od węzła i do węzła j to długość najkrótszej ścieżki od i do j.

Sposoby w jaki modelowane jest rozprzestrzenianie informacji w sieciach społecznych obejmują kilka podstawowych kierunków. Wyróżnić tu trzeba model kaskadowy (ang. independent cascade model [16]), model progowy (ang. linear threshold model [17]), jak również modele epidemiczne [15]. W rozprawie podczas badań oparto się na modelu kaskadowym, w którym informacja rozprzestrzenia się w sieci poprzez kaskady. Każdy węzeł w sieci znajduje się w jednym z dwóch stanów: aktywnym lub nieaktywnym. Początkowo wszystkie węzły są w stanie nieaktywnym. Na początku symulacji inicjalizuje się proces poprzez przekazanie informacji do małej części (tzw. ang. seeding fraction) węzłów (ang. seeds). W momencie otrzymania informacji, stają się one aktywne. W każdym dyskretnym kroku symulacji, węzły aktywne starają się przekazać informację do węzłów nieaktywnych. O sukcesie lub porażce przekazania informacji stanowi prawdopodobieństwo propagacji (ang. propagation probability) cechujące krawędź sieci łączącą węzeł aktywny z węzłem nieaktywnym. Każda krawędź ma swoją własną wartość prawdopodobieństwa propagacji. W całości sieci określa się średnią wartość prawdopodobieństwa propagacji. Każdy aktywny węzeł ma tylko jedną szansę przekazać informację do węzła nieaktywnego. Węzły do których uda się przekazać informację, zostają aktywowane. Powyższy proces trwa tak długo, jak aktywowane są kolejne węzły, a kończy się gdy żaden nowy węzeł nie zostanie aktywowany. Liczba węzłów aktywowanych w sieci po zakończeniu procesu kaskadowego to tzw. zasięg kampanii (ang. coverage).



Rysunek 4.1. Grafy przedstawiające przykład 16-węzłowych sieci syntetycznych: A) BA, B) ER, C) WS; oraz rozkład stopni węzłów przykładowych 2000-węzłowych sieci syntetycznych: D) BA, E) ER, F) WS.

Większość strategii inicjalizacji kampanii w sieciach opiera się na doborze początkowych węzłów (ang. seeds) na bazie rankingów uzyskanych na podstawie miar centralności węzłów. Zakłada się tutaj, że im węzeł bardziej centralny, tym większy zasięg w sieci pozwoli osiągnąć. Jako najbardziej podstawowe miary centralności można wskazać centralność stopnia (ang. degree centrality), centralność bliskości (ang. closeness centrality), centralność pośrednictwa (ang. betweenness centrality) i centralność wektora własnego (ang. eigenvector centrality). Ponadto, stosowane są bardziej zaawansowane rozwiązania, jak podejście zachłanne [16].

Niestety, wiedza o sieciach społecznych, w których prowadzone mają być kampanie, często ogranicza się do kilku podstawowych cech. Podczas gdy zbieranie informacji o rzeczywistych sieciach jest trudne, można wykorzystać syntetyczne sieci oparte na modelach teoretycznych. Dodatkowym atutem sieci syntetycznych jest możliwość dostosowywania ich struktury w trakcie procesu ich generowania, co pozwala na głębszą analizę procesów zachodzących w złożonych sieciach. W badaniach symulacyjnych często wykorzystuje się sieci oparte na modelu bezskalowym (ang. free-scale) zaproponowanym przez Barabasi-Alberta (BA) [22], modelu małego świata zaproponowanego przez Wattsa-Strogatza (WS) [23] oraz modelu grafu losowego wprowadzonego przez zespół Erdos-Renyi (ER) [24].

Charakterystyki modeli teoretycznych BA i WS są zbliżone do rzeczywistych systemów. Model Barabasiego-Alberta powstał w 1999 roku, w wyniku badania ówczesnej struktury sieci WWW. Budowa sieci BA opiera się na dwóch komplementarnych mechanizmach: rozwoju sieci oraz mechanizmie preferencyjnego dołączania. Model BA jest podobny do wielu systemów naturalnych i stworzonych przez człowieka, takich jak Internet, WWW,

sieci cytowań czy sieci społeczne. W systemach takich kilka wybranych węzłów (ang. hubs) ma wysoki stopień w porównaniu z pozostałymi węzłami sieci. Na rys. 4.1 (A) przedstawiono przykład sieci BA, a wykres na rys. 4.1 (D) przedstawia rozkład stopni węzłów przykładowego modelu.

Model sieci ER został po raz pierwszy opisany w 1959 r. Podczas jego konstrukcji, najpierw zdefiniowana jest liczba N węzłów, a następnie ze wszystkich $\binom{N}{2}$ par węzłów, wybiera się losowe E par, pomiędzy którymi tworzone są krawędzie. Przykładowy model ER i rozkład stopni przykładowego modelu ER przedstawiono odpowiednio na rys. 4.1 (B) i rys. 4.1 (E).

Model ER oferuje uniwersalny model o wielu zastosowaniach. Może on jednak być nieodpowiedni do modelowania niektórych zjawisk rzeczywistych, ponieważ nie generuje lokalnych klastrów węzłów. Aby rozwiązać ten problem, w 1998 roku powstał model Wattsa-Strogatza. Model WS uwzględnia klastrowanie, ale zachowuje krótkie średnie długości ścieżek z modelu ER. Rys. 4.1 (C) przedstawia przykład sieci WS, a wykres na rys. 4.1 (F) przedstawia rozkład stopni węzłów przykładowego modelu WS.

4.2. Wielokryterialne planowanie i ewaluacja procesów rozprzestrzeniania informacji w sieciach społecznych [A1, A2][1]

W przypadku procesów rozprzestrzeniania informacji w serwisach społecznościowych inicjator procesu może być zainteresowany nie tylko maksymalizacją zasięgu kampanii, ale także wpływaniem na jej dynamikę oraz utrzymaniem kosztów na zadanym poziomie. W związku z powyższym, planowanie takich kampanii jest problemem wielokryterialnym, który można przedstawić jako: (4.1) [25]:

$$max\left\{c_1(a), c_2(a), \ldots, c_k(a) | a \in A\right\}, \tag{4.1}$$

gdzie A oznacza zbiór możliwych strategii $\{a_1, a_2, \ldots, a_n\}$, a $\{c_1(\cdot), c_2(\cdot), \ldots, c_k(\cdot)\}$ oznacza zbiór kryteriów wykorzystywanych do ewaluacji tych strategii. Część kryteriów może być maksymalizowana, a część minimalizowana. Kryterialne wyniki każdej strategii w odniesieniu do każdego kryterium można wyrazić w postaci tabeli wyników. Intuicyjnie oczekuje się od decydenta określenia strategii optymalizującej wszystkie kryteria. Jednak zwykle nie ma alternatywy, która optymalizowałaby wszystkie kryteria jednocześnie.

Rozważmy przykładowy proces, dla którego przygotowano wiele alternatywnych strategii. Strategie te charakteryzują się trzema kryteriami: liczbą węzłów zasiewowych (ang. seeding fraction), prawdopodobieństwem propagacji informacji (ang. propagation probability) i potencjalnym zasięgiem, jaki można uzyskać (ang. coverage). W klasycznym homogenicznym podejściu maksymalizowałoby się zasięg. Zasięg jest bardzo ważnym kryterium, jednak generalnie strategia zapewniająca 100% zasięgu nie zawsze jest wybierana, ponieważ wymagałaby zainfekowania ogromnej liczby początkowych węzłów w

[1]Sekcja powstała na podstawie opublikowanego artykułu **A1** o współczynniku IF 2.776 i 33 cytowaniach oraz **A2** opublikowanego w renomowanym czasopiśmie Omega wydawnictwa Elsevier o współczynniku IF 5.341 i 167 cytowaniach.

sieci lub zapewnienia wielu bodźców w celu zwiększenia prawdopodobieństwa propagacji pomiędzy węzłami pośrednimi w sieci. Z drugiej strony, jeśli zostanie wybrana strategia z minimalną liczbą inicjalnych węzłów i minimalnym prawdopodobieństwem propagacji, nie można oczekiwać, że obejmie ona całą sieć. Dlatego należy wybrać spośród strategii rozwiązanie kompromisowe. W proponowanym heterogenicznym podejściu, oprócz maksymalizacji zasięgu, rozważa się także inne kryteria.

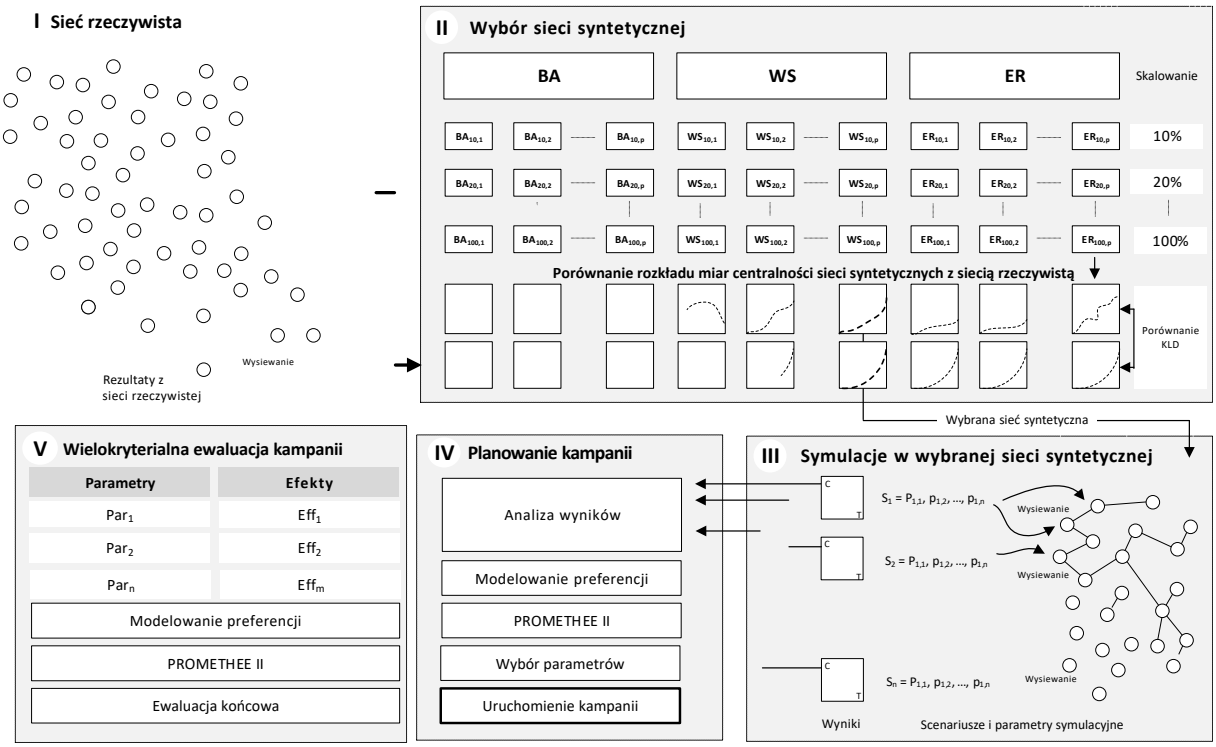
Należy zwrócić uwagę, że rozwiązanie problemu wielokryterialnego zależy nie tylko od wartości kryteriów każdej alternatywy, ale także od inicjatora procesu. Nie istnieje strategia absolutnie najlepsza dla wszystkich kampanii, natomiast najlepsza strategia kompromisowa zależy od preferencji decydenta.

Z przedstawionym wzorem (4.1) problemem decyzyjnym o charakterze wielokryterialnym można powiązać trzy relacje naturalnej dominacji: obojętność, preferencję i nieporównywalność. Rozważmy dwie alternatywy a i b. Jeśli dla każdego kryterium c_i, a jest tak samo dobre jak b, to obie strategie są obojętne (aIb). Jeśli dla każdego kryterium c_j, a jest tak samo dobre lub równe b i istnieje przynajmniej jedno kryterium c_k, dla którego a jest lepsze niż b, to a jest preferowane od b (aPb). Wreszcie, jeśli istnieje kryterium c_m, dla którego a jest lepsze niż b, ale istnieje również kryterium c_n, dla którego b jest lepsze niż a, to obie strategie są nieporównywalne (aRb).

Strategie, które są najlepsze według każdego kryterium mogą nie występować. W związku z tym, zazwyczaj większość strategii jest nieporównywalna bez dodatkowych informacji od decydenta. Informacje te mogą obejmować między innymi wagi wyrażające względne znaczenie każdego kryterium lub preferencje pomiędzy porównywanymi parami strategii, gdy każde kryterium jest rozpatrywane osobno [25]. Metody wielokryterialnego wspomagania decyzji (ang. Multi-Criteria Decision Analysis, MCDA) pomagają zredukować liczbę nieporównywalności (R) w grafie decyzyjnym między rozważanymi strategiami kampanii.

Metody MCDA generalnie dzieli się na dwie rodziny. Pierwsza, tzw. amerykańska, obejmuje metody agregujące wszystkie kryteria do pojedynczego kryterium – funkcji użyteczności. Druga, tzw. europejska, bazuje na relacjach przewyższania (ang. outranking) poprzez wzmacnianie relacji dominacji między alternatywami wszędzie tam, gdzie jest to możliwe. Przy takim podejściu nie wszystkie nieporównywalności są eliminowane, jednakże możliwy jest rzetelny wybór najlepszej alternatywy. Istnieją także metody łączące cechy obydwu rodzin. Szczegółowe zestawienie i ramy decyzyjne do wyboru metody MCDA dopasowanej do problemu decyzyjnego opisane zostały w publikacji **A2**. Na rysunku 4.2 przedstawiono schemat poglądowy proponowanego podejścia.

Na podstawie ram decyzyjnych opracowanych w **A2** i bazy reguł opublikowanej w [26] zdecydowano o wykorzystaniu metody PROMETHEE II w dalszej części tej sekcji. PROMETHEE to rodzina metod MCDA, które wykorzystują porównania alternatyw parami i przepływy przewyższania w celu stworzenia rankingu najlepszych wariantów decyzyjnych. Wagi wyrażające względne znaczenie każdego kryterium zostają określone przez decydenta. To skomplikowany proces oparty na priorytetach i spostrzeżeniach decydenta. Rzeczywiste wartości wag kryteriów mogą być dowolnie wybierane przez zamawiającego kampanię. Na szczęście metody wielokryterialnego wspomagania decyzji dostarczają narzędzi, takich jak analizy wrażliwości i odporności (ang. sensitivity and robustness analyses), które pozwalają zweryfikować wpływ wybranych wartości na otrzymane rankingi



Rysunek 4.2. Schemat poglądowy podejścia planowania i ewaluacji procesów rozprzestrzeniania informacji w sieciach społecznych.

i sekwencyjnie dostosowywać wybrane wagi. Szczegółowy opis metody PROMETHEE II można znaleźć w publikacji [25].

Wybór najlepszej strategii kampanii wirusowej w sieciach społecznych to złożony problem decyzyjny oparty na wielu kryteriach. Przeprowadzenie symulacji w rzeczywistej sieci jest najczęściej czasochłonne, a czasami niemożliwe. W związku z powyższym, w prezentowanym podejściu proponuje się prowadzenie procesu planowania na modelu syntetycznym, który ma podobne właściwości do docelowej sieci rzeczywistej, ale pozwala na przeprowadzenie wielu symulacji. Wynikiem tych symulacji jest zestaw danych zasilających tabelę wydajności (ang. performance table) wszystkich analizowanych strategii pod kątem kryteriów oceny. Tabela ta stanowi dane wejściowe do procesu oceny strategii.

W celu uzyskania modelu syntetycznego jak najbardziej przypominającego docelową sieć rzeczywistą, proponuje się wygenerować szereg sieci BA, ER i WS o zróżnicowanych parametrach oraz liczbie węzłów równą $10\%, 20\%, \ldots, 100\%$ sieci rzeczywistej. Następnie, wskaźnik KLD (ang. Kullback-Leibler Divergence, [27]) może zostać użyty do weryfikacji która z wygenerowanych sieci jest najbliższa sieci rzeczywistej. Dodatkowo rozważone mogą być kryteria wywierające wpływ na złożoność obliczeniową symulacji, takie jak liczba węzłów i krawędzi. W zależności od potrzeb decydenta do procesu decyzyjnego można również dodać dodatkowe kryteria. Gdy tabela wydajności dla wszystkich sieci syntetycznych i wszystkich kryteriów zostanie już stworzona, należy wybrać sieć najbardziej preferowaną z wykorzystaniem metod MCDA.

Kolejnym elementem proponowanego podejścia jest proces strukturyzacji modelu decyzyjnego. W trakcie tego procesu dobierane są kryteria decyzyjne oceny możliwych strategii kampanii. W proponowanym podejściu kryteria można podzielić na dwie grupy. Pierwsza

grupa zawiera kryteria wejściowe do budowy strategii – Par1, Par2, ..., Par$m$. Druga grupa zawiera kryteria oceny efektywności strategii Eff1, Eff2, ..., Eff$n$, których wartości opierają się na osiąganych efektach, a ich wartości można uzyskać z symulacji każdej strategii na wybranej sieci syntetycznej. Niemniej jednak proponowane podejście zakłada swobodę decydenta w doborze kryteriów decyzyjnych i grupowaniu ich w klastry w zależności od wymagań inicjatora procesu.

Gdy wszystkie kryteria są już dobrane, a model decyzyjny został ustrukturyzowany, na wybranej sieci syntetycznej przeprowadza się szereg symulacji dla każdej potencjalnej strategii. W podejściu prezentowanym w publikacji A1 zastosowano model IC (ang. independent cascade, [16]). Wybór został podyktowany względnie niewielką liczbą węzłów inicjalnych (ang. seeds) wymaganych do wzbudzenia propagacji informacji, co może być istotne w niewielkich sieciach. W przypadku modelu LT (ang. linear threshold) mała liczba węzłów początkowych nie przyniosłaby efektu.

W wyniku przeprowadzenia symulacji otrzymuje się macierz wydajności kryterialnej wszystkich ewaluowanych strategii. Macierz tę wykorzystuje się następnie w metodzie PROMETHEE II do przeprowadzenia wielokryterialnej oceny strategii. Analiza ta obejmuje w szczególności:

— wygenerowanie kompletnego rankingu strategii, w oparciu o różne funkcje preferencji;

— wykorzystanie płaszczyzny GAIA (ang. Geometrical Analysis for Interactive Assistance, [25]) w celu weryfikacji jak poszczególne kryteria oddziaływują na dobór strategii;

— przeprowadzenie analizy wrażliwości w celu weryfikacji stabilności uzyskanych rankingów dla wiodących strategii.

Należy zauważyć, że podczas analizy krok modelowania preferencji jest powtarzany wielokrotnie. Początkowe wagi preferencji kryteriów można następnie modyfikować w celu zweryfikowania odporności uzyskanego rozwiązania problemu wyboru strategii. Ostatecznie analityk rekomenduje, jaką strategię, czyli zbiór parametrów, zastosować do realizacji procesu w docelowej sieci rzeczywistej.

W celu weryfikacji proponowanego podejścia, zaproponowany został zestaw pięciu parametrów, z czego trzy pierwsze to kryteria dotyczące parametrów uruchomienia procesu, a pozostałe dwa to kryteria wydajnościowe:

Par1 Liczba węzłów (użytkowników) inicjalnych w kampanii (ang. seeds) wyrażona jako ułamek całkowitej liczby węzłow w sieci (ang. seeding fraction).

Par2 Motywacja do przekazywania treści - wyrażona jako średnie prawdopodobieństwo propagacji (ang. propagation probability).

Par3 Miara centralności wykorzystywana do wyboru początkowych użytkowników sieci w kampanii, takie jak centralność stopnia (ang. degree centrality), centralność wektora własnego (ang. eigenvector centrality) itp. Z doborem węzłów początkowych może wiązać się koszt. Przykładowo węzły o wysokiej wartości centralności wektora własnego uznawane są za użytkowników wpływowych w sieci społecznej, więc koszt ich pozyskania może być wyższy.

Eff4 Czas wymagany na dotarcie do założonej liczby użytkowników sieci. W przypadku modeli syntetycznych wartość ta reprezentowana jest przez liczbę kroków symulacji.

Eff5 Uzyskany zasięg w sieci, czyli liczba użytkowników sieci, do której dotarto przy założonej strategii kampanii względem całkowitej liczby użytkowników sieci.

Badania empiryczne proponowanego podejścia do planowania procesów rozprzestrzeniania zostały przedstawione w sekcji 3 publikacji **A1**. Badania te oparte zostały na sieci rzeczywistej [28] o 7610 węzłach i 15751 krawędziach, o średnim stopniu węzłów wynoszącym 4.14. W ramach badań empirycznych wygenerowano 150 sieci syntetycznych (po 50 każdego typu BA, ER i WS) zbudowanych z $10\%, 20\%, \ldots, 100\%$ węzłów w stosunku do sieci rzeczywistej. Z wykorzystaniem miary KLD oraz analizy wielokryterialnej wybrano do dalszej analizy sieć BA o 761 węzłach i 3034 krawędziach. Następnie przeprowadzono łącznie 4000 symulacji dla 400 zestawów kryteriów Par1-Par3. W efekcie uzyskano wartości Eff4 i Eff5 dla wszystkich 400 strategii. Uzyskane wartości posłużyły do przeprowadzenia analizy wielokryterialnej poszczególnych kampanii z wykorzystaniem metody PROMETHEE II z różnymi funkcjami preferencji. Analiza GAIA pozwoliła określić wpływ poszczególnych kryteriów na wybór najlepszej dla decydenta strategii, a analiza wrażliwości pozwoliła określić stabilność rankingu.

Należy zwrócić uwagę, że zaproponowane podejście może być również wykorzystane do monitorowania wyników przeprowadzonej kampanii, a także do przeprowadzenia wielokryterialnej oceny strategii kampanii w sieci rzeczywistej. Empiryczne badanie proponowanego podejścia do ewaluacji kampanii rozprzestrzeniania informacji w rzeczywistej sieci [28] zostało przedstawione w sekcji 4 publikacji **A1**.

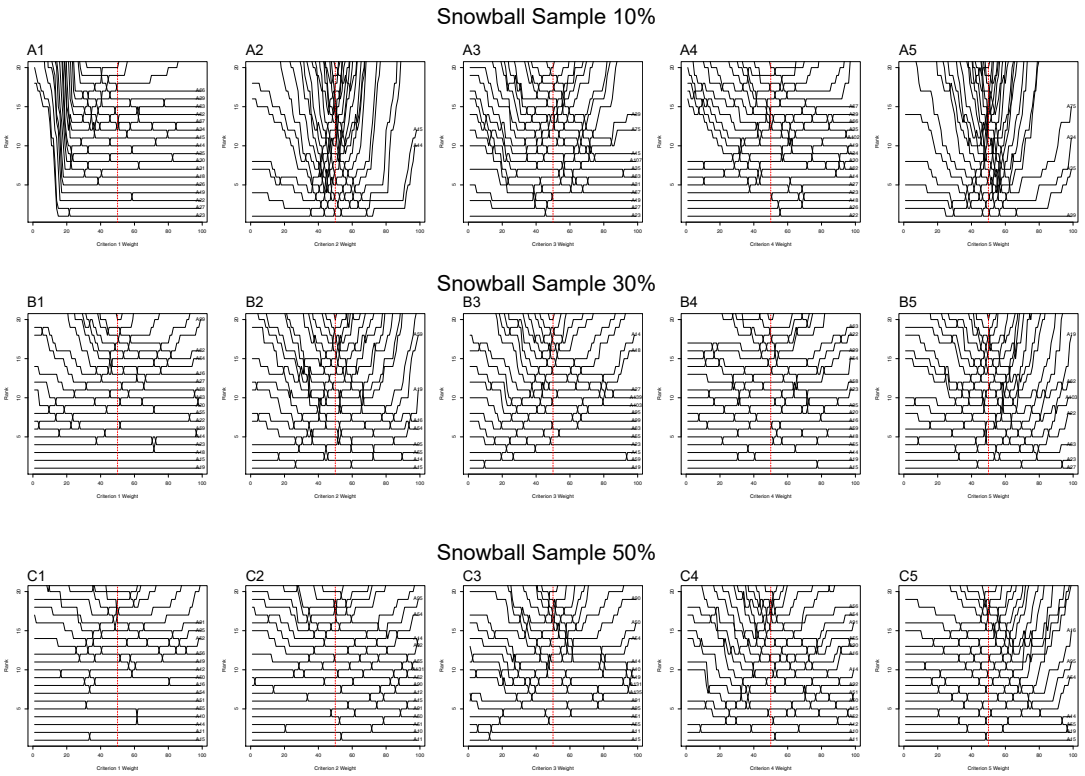
4.3. Oddziaływanie na inicjalizację procesów rozprzestrzeniania z udziałem próbkowania i model doboru wielkości próbek w ujęciu wielokryterialnym [A3, A4][2]

W sekcji 4.2 przedstawiono podejście do planowania procesów propagacji z wykorzystaniem sieci syntetycznych opartych na modelach teoretycznych. Zastosowanie sieci syntetycznych o odpowiednio zmniejszonej liczbie węzłów pozwoliło przeprowadzić zaawansowane planowanie kampanii przy zredukowanej złożoności obliczeniowej symulacji. W celu jak najlepszego dopasowania sieci syntetycznej do docelowej sieci rzeczywistej zastosowano miarę KLD oraz analizę wielokryterialną. Powstała jednakże obawa, że modele teoretyczne w niektórych przypadkach mogą nie być wystarczająco dobrze dopasowane do sieci rzeczywistych. W związku z powyższym, w tej sekcji przedstawiono podejście do planowania i uruchamiania kampanii rozprzestrzeniania informacji w sieciach społecznych z wykorzystaniem próbek sieci rzeczywistej.

W publikacji **A3** rozszerzono podejście proponowane w sekcji 4.2 o wykorzystanie próbek sieci rzeczywistej. Badanie empiryczne oparto na sieci rzeczywistej [29], przedstawiającej części topologii sieci Gnutella. Odwzorowana sieć składa się z 8846 węzłów i 31839 krawędzi. Węzły reprezentują hosty w topologii sieci Gnutella, a krawędzie reprezentują połączenia pomiędzy hostami Gnutella w jednej z migawek sieci zebranych w sierpniu 2002 r. Średni stopień węzłów w sieci rzeczywistej wynosi 7.1985.

W badaniu wykorzystano taki sam zestaw kryteriów ewaluacyjnych strategii Par1-Eff5 jak w sekcji 4.2. Dodatkowo, w tym przypadku do ewaluacji poszczególnych 400 stra-

[2]Sekcja powstała w oparciu o publikacje **A3** i **A4**.



Rysunek 4.3. Analiza wrażliwości rankingu dla 20 najlepszych strategii z ewaluacji metodą TOPSIS na próbkach sieci rzeczywistej [29]. A1-A5 – sieć 10%, B1-B5 – sieć 30%, C1-C5 – sieć 50%.

tegii zamiast PROMETHEE II wykorzystano metodę TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution, [30]). Metoda TOPSIS jest reprezentantem amerykańskiej szkoły MCDA, która przekształca wszystkie kryteria problemu decyzyjnego w pojedynczą wartość punktową. W przypadku metody TOPSIS, w oparciu o wartości ocenianych kryteriów, tworzone są idealne i antyidealne strategie, czyli takie, które są najlepsze pod względem każdego z kryteriów i takie, które są najgorsze pod względem każdego z kryteriów. Następnie każdej ocenianej strategii przypisuje się punkty obliczane jako relatywna odległość między ocenianą strategią a zarówno rozwiązaniem idealnym, jak i antyidealnym. Gdy wszystkim strategiom przypisane zostaną punkty, wybierana jest taka strategia, która jest najbliższa strategii idealnej, ale jednocześnie jest jej tak daleko jak to możliwe od strategii antyidealnej pod względem wartości poszczególnych kryteriów.

Jedną z zalet metody TOPSIS jest to, że pozwala ona na zbudowanie idealnego modelu referencyjnego dla zadanego problemu ewaluacyjnego. W przypadku wykorzystanej w badaniu sieci rzeczywistej, strategia idealna opierałaby się na miarze centralności stopnia (ang. degree centrality) do wyboru inicjalnych użytkowników sieci (ang. seeds). Dodatkowo, kampania uruchomiona zostałaby przez przekazanie informacji do 1% użytkowników. Ponadto, zachęty do propagowania treści zorganizowane byłyby w taki sposób, żeby osiągać średnie prawdopodobieństwo propagacji na poziomie 1%. Z drugiej strony, przy takich parametrach sieci, strategia idealna skutkowałaby wynikami pokrycia sieci na poziomie 97.22% przy średniej długości 19.6 iteracji. Należy zauważyć, że choć taka strategia byłaby idealna, to jednak jest tylko modelem referencyjnym i nie istnieje w rzeczywistości.

Tablica 4.1. Macierz korelacji pomiędzy rankingami strategii obliczonymi na sieci rzeczywistej i jej próbkach.

	Rzeczywista	**Próbka 10%**	**Próbka 30%**	**Próbka 50%**
Rzeczywista	x	0.4629	0.6837	0.9222
Próbka 10%	0.4629	x	0.8227	0.6159
Próbka 30%	0.6837	0.8227	x	0.8718
Próbka 50%	0.9222	0.6159	0.8718	x

W badaniu empirycznym planowanie procesu na sieci rzeczywistej w oparciu o jej próbki przeprowadzono z wykorzystaniem trzech próbek sieci różnych wielkości: 10%, 30% i 50% węzłów oryginalnej sieci. Próbki uzyskano z wykorzystaniem metody snowball sampling (zastosowano implementację *snowball.sampling* języka R z biblioteki *netdep* [31]). Następnie z wykorzystaniem metody TOPSIS dokonano ewaluacji wszystkich strategii na próbkach sieci rzeczywistej (zob. sekcja IV.D publikacji **A3**). Ponadto, wykonano analizę wrażliwości rankingów strategii. Wynik tej analizy (dla 20 najlepszych strategii) przedstawiony jest na rys. 4.3. Zwrócić należy uwagę, że o ile dla próbki 50% i 30% rankingi są dosyć stabilne, to przy próbce 10% niewielkie wahania wagi kryteriów Par2-Eff5 spowodowałyby znaczne zmiany kolejności strategii w rankingu.

W tabeli 4.1 przedstawiono wartości współczynników korelacji pozycji poszczególnych strategii w rankingach uzyskanych dla sieci rzeczywistej oraz poszczególnych jej próbek. Należy zwrócić uwagę, że o ile dla próbki 50% korelacja wynosi 0.9222, co wskazuje na wysoką korelację pomiędzy rankingami dla sieci rzeczywistej i jej 50-procentowej próbki, to dla mniejszych próbek wartość współczynnika korelacji nie jest zadowalająca.

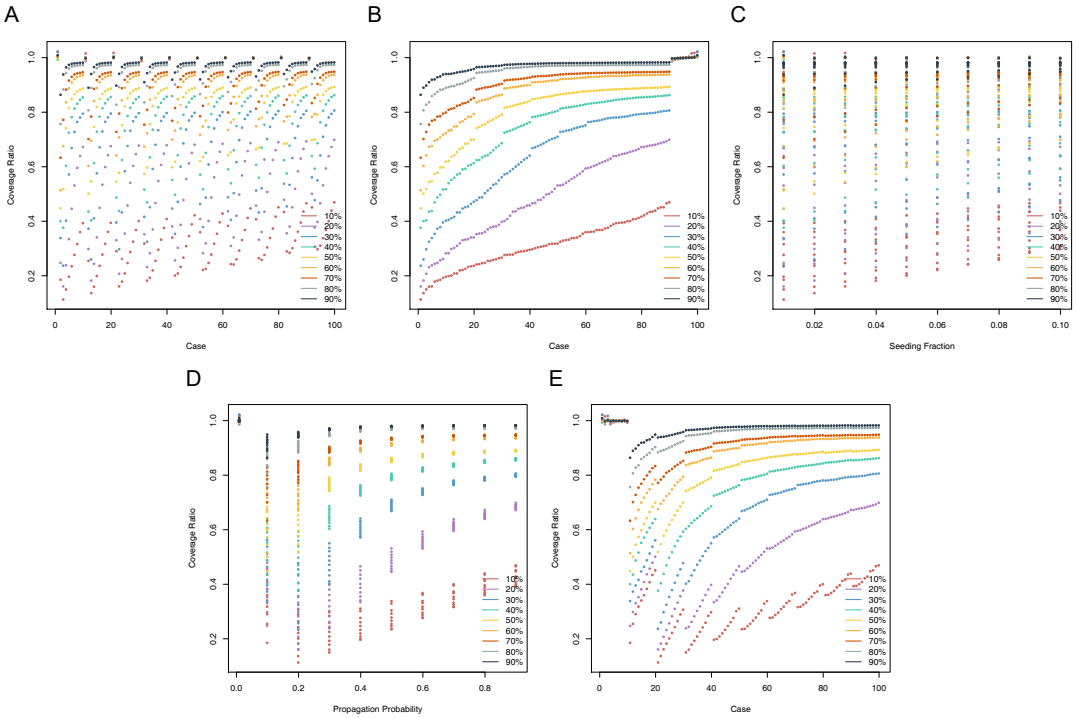
Wspomniana powyżej rozbieżność pomiędzy rankingami strategii dla sieci rzeczywistej i jej próbek różnych wielkości stanowiła podstawę opracowania publikacji **A4**. W publikacji tej przeprowadzono szereg symulacji dla sieci rzeczywistej Gnutella [29] oraz jej $10\%, 20\%, \ldots, 90\%$ próbek. Dla każdej z tych 10 sieci przeprowadzono 1000 symulacji dla:

— 10 wartości liczby inicjalnych węzłów kampanii (ang. seeding fraction, SF): 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10,

— 10 wartości średniego prawdopodobieństwa propagacji (ang. propagation probability, PP): 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10,

— 10 uprzednio wylosowanych scenariuszy ustawień wag w sieci. Dla każdego węzła została wylosowana wartość. Jeśli wartość była mniejsza lub równa wartości PP podczas symulacji, informacja była przekazywana przez określony węzeł. Jeśli wartość była wyższa od wartości PP, informacja nie była propagowana.

Inicjalne węzły (ang. seeds) do uruchomienia procesu wybierane były na podstawie rankingu opartego na centralności stopnia (ang. degree centrality) poszczególnych węzłów. Szczegółowe analizy znajdują się w sekcji 4 publikacji **A4**. Porównanie wartości zasięgu dla każdej próbki względem sieci rzeczywistej przedstawione zostało na rysunku 4.4. Jak można zaobserwować, wyniki najbardziej zbliżone do sieci rzeczywistej uzyskano dla próbek o jak najwyższej liczbie węzłów.



Rysunek 4.4. Stosunek zasięgu próbek do sieci rzeczywistej [29]. A: uporządkowane według przypadku symulacji, B: uporządkowane według stosunku zasięgu rosnąco, C: zgrupowane i uporządkowane według SF, D: zgrupowane i uporządkowane według PP, E: uporządkowane według rosnącej wartości zasięgu w sieci rzeczywistej.

W przypadku badanej sieci rzeczywistej, wybór 90% próbki sieci do planowania strategii rozprzestrzeniania informacji wydaje się najwłaściwszą opcją. Jednak w rzeczywistych zastosowaniach inicjator może zdecydować o rezygnacji z idealnej dokładności planowania kampanii, jeśli przyniosłoby to inne korzyści, które zrekompensowałyby utratę dokładności (kompensacja kryteriów). W publikacji **A4** zaproponowano wykorzystanie komponentu MCDA, aby ułatwić dobór wielkości próbki przy zmiennych preferencjach inicjatora procesu (Rys. 4.5).

Zaproponowano cztery kryteria do wyboru wielkości próbki sieci rzeczywistej, podzielone na dwie grupy – koszt i dokładność:

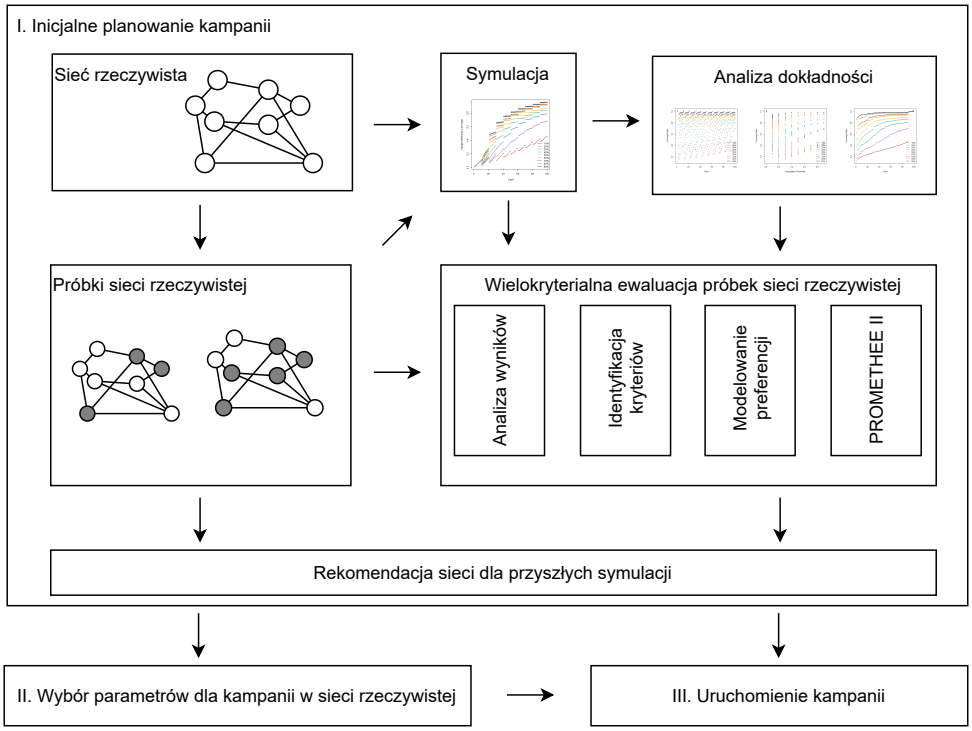
— C1 – grupa kosztów – wielkości próbki sieci, wyrażona jako stosunkowi liczby węzłów próbki do liczby wszystkich węzłów w sieci rzeczywistej;

— C2 – grupa kosztów – czas potrzebny na wygenerowanie próbki o określonej wielkości;

— C3 – grupa dokładności – różnica stosunku zasięgu pomiędzy próbką a siecią rzeczywistą od idealnego stosunku 1/1;

— C4 – grupa dokładności – różnica stosunku czasu trwania symulacji pomiędzy próbką a siecią rzeczywistą od idealnego stosunku 1/1.

Badanie empiryczne dla zaproponowanego modelu decyzyjnego przedstawiono w sekcji 4.3 publikacji **A4**. Na rysunku 4.6 przedstawiono analizę wizualną GAIA dla wyboru wielkości próbki sieci rzeczywistej przy równych wagach wszystkich czterech kryteriów. Można zwrócić uwagę, że dla zwykłej funkcji preferencji (ang. usual preference function) można



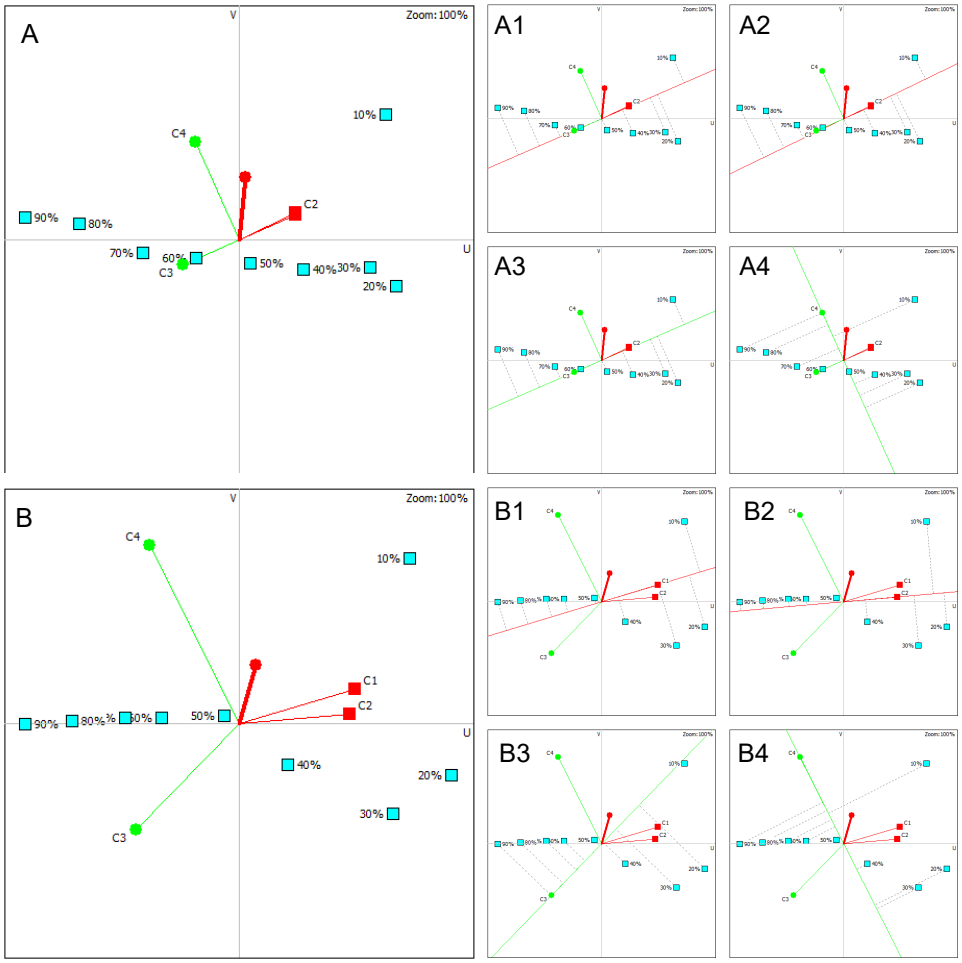
Rysunek 4.5. Schemat poglądowy proponowanego podejścia do wielokryterialnego doboru wielkości próbki przy zmiennych preferencjach inicjatora procesu.

wskazać zwycięską próbkę 10%, to wskazanie kolejności pozostałych wielkości próbek jest trudne – wiele z nich otrzymało taką samą ocenę w rankingu. Problem ten rozwiązany może zostać przez zastosowanie liniowej funkcji preferencji (ang. linear preference function) z uwzględnieniem wartości preferencji (ang. preference, aPb) i obojętności (ang. indifference, aIb) przy porównywaniu parami dowolnych dwóch wariantów. Na rysunku 4.6 (B) po zastosowaniu liniowej funkcji preferencji wyraźnie można wskazać, że drugim najlepszym wyborem byłoby użycie próbki 50%, a następnie kolejno 40%, 60% i 70%.

4.4. Oddziaływanie na proces propagacji informacji poprzez wielokryterialny dobór rankingów dla węzłów zasiewowych w podejściu sekwencyjnym [A5, A10][3]

Wiele dotychczasowych badań opiera się na jednoetapowym doborze węzłów początkowych w celu maksymalizacji zasięgu w sieci. Sprowadza się to na ogół do wyboru użytkowników początkowych sieci (ang. seeds), uruchomienia kampanii i ewaluacji wyników. W sekcji 4.2 zademonstrowano, że kampanie rozprzestrzeniania informacji w sieciach społecznych mogą być planowane z wykorzystaniem mniejszych modeli teoretycznych, co znacznie obniża wymagane moce obliczeniowe. Przeprowadzone badania pokazały, że mimo iż wykorzystane sieci syntetyczne były znacznie mniejsze i mniej skomplikowane obliczeniowo, współczynnik korelacji wyników na sieci syntetycznej i sieci rzeczywistej przekroczył wartość 0.9.

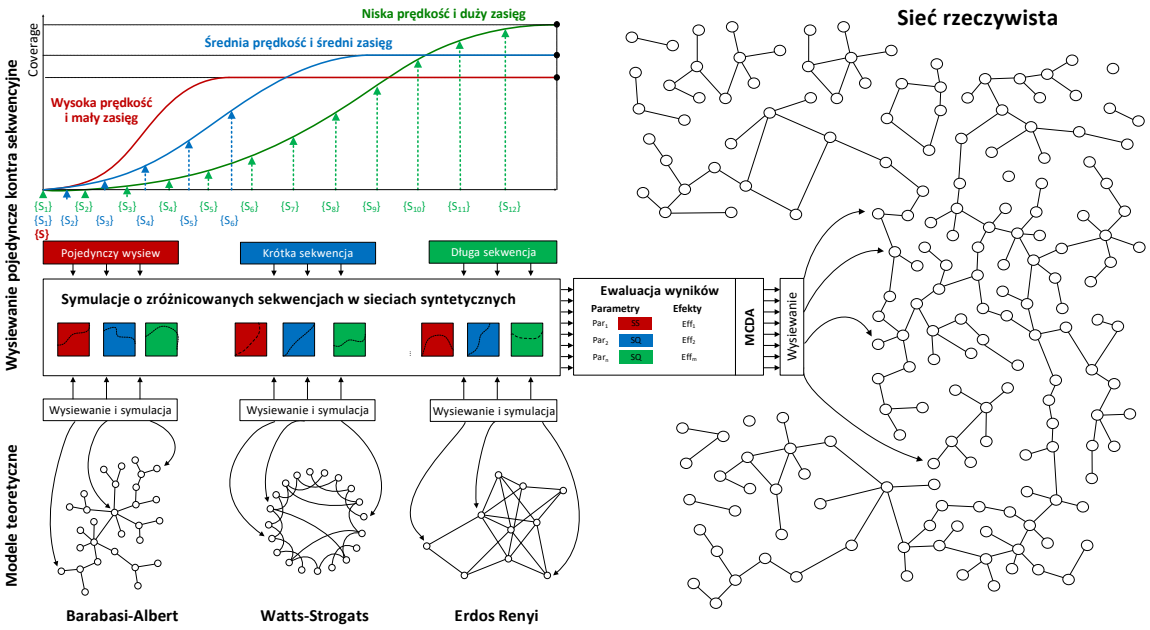
[3]Sekcja powstała w oparciu o publikację **A5** i wysoko-punktowaną publikację **A10** (200 punktów ministerialnych)



Rysunek 4.6. Analiza wizualna GAIA dla wyboru wiekości próbki sieci rzeczywistej. A - zwykła funkcja preferencji (ang. usual preference function). B - liniowa funkcja preferencji (ang. linear preference function).

Przegląd literatury pokazuje, że kryteria oceny strategii wykorzystane w sekcji 4.2 można dodatkowo rozszerzyć o sekwencyjną inicjalizację kampanii (ang. sequential seeding, [32]). Oznacza to, że strategia może przewidywać wystąpienie więcej niż jednej iteracji wprowadzania informacji do sieci (ang. seeding). Co więcej, takie iteracje mogą następować po sobie bez przerw lub być rozproszone po kampanii w określonych odstępach czasowych. To z kolei rodzi ciekawe pytanie badawcze, czy rozszerzenie modelu decyzyjnego doboru strategii kampanii rozprzestrzeniania informacji w sieciach społecznych o te nowe parametry pozwoliłoby na lepsze dopasowanie kampanii do potrzeb inicjatorów kampanii. W ramach prac badawczych początkowe podejście sekwencyjne z [32] rozszerzono poprzez zestawienie sieci syntetycznych ze zróżnicowanymi sekwencjami inicjalizacji kampanii w celu wielokryterialnego wyboru strategii rozprzestrzeniania informacji w sieci rzeczywistej.

W części metodycznej publikacji **A5** zaproponowano rozszerzenie wielokryterialnego podejścia z sekcji 4.2 o sekwencyjną inicjalizację kampanii (rys. 4.7). Założoną część węzłów początkowych (ang. seeding fraction) można wykorzystać do inicjalizacji kampanii na różne sposoby. Całość wyselekcjonowanych węzłów można zasilić informacją od razu na początku kampanii. Z drugiej strony, jednakże, można podzielić zbiór wyselekcjonowanych



Rysunek 4.7. Schemat poglądowy proponowanego podejścia do sekwencyjnej inicjalizacji węzłów z wykorzystaniem symulacji w sieciach syntetycznych.

węzłów na mniejsze podzbiory i aktywować je w kilku rzutach. Iteracje zasilania sieci w węzły początkowe mogą następować po sobie jedna po drugiej lub z przerwami.

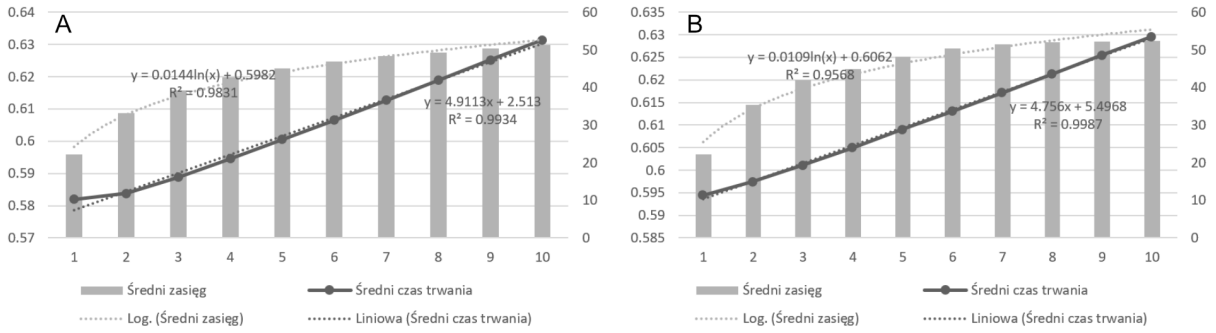
Uwzględniając propozycję sekwencyjnej inicjalizacji, zaproponowano nowy zestaw parametrów do planowania strategii kampanii rozprzestrzeniania informacji w sieciach społecznych. Uwzględniono pięć kryteriów dotyczących parametrów kampanii (Par1 - Par5) oraz dwa kryteria wydajnościowe (Eff6 - Eff7):

Par1 – odsetek węzłów inicjalnych (ang. seeding fraction)
Ułamek wszystkich węzłów, które zostały wybrane do pierwotnego dostarczenia informacji celem dalszego przekazywania w sieci.

Par2 – prawdopodobieństwo propagacji (ang. propagation probability)
Zakładane prawdopodobieństwo przekazania informacji z jednego zainfekowanego węzła do innych niezainfekowanych węzłów. Poziom prawdopodobieństwa propagacji można dostosować, stosując zachęty użytkowników do przekazywania informacji.

Par3 – liczba iteracji inicjalizacji procesu (ang. seeding iterations' count)
W pierwotnym modelu decyzyjnym z sekcji 4.2, do uruchomienia procesu propagacji informacji w sieci wykorzystywany był pojedynczy rzut informacji do wyselekcjonowanych węzłów inicjalnych. Bazując na początkowych sukcesach uzyskanych w badaniach z sekcji 4.2, postanowiono rozszerzyć oryginalny model o procedury sekwencyjnej inicjalizacji kampanii (ang. sequential seeding). Ten parametr określa, ilokrotnie informacje będą wprowadzone do sieci (uwzględniając zarówno początkowe, jak i kolejne rzuty informacji do sieci).

Par4 – interwał pomiędzy sekwencjami wprowadzania informacji do sieci
Wstępna inicjalizacja kampanii jest zawsze wykonywana w pierwszej iteracji. Jeśli istnieje więcej niż jedna iteracja zasilania sieci w informacje (patrz Par3), Par4 określa



Rysunek 4.8. Wpływ A) liczby iteracji, B) interwału pomiędzy iteracjami inicjalnego zasilania sieci informacjami na zasięg i czas trwania kampanii na przykładzie sieci rzeczywistej Gnutella [29]

jaki jest odstęp czasu pomiędzy każdą procedurą wprowadzania informacji do sieci wewnątrz pojedynczej kampanii.

Par5 – miara centralności wykorzystywana do budowy rankingu węzłów
Węzły wybrane do początkowego umieszczenia informacji nie są wybierane losowo. Najpierw są sortowane według wybranej metryki, a następnie wybierane są te najlepsze. Każda możliwa metryka charakteryzuje się określonym kosztem obliczeniowym.

Eff6 – liczba iteracji
Parametr ten reprezentuje moment w symulacji, w którym doszło do ostatniej infekcji, tj. kiedy wygasł proces propagacji informacji. Wartość tego parametru wynosi co najmniej $1 + (Par3 - 1) \times Par4$.

Eff7 – uzyskany zasięg w sieci
Jest to całkowity zasięg osiągnięty przez symulacje strategii opartej na parametrach Par1-Par5, tj. stosunek zainfekowanych węzłów do całkowitej liczby węzłów w sieci.

W badaniach empirycznych ponownie oparto się na sieci Gnutella [29]. Do planowania kampanii wykorzystano wstępnie 15 sieci syntetycznych o połowie wielkości sieci rzeczywistej. Za pomocą miary KLD [27] wybrano jedną z sieci BA do dalszych analiz. Następnie zademonstrowano działanie proponowanego podejścia dla dwóch przeciwnych celów kampanii. W pierwszym przypadku oczekiwano maksymalizacji zasięgu przy dużej dynamice kampanii skutkującej krótkim trwaniem kampanii. W drugim – maksymalizacji zasięgu w sieci, ale przy jak najdłuższym podtrzymywaniu trwania kampanii. Do ewaluacji wykorzystano metodę TOPSIS, co umożliwiło nie tylko łatwą ewaluację 9100 możliwych strategii, lecz również późniejsze przeprowadzenie analizy wrażliwości uzyskanych rozwiązań.

Prace empiryczne zakończono zbadaniem wpływu dwóch nowych kryteriów Par3 i Par4 na ostateczny zasięg i czas trwania procesu rozpowszechniania informacji. Dane z symulacji przeprowadzonych w sieci rzeczywistej Gnutella [29] zostały zagregowane i przedstawione na rysunku 4.8.

Wykres na A pokazuje, że wraz ze wzrostem liczby iteracji inicjalizacji kampanii (ang. seeding iterations) wzrastało zarówno średnie pokrycie, jak i czas trwania procesu. Średni wzrost czasu trwania symulacji można przybliżyć funkcją liniową $y = 4.9113x + 2.513$ z $R^2 = 0.9934$, natomiast średni wzrost zasięgu można przybliżyć funkcją logarytmiczną $y = 0.0144ln(x) + 0.5982$ z $R^2 = 0.9831$. Podobny wzrost średniego zasięgu i średniego

czasu trwania procesu rozprzestrzeniania informacji można zaobserwować, gdy interwał pomiędzy iteracjami inicjalizacji kampanii jest zwiększony (patrz wykres B). Wzrost czasu trwania można przybliżyć funkcją liniową $y = 4.756x + 5.4968$ z $R^2 = 0.9987$, podczas gdy średni wzrost zasięgu można przybliżyć funkcją logarytmiczną $y = 0.0109ln(x) + 0.6062$ with $R^2 = 0.9568$.

4.5. Oddziaływanie poprzez nierównomierny rozrzut prawdopodobieństwa propagacji informacji [A6][4]

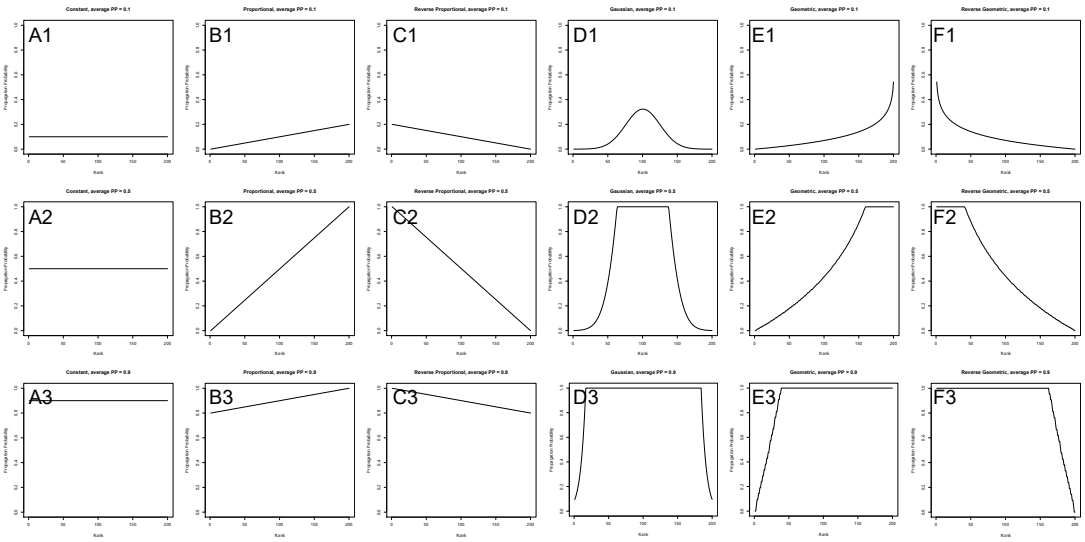
W tej sekcji przedstawione zostało wywieranie wpływu na dynamikę i zasięg procesu dyfuzji informacji poprzez nierównomierny rozrzut prawdopodobieństwa propagacji (ang. propagation probability spraying). W proponowanym podejściu, podobnie jak miało to miejsce we wcześniejszych sekcjach rozprawy, założono, że raz zainicjalizowane kampanie w sieciach społecznych mogą być następnie wspierane poprzez zwiększanie motywacji węzłów w sieci do przekazywania informacji do kolejnych węzłów. O ile we wcześniejszych badaniach zakładano, że prawdopodobieństwo propagacji w sieci rozrzucone jest losowo (lub jest jednolite), tak aby łącznie osiągnąć określony poziom średni (ang. average propagation probability), w publikacji **A6** zaproponowane zostało nowatorskie podejście.

W pracy zauważono, że na poszczególne węzły można oddziaływać np. zachętami w taki sposób, żeby wpływać na ich prawdopodobieństwo propagacji. Można rozważyć kilka strategii skupionych na zwiększeniu aktywności węzłów o wysokiej centralności. Jednak tacy użytkownicy (węzły) mogą być wymagający i trudno do nich dotrzeć. W związku z tym inną możliwością może być zwiększenie aktywności użytkowników o średnich lub nawet niewielkich wartościach miar centralności. Zwiększanie motywacji użytkowników o niskich wartościach miar centralności może być wydajniejsze pod względem kosztów, w związku z potencjalnie mniejszymi zachętami niż dla popularnych węzłów centralnych (ang. hubs).

Załóżmy, że prawdopodobieństwo rozpowszechnienia treści jest bezpośrednio związane z motywacją użytkownika. Z założonego zbioru i rozkładów D wybieramy rozkład D_i. Tworzony jest wektor $P_i[p_1, p_2, p_n]$ z n elementami, gdzie n odpowiada liczbie węzłów w sieci. Wektor P_i zawiera rozkład prawdopodobieństw, a każdy element reprezentuje prawdopodobieństwo, które należy przypisać do odpowiedniej pozycji w rankingu. Dla celów inicjalizacji kampanii w sieci (ang. seeding), w sieci z n węzłów tworzymy ranking węzłów reprezentowanych przez wektor $R_j[r_1, r_2, r_n]$ z węzłami uporządkowanymi według ich miary centralności typu j. Funkcja $f(p_i, r_i)$ przypisuje prawdopodobieństwo p_i elementowi o rankingu r_i. Funkcja $f(r_i, v_i)$ odwzorowuje prawdopodobieństwa przypisane węzłom w rankingu na poszczególne wierzchołki w sieci.

W trakcie badań nad proponowanym podejściem opracowano szereg algorytmów do uzyskiwania wektorów rozkładów prawdopodobieństw P_i dla zadanej liczby węzłow n i średniego prawdopodobieństwa propagacji P_{avg}. Przykładowe wyniki działania tych algorytmów dla różnorodnych rozkładów prawdopodobieństw przedstawiono na rysunku 4.9.

[4]Sekcja powstała na podstawie recenzowanej publikacji konferencyjnej **A6** indeksowanej w bazach WoS i Scopus

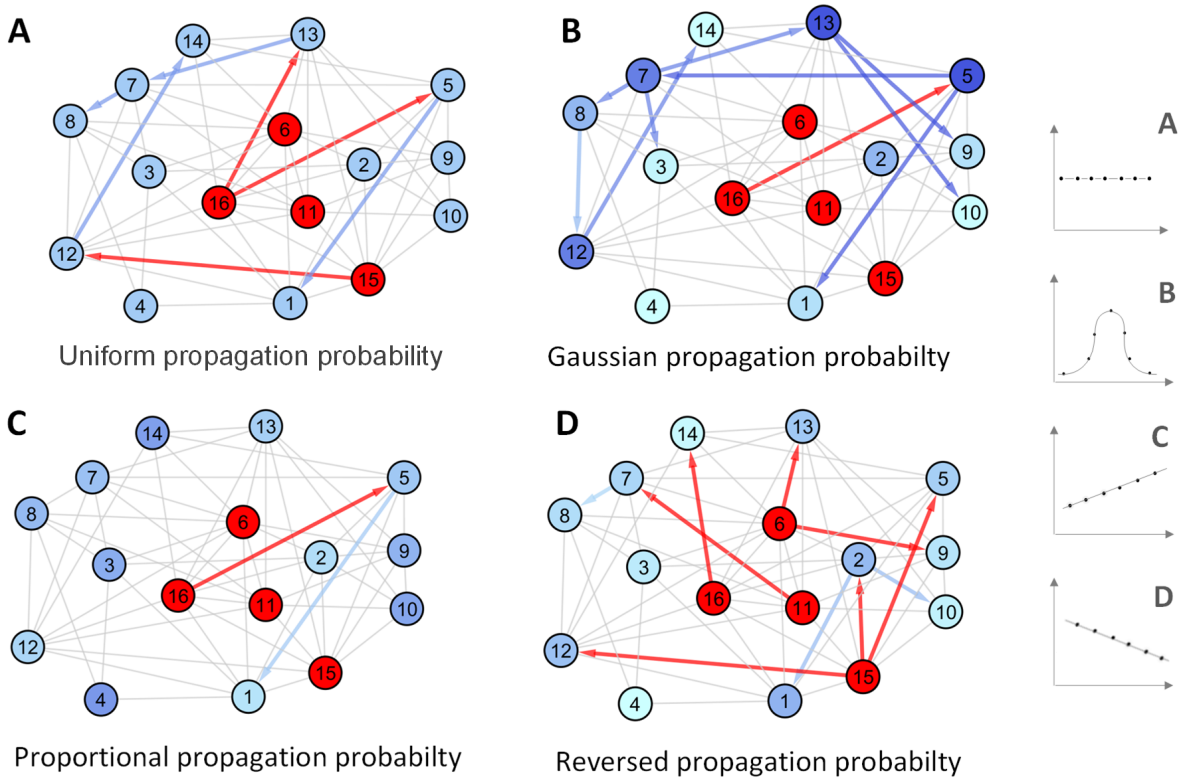


Rysunek 4.9. Przykładowe wyniki działania algorytmów generacji wektora P_i dla rozkładów **A** jednostajnego (ang. uniform), **B** proporcjonalnego (ang. proportional), **C** odwrotnie proporcjonalnego (ang. reversed proportional), **D** normalnego (ang. Gaussian), **E** geometrycznego (ang. geometric) i **F** odwróconego geometrycznego (ang. reversed geometric) przy założonym średnim prawdopodobieństwie propagacji równym **1** – 0.1, **2** – 0.5 i **3** – 0.9.

Tablica 4.2. Przedstawienie wartości prawdopodobieństwa dla wszystkich węzłów sieci [33] dla średniego prawdopodobieństwa propagacji wynoszącego 0.2

Ranking	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Węzeł	6	11	15	16	1	2	12	13	5	7	8	9	3	10	14	4
Stopień	10	9	9	9	8	8	8	8	7	7	7	7	6	5	5	3
r. jednorodny	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000
r. normalny	0.0002	0.0016	0.0092	0.0383	0.1195	0.2806	0.4957	0.6589	0.6589	0.4957	0.2806	0.1195	0.0383	0.0092	0.0016	0.0002
r. proporcjonalny	0.0000	0.0266	0.0533	0.0800	0.1066	0.1333	0.1600	0.1866	0.2133	0.2400	0.2666	0.2933	0.3200	0.3466	0.3733	0.4000
r. odwr. prop.	0.4000	0.3733	0.3466	0.3200	0.2933	0.2666	0.2400	0.2133	0.1866	0.1600	0.1333	0.1066	0.0800	0.0533	0.0266	0.0000

Efekty różnych strategii nierównomiernego rozrzutu prawdopodobieństwa zostały zilustrowane na prostym przykładzie na rysunku 4.10. Przedstawiony przykład jest oparty na symulacji w ramach sieci rzeczywistej składającej się z 16 węzłów [33]. W symulacjach założono, że inicjalnie zainfekowane zostanie 25% sieci, co przekładało się na 4 węzły (ang. seeds) o najwyższej centralności stopnia (ang. degree centrality). Na przykładzie przedstawiono cztery strategie rozrzutu prawdopodobieństwa propagacji (ang. propagation probability spraying), w oparciu o rozkłady jednorodny, normalny, proporcjonalny i odwrotnie proporcjonalny. Tabela 4.2 zawiera wartości prawdopodobieństwa propagacji (PP) dla każdego węzła w sieci, zgodnie z wybranymi dystrybucjami. Na rysunku 4.10 węzły wybrane do inicjalizacji kampanii oznaczono na czerwono. Odcień niebieskiego wskazuje na wielkość prawdopodobieństwa propagacji w odniesieniu do określonego rozkładu. Graf A przedstawia proces oparty na jednorodnym prawdopodobieństwie propagacji, w którym do każdego węzła przypisana jest ta sama wartość PP, równa 0.2. Proces kończy się z zasięgiem 68.75% z 11 zainfekowanymi węzłami w 4 krokach. Graf B ilustruje proces oparty na normalnym rozkładzie prawdopodobieństwa. Ostateczny zasięg 87.5% został osiągnięty w 6 krokach, przy 14 zainfekowanych węzłach. Dla rozkładu proporcjonalnego zilustrowanego na grafie C zasięg wynosi 37.5%, czyli 6 zainfekowanych węzłów w 3 krokach. Odwrócony rozkład proporcjonalny przedstawiony na grafie D skutkował zasięgiem 87.5%, czyli czternastoma zainfekowanymi węzłami w trzech krokach.



Rysunek 4.10. Przykładowy proces rozprzestrzeniania informacji z rozkładem **A** jednorodnym (ang. uniform); **B** normalnym (ang. Gaussian); **C** proporcjonalnym (ang. proportional); i **D** odwrotnie proporcjonalnym (ang. reversed proportional) prawdopodobieństwa propagacji informacji.

Powyższy przykład poglądowy pokazuje, jak oddziaływanie na procesy rozprzestrzeniania informacji w sieciach społecznych poprzez nierównomierny rozrzut prawdopodobieństwa propagacji informacji może wpływać na zasięg i czas trwania procesu. W kolejnym etapie badań przeprowadzone zostały symulacje na 25 sieciach syntetycznych i 5 sieciach rzeczywistych w celu oceny wpływu rozrzucanego rozkładu prawdopodobieństwa propagacji na końcowy zasięg kampanii. Opis badanych sieci oraz szczegółowe analizy przedstawione zostały w sekcji IV publikacji **A6**. Dwa z analizowanych podejść oparto na odwróconym rozkładzie geometrycznym i odwróconym rozkładzie proporcjonalnym, z większym wzrostem prawdopodobieństwa dla węzłów o wysokim rankingu. Skutkowały one największym wzrostem zasięgu w sieci w porównaniu z rozkładem jednorodnym prawdopodobieństwa. Zastosowanie geometrycznego i proporcjonalnego rozkładu prawdopodobieństw spowodowało natomiast zmniejszenie zasięgu w porównaniu z jednorodnym prawdopodobieństwem propagacji, ale ogólne koszty kampanii mogą być niższe. Podejście z rozkładem Gaussa okazało się skutkować najsłabszym zasięgiem, prawdopodobnie w związku z faktem, że znaczna część zachęt przeznaczona została w nim dla węzłów o przeciętnych rankingach.

4.6. Oddziaływanie poprzez wielokryterialne targetowanie [A7, A8, A9][5]

Większość dotychczasowych badań nad procesami rozprzestrzeniania informacji w sieciach społecznych zakłada jednorodność wszystkich węzłów (użytkowników) w sieci. Oznacza to koncentrację na dotarciu do jak największej grupy obojętnie których węzłów w sieci. Nieliczne spośród najnowszych badań zaczęły koncentrować się na kampaniach targetowanych [19, 20]. W odróżnieniu od tradycyjnych kampanii, spośród wszystkich użytkowników sieci wskazuje się tutaj podzbiór tych użytkowników, do których zamawiający kampanię chce dotrzeć. Miarą efektywności takiej kampanii targetowanej nie jest globalny zasięg w sieci, lecz zasięg osiągnięty bezpośrednio w grupie docelowej. Część badań koncentruje się również na unikaniu powtarzania wiadomości, w celu uniknięcia efektu habituacji [34], czy też przeładowania użytkowników informacjami.

Należy jednak zwrócić uwagę, że dotychczasowe badania w dziedzinie targetowanych kampanii wirusowych w sieciach społecznych bazowały homogenicznie na pojedynczym atrybucie lub mierze centralności w celu doboru węzłów początkowych do uruchomienia kampanii (ang. seeds). W praktyce jednak, rzeczywiste zastosowania sieci społecznych w wirusowych kampaniach marketingowych często opierają się heterogenicznie na wyborze wielu atrybutów użytkowników, takich jak wiek, płeć czy lokalizacja grupy docelowej.

W tej sekcji przedstawiono nowatorskie podejście do badania procesów rozprzestrzeniania informacji w sieciach społecznych. Proponowana metodologia uzupełnia szeroko stosowany model Independent Cascade (IC) [16] poprzez uwzględnienie problemu docierania do docelowych węzłów wieloatrybutowych w sieciach społecznych. W proponowanym podejściu zakłada się, że węzły sieci charakteryzują się nie tylko relacjami centralności pomiędzy nimi a innymi węzłami, ale także zestawem niestandardowych atrybutów $C_1, C_2, \ldots, C_n$.

Wartości tych atrybutów dla poszczególnych węzłów można wyrazić jako dokładne wartości liczbowe, takie jak wiek [lata] czy dochód [dolary]. Alternatywnie, jeśli atrybuty reprezentują jakościowe właściwości węzłów, ich wartości można przeliczyć na wartości liczbowe za pomocą 5-stopniowej skali Likerta [35, 36] (1 - zdecydowanie się nie zgadzam, 5 - zdecydowanie się zgadzam) lub wyliczenia (np. wiek: 1 - młody, 2 - w średnim wieku, 3 - stary; lub płeć: 1 - mężczyzna, 2 - kobieta).

Węzły można również scharakteryzować za pomocą miar centralności, takich jak centralność stopnia, bliskości, pośrednictwa czy wektora własnego. Dodatkowe atrybuty można również wyprowadzić jako złożenie dwóch wyżej wymienionych typów atrybutów. Na przykład, jeśli atrybut C_i reprezentuje stopień węzła, tj. całkowitą liczbę jego sąsiadów, C_{i_1} może reprezentować liczbę mężczyzn w jego sąsiedzwtie, a C_{i_2} liczbę kobiet w jego sąsiedztwie.

W proponowanym podejściu podjęto próbę dotarcia do węzłów o określonych wartościach wybranych atrybutów. Przykładowo, w programie profilaktyki raka piersi [37] podejmuje się próbę dotarcia do kobiet w średnim wieku, tj. między 50 a 69 rokiem życia.

[5]Sekcja powstała w oparciu o artykuł **A7** opublikowany w czasopiśmie Symmetry o współczynniku IF 2.645 oraz o dwie wysoko punktowane publikacje **A8** i **A9** (140 punktów ministerialnych każda).

W modelu IC [16] proces propagacji informacji w sieci poprzedzony jest selekcją inicjalnych węzłów do uruchomienia kampanii (ang. seeds). Zwykle nasiona reprezentują określoną część wszystkich węzłów sieci. Na przykład założone może być zainicjalizowanie kampanii poprzez przekazanie informacji do 5% użytkowników w sieci. Istnieje wiele sposobów wyboru inicjalnych węzłów sieci, które generalnie sprowadzają się do opracowania rankingu wszystkich węzłów w sieci i wyboru tych najwyżej punktowanych. Podczas gdy inne podejścia koncentrują się na generowaniu rankingu na podstawie pojedynczej miary centralności, takiej jak stopień, w podejściu proponowanym w tej sekcji rozważa się wiele atrybutów w celu wybrania węzłów inicjalnych.

Należy zauważyć, że w proponowanym podejściu ostateczny globalny zasięg sieci, czyli odsetek węzłów, do których dotarły informacje, może być niższy niż w przypadku tradycyjnych podejść opartych na miarach centralności. Jednak, co ukazały badania w publikacjach **A7**, **A8**, proponowane podejście daje szansę na zwiększenie zasięgu w grupach węzłów docelowych wewnątrz sieci.

W badaniach empirycznych w publikacjach **A7** i **A8** oparto się na sieciach rzeczywistych i syntetycznych. Ponieważ znalezienie modelu sieci o węzłach scharakteryzowanych wieloma atrybutami okazało się niemożliwe, postanowiono na zwykłe grafy, reprezentowane przez macierz sąsiedztwa, sztucznie nałożyć atrybuty na podstawie danych demograficznych [38] - płci (kobieta, mężczyzna) i grupy wiekowej (młodsi, średni, starsi). Następnie próbowano dotrzeć do różnych grup docelowych, przykładowo do kobiet w średnim wieku, bądź młodszych mężczyzn.

Dwa wyżej wspomniane atrybuty węzłów sieci, tj. płeć i wiek naturalnie zostały wybrane jako kryteria decyzyjne do procesu wyboru węzłów inicjalnych (ang. seeds) w kampanii. Kolejnym wybranym kryterium był stopień każdego węzła, jako najbardziej tradycyjna miara centralności do wyboru węzłów startowych. Ponadto, utworzono dodatkowe kryteria ewaluacyjne jako złożenie miary stopnia i pozostałych atrybutów sieci. Finalnie zaproponowano model decyzyjny oparty na 8 kryteriach:

C1 – stopień węzła, tj. liczba jego sąsiadów;

C2 – płeć użytkownika;

C3 – liczba sąsiadów płci męskiej;

C4 – liczba sąsiadów płci żeńskiej;

C5 – wiek użytkownika;

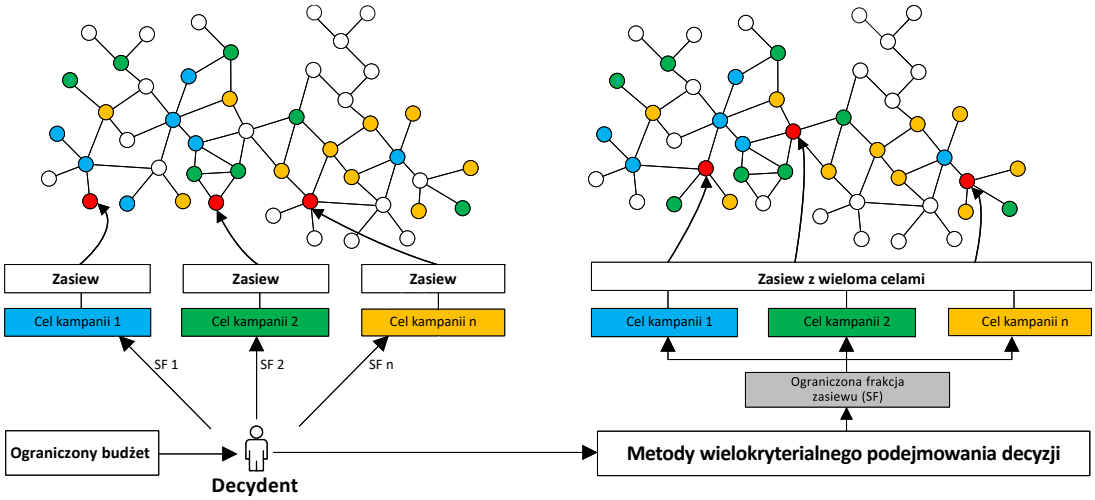
C6 – liczba sąsiadów w młodszym wieku;

C7 – liczba sąsiadów w średnim wieku;

C8 – liczba sąsiadów w starszym wieku.

W badaniach empirycznych w publikacjach **A7** i **A8** do wyboru węzłów inicjalnych zastosowano metody wielokryterialne TOPSIS i PROMETHEE (do ewaluacji wezłów) oraz AHP (do doboru wag kryteriów). Zastosowanie proponowanego podejścia umożliwiło uzyskanie w grupie docelowej zasięgu większego o 7.14% niż w przypadku tradycyjnego podejścia opartego o miarę centralności stopnia.

W związku z dobrymi osiągami proponowanego podejścia, zdecydowano o dalszym jego badaniu. Poskutkowało to jego rozszerzeniem w kierunku zrównoważenia (ang. susta-



Rysunek 4.11. Schemat poglądowy proponowanego zrównoważonego podejścia do realizacji pojedynczej kampanii realizującej n celów zamiast n kampanii realizujących pojedyncze cele.

inability) procesu doboru węzłów inicjalnych dla komplementarnych celów kampanii w sieciach społecznych (publikacja **A9**). W tak rozszerzonym podejściu zakłada się, że inicjator procesu rozprzestrzeniania informacji w sieci społecznej stara się osiągnąć n celów, tj. chce zmaksymalizować wpływ w n zbiorach docelowych węzłów (ang. targeted nodes). W tradycyjnym podejściu można byłoby przeprowadzić n kampanii, po jednej dla każdej grupy docelowej. Jednakże w proponowanym podejściu zakłada się ograniczony budżet, a tym samym ograniczony zbiór użytkowników początkowych sieci (ang. seeds), na bazie których można oprzeć kampanię. W związku z tym, proponuje się podejście zrównoważone, w którym decydent (ang. decision-maker, DM) stosuje metodę MCDA i na podstawie celów kampanii $1, 2, \ldots, n$, oceny eksperckiej oraz doświadczenia analityka wybiera ograniczoną część węzłów sieci do zainicjowania pojedynczej kampanii w celu dotarcia do węzłów odpowiadających wielu uzupełniającym się celom zamawiającego (zob. rys. 4.11).

Ze względu na zrównoważony charakter proponowanego podejścia zakłada się, że globalny zasięg sieci może zostać zmniejszony, lecz celem jest zwiększenie zasięgu wśród docelowych węzłów sieci. W badaniach empirycznych, przedstawionych w publikacji **A9**, dla przykładowych kampanii łączących dwa rozłączne cele udało się osiągnąć zasięg w grupach docelowych większy o 1.74% od tradycyjnego podejścia, przy zmniejszonym zasięgu globalnym o 2.1%.

5. Otwarte zorientowane obiektowo środowisko symulacyjne do badania procesu dyfuzji informacji w sieciach złożonych [A10][1]

Grafy i złożone sieci, a także zachodzące w nich procesy rozprzestrzeniania informacji to interdyscyplinarny temat badawczy, znajdujący zainteresowanie w takich dyscyplinach jak informatyka, fizyka, medycyna, epidemiologia [15, 7, 3]. Model kaskadowy (ang. independent cascade model, IC, [16]) oraz symulacje agentowe pozwalają na badanie procesów propagacji informacji w sieciach złożonych.

Istnieją biblioteki w języku R, takie jak igraph[2] oraz netdep[3], które umożliwiają reprezentację grafów i złożonych sieci w środowisku skryptowym języka R. Jednak aby przeprowadzić eksperymenty w modelu IC, badacze zmuszeni dotychczas byli do pisania własnych skryptów, zwłaszcza jeśli ich badania koncentrowały się na adaptacyjnym lub sekwencyjnym inicjalizowaniu kampanii (ang. adaptive and sequential seeding). Przemieszanie logiki modelu IC z właściwą logiką rozpowszechniania informacji wymaga od naukowców ponownego implementowania całości skryptu przy każdym kolejnym projekcie badawczym. Co więcej, podejście takie pomija osiągnięcia i zalety paradygmatu programowania obiektowego [39]. Stworzyło to interesującą lukę, którą w trakcie prac nad rozprawą udało się wypełnić zorientowaną obiektowo biblioteką wraz z wykorzystującym ją środowiskiem do symulacji, badania oraz oddziaływania na procesy rozprzestrzeniania informacji w sieciach złożonych – w tym sieciach społecznych.

Zastosowanie paradygmatów programowania obiektowego pozwoliło na enkapsulację warstwy złożonej i powtarzalnej logiki symulacji procesów rozprzestrzeniania informacji w sieciach do postaci łatwo reużywalnej biblioteki OONIS[4]. Ponadto zastosowane podejścia modułowego i podziału odpowiedzialności pozwoliło na zbudowanie środowiska, w którym tworzenie scenariuszy symulacyjnych jest łatwe, skalowalne i szybkie. Wdrożenie w projektach badawczych biblioteki OONIS pozwoliło podczas tworzenia tej rozprawy na tworzenie scenariuszy eksperymentalnych badających różne podejśćia, bez potrzeby utrzymywania warstwy logiki symulacyjnej, a w szczególności – mechaniki symulacji wysiewu informacji (ang. information seeding), infekowania węzłów i rejestrowania wyników. Przygotowana i przetestowana wielokrotnie podczas prac nad tą rozprawą biblioteka OONIS została następnie udostępniona do użytku publicznego na licencji GNU GPLv3 za sprawą publikacji o otwartym dostępie **A10**.

Zaprojektowane oprogramowanie składa się z dwóch części. Pierwsza, to zorientowana obiektowo biblioteka w R do symulacji procesów rozprzestrzeniania informacji w sieciach

[1] Sekcja powstała w oparciu o wysoko punktowaną publikację **A10** (200 punktów ministerialnych).
[2] https://igraph.org
[3] https://cran.r-project.org/web/packages/netdep/index.html
[4] OONIS – Object-Oriented Network Infection Simulator

```
# SEEDER
# choose and configure the simulation seeder component
seeder <- SomeDefaultOrCustomSeeder(...);

# CONTAMINATOR
# choose and configure the simulation contamination component
contaminator <- SomeDefaultOrCustomContaminator(...)

# RESULT PRINTER
resultPrinter <- SomeDefaultOrCustomResultPrinter(...);

# SIMULATION
maxIterations <- 100; minIterations <- 1;
ir <- InfectionRunner(infectionSeeder = seeder, contaminator = contaminator, resultPrinter = resultPrinter)
ir$readFromEdgesTxt('/path/to/network.txt', FALSE)
ir$run(maxIterations, minIterations)
```

Rysunek 5.1. Przykładowy kod źródłowy prostego scenariusza symulacji procesu rozprzestrzeniania informacji w sieci złożonej z wykorzystaniem proponowanego środowiska [40].

złożonych. Druga część to proponowane środowisko do przeprowadzania eksperymentów z wykorzystaniem owej biblioteki. Biblioteka składa się z głównej klasy *InfectionRunner*, która odpowiada za mechanikę symulacji w modelu IC [16]. Przed uruchomieniem symulacji, do obiektu klasy *InfectionRunner* muszą zostać podłączone instancje trzech modułów: *seeder* (do wprowadzania informacji do sieci), *contaminator* (do przekazywania informacji między węzłami) oraz *result printer* (do rejestrowania wyników).

W skład opublikowanej biblioteki wchodzi 7 przykładowych modułów zasiewania oraz trzy przykładowe moduły do obsługi przekazywania informacji. Domyślnie biblioteka umożliwia rejestowanie wyników do postaci plików rozdzielanych przecinkami.

Opublikowana biblioteka, wraz z bazującym na niej środowiskiem symulacyjnym, za sprawą enkapsulacji oraz podziału odpowiedzialności poszczególnych wymiennych modułów, umożliwia w łatwy sposób przeprowadzać eksperymenty heterogenicznych scenariuszy oddziaływania na procesy rozprzestrzeniania informacji w sieciach społecznych, takich jak:

— oddziaływanie przez zmienną frakcję inicjalnych węzłów w kampanii;

— oddziaływanie przez wysiewanie pojedyncze lub sekwencyjne w wielu iteracjach (publikacja **A5**);

— oddziaływanie przez nierównomierny rozrzut prawdopodobieństwa propagacji w sieci (publikacja **A6**);

— ewaluacja i planowanie kampanii marketingu wirusowego w sieciach społecznych w oparciu o sparametryzowane wartości frakcji wysiewu, prawdopodobieństwa propagacji, miar centralności węzłów i rankingi (publikacje **A1**, **A2**, **A3**, **A4**)

Na rysunku 5.1 zaprezentowano przykład wykorzystania proponowanego środowiska, ukazujący jak łatwe jest tworzenie rozmaitych scenariuszy symulacyjnych dla sieci złożonych z wykorzystaniem biblioteki i środowiska powstałych w trakcie tworzenia tej rozprawy. Wystarczy zainicjalizować obiekt InfectionRunner jednym z domyślnych lub samodzielnie oprogramowanych modułów wysiewu, zarażania i zbierania wyników, wczytać sieć za pomocą listy krawędzi i uruchomić symulację.

6. Podsumowanie

Procesy rozprzestrzeniania informacji w sieciach społecznych generują szereg wyzwań dla informatyki w obszarach modelowania, optymalizacji i efektywności obliczeniowej. Aktualne kierunki badawcze odnoszą się między innymi do rozkładu zasiewu w czasie, procesów adaptacyjnych, optymalizacji zasobów czy modelowania procesów konkurujących. Badania zostały posadowione w dyscyplinie *Informatyka techniczna i telekomunikacja*, a w szczególności zgodnie z klasyfikacją ACM Digital Library Computing Classification System w obszarach *Human-centered computing · Collaborative and social computing · Collaborative and social computing theory, concepts and paradigms · Social networks.*

W ramach realizowanych badań cel rozprawy osiągnięto poprzez wprowadzenie szeregu rozszerzeń do aktualnych rozwiązań. Dotychczas uwaga środowiska badawczego była zwrócona na problem maksymalizacji zasięgu poprzez odpowiedni dobór zbioru węzłów zasiewowych wykorzystywanych do jednorazowej inicjalizacji procesu. W rzeczywistych procesach poza samym zasięgiem istotne są też parametry czasowe, charakterystyka grup docelowych czy dostępne zasoby. W niniejszej rozprawie podjęto próbę zbliżenia modeli teoretycznych do potrzeb systemów rzeczywistych i zaproponowano ramy do wielokryterialnego planowania i ewaluacji procesów rozprzestrzeniania informacji w sieciach społecznych. Następnie przedstawiono szereg sposobów heterogenicznego oddziaływania na procesy rozprzestrzeniania informacji w nich zachodzące. W rozprawie oddziaływano na procesy rozprzestrzeniania informacji poprzez sekwencyjność inicjalizacji, nierównomierny rozrzut prawdopodobieństwa propagacji informacji oraz targetowanie wielokryterialne.

W szczególności osiągnięcia uzyskane w rozprawie, które stanowią wkład do *Informatyki technicznej i telekomunikacji* obejmują:

— opracowanie autorskiego podejścia do wielokryterialnego planowania i ewaluacji procesów rozprzestrzeniania informacji w sieciach społecznych;

— zaprezentowanie podejścia wykorzystującego sieci syntetyczne i próbki sieci rzeczywistych do planowania procesów rozprzestrzeniania informacji w strukturach sieciowych;

— opracowanie wielokryterialnego podejścia do wyboru rozmiaru próbek sieci rzeczywistej w celu uproszczonego obliczeniowo pozyskiwania przyszłych rekomendacji odnośnie parametrów inicjalizacji procesów w sieci rzeczywistej;

— opracowanie podejścia wykorzystującego sieci syntetyczne i zróżnicowanie sekwencji inicjalizacji węzłów wysiewowych w celu doboru strategii do inicjalizacji i realizacji procesu w sieci rzeczywistej;

— zbadanie oddziaływania nierównomiernego rozrzutu prawdopodobieństwa propagacji informacji w sieciach na osiągany zasięg i dynamikę procesu rozprzestrzeniania informacji;

— opracowanie szeregu algorytmów do generowania wektorów nierównomiernego rozrzutu prawdopodobieństwa propagacji informacji (ang. probability spraying);

— zaproponowanie podejścia do wielokryterialnego targetowania w wieloatrybutowe (heterogeniczne) węzły w sieciach społecznych;

— zaproponowanie zrównoważonego podejścia do adresowania wielu celów działań w sieciach o wieloatrybutowych węzłach z wykorzystaniem pojedynczych kampanii;

— opracowanie otwartego zorientowanego obiektowo środowiska symulacyjnego do badania procesu dyfuzji informacji w sieciach złożonych.

Dodatkowym efektem rozprawy doktorskiej w dziedzinie Informatyka techniczna i telekomunikacja było opublikowanie otwartej zorientowanej obiektowo biblioteki oraz środowiska w języku R do przeprowadzania rozmaitych eksperymentów symulacyjnych z zakresu oddziaływania na procesy rozprzestrzeniania informacji w sieciach złożonych.

Przeprowadzone badania pozwoliły na potwierdzenie postawionej w rozprawie tezy. Oddziaływanie na procesy rozprzestrzeniania informacji w sieciach złożonych zróżnicowanymi metodami umożliwiło zwiększenie zasięgu procesu i jego dynamiki oraz innych charakterystyk zgodnie z preferencjami decydenta, które nie znajdwały odzwierciedlenia we wcześniejszych rozwiązaniach.

Prace prowadzone w ramach rozprawy były częściowo realizowane w ramach grantów NCN (OPUS) numer 2016/21/B/HS4/01562 (**A1**, **A3**, **A4**, **A6**, **A10**) oraz 2017/27/B/HS4/01216 (**A7**, **A8**, **A9**), w których doktorant był wykonawcą/stypendystą.

W trakcie prac nad rozprawą doktorską zidentyfikowano potencjalne obszary dalszych prac badawczych. Korzystne do dalszych badań byłoby mapowanie rzeczywistych sieci o węzłach charakteryzowanych wieloma atrybutami na środowisko realizacji procesów propagacji. Planowany jest także dalszy rozwój zrównoważonego podejścia do inicjalizacji pojedynczych procesów realizujących jednocześnie wiele celów, tradycyjnie uzyskiwanych przez odrębne kampanie.

7. Dorobek akademicki

Niniejszy rozdział prezentuje dorobek akademicki mgr inż. Artura Karczmarczyka, kandydata do stopnia naukowego doktora. W ramach prezentowanego dorobku wyodrębniono osiągnięcia naukowe, dydaktyczne i organizacyjne.

7.1. Dorobek naukowy

7.1.1. Profile internetowe[1]

Wskaźniki z profili internetowych Web of Science, Scopus oraz Google Scholar zebrane zostały w tabeli 7.1.

Tablica 7.1. Profile internetowe (stan na dzień 6.05.2021).

Profil	Liczba artykułów	Liczba cytowań	h-index
Web of Science	16	201	9
Scopus	27	429	12
Google Scholar	30	547	13

7.1.2. Wykaz prac naukowych[2]

Poniżej przedstawiono wykaz prac naukowych opublikowanych oraz w druku. Łączny uzyskany współczynnik IF wynosi 19.663. Łączna liczba punktów ministerialnych za publikacje bez uwzględnienia oświadczeń: 295 (do 2017) oraz 760 (od 2018).

1. Wątróbski, J., Jankowski, J., Ziemba, P., **Karczmarczyk, A.**, Zioło, M. (2019). Generalised framework for multi-criteria method selection. Omega, 86, 107-124.
   [IF: 5.341, 140 pkt, 170 cytowań]

2. Wątróbski, J., Małecki, K., Kijewska, K., Iwan, S., **Karczmarczyk, A.**, Thompson, R. G. (2017). Multi-criteria analysis of electric vans for city logistics. Sustainability, 9(8), 1453
   [IF: 2.075, 20 pkt, 55 cytowań]

[1] Stan na dzień 4.05.2021 r.
[2] Stan na dzień 4.05.2021 r.

3. Ziemba, P., Wątróbski, J., Zioło, M., **Karczmarczyk, A.** (2017). Using the PROSA method in offshore wind farm location problems. Energies, 10(11), 1755.
[IF: 2.676, 25 pkt, 51 cytowań, okładka]

4. Wątróbski, J., Ziemba, E., **Karczmarczyk, A.**, Jankowski, J. (2018). An index to measure the sustainable information society: the Polish households case. Sustainability, 10(9), 3223.
[IF: 2.075, 20 pkt, 35 cytowań]

5. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2018). Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one, 13(12), e0209372.
[IF: 2.776, 100 pkt, 34 cytowania]

6. Sałabun, W., **Karczmarczyk, A.** (2018). Using the comet method in the sustainable city transport problem: an empirical study of the electric powered cars. Procedia computer science, 126, 2248-2260.
[WOS 15 pkt, 30 cytowań]

7. Wątróbski, J., Jankowski, J., Ziemba, P., **Karczmarczyk, A.**, Zioło, M. (2019). Generalised framework for multi-criteria method selection: Rule set database and exemplary decision support system implementation blueprints. Data in brief, 22, 639.
[40 pkt, 29 cytowań]

8. Wątróbski, J., Sałabun, W., **Karczmarczyk, A.**, Wolski, W. (2017, September). Sustainable decision-making using the COMET method: An empirical study of the ammonium nitrate transport management. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 949-958). IEEE.
[WOS 20 pkt, 28 cytowań]

9. Sałabun, W., **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2018, November). Handling data uncertainty in decision making with COMET. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1478-1484). IEEE.
[WOS 20pkt, 27 cytowań]

10. Sałabun, W., **Karczmarczyk, A.**, Wątróbski, J. (2018, November). Decision-making using the hesitant fuzzy sets COMET method: An empirical study of the electric city buses selection. In 2018 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1485-1492). IEEE.
[20 cytowań]

11. Jankowski, J., Zioło, M., **Karczmarczyk, A.**, Wątróbski, J. (2018). Towards sustainability in viral marketing with user engaging supporting campaigns. Sustainability, 10(1), 15.
[IF: 2.075, 20 pkt, 18 cytowań]

12. Ziemba, P., Wątróbski, J., **Karczmarczyk, A.**, Jankowski, J., Wolski, W. (2017, September). Integrated approach to e-commerce websites evaluation with the use of surveys and eye tracking based experiments. In 2017 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 1019-1030). IEEE.
[WOS 15 pkt, 15 cytowań]

13. Wątróbski, J., Jankowski, J., **Karczmarczyk, A.**, Ziemba, P. (2017, September). Integration of eye-tracking based studies into e-commerce websites evaluation process with eQual and TOPSIS methods. In EuroSymposium on Systems Analysis and Design (pp. 56-80). Springer, Cham.
[WOS 15 pkt, 14 cytowań]

14. Wątróbski, J., **Karczmarczyk, A.** (2017). Application of the fair secret exchange protocols in the distribution of electronic invoices. Procedia computer science, 112, 1819-1828.
[WOS 15 pkt, 4 cytowania]

15. **Karczmarczyk, A.**, Jankowski, J., Sałabun, W. (2017). Linguistic query based quality evaluation of selected image search engines. Procedia computer science, 112, 1809-1818.
[WOS 15 pkt, 3 cytowania]

16. **Karczmarczyk, A.**, Wątróbski, J., Ladorucki, G., Jankowski, J. (2018, September). MCDA-based approach to sustainable supplier selection. In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 769-778). IEEE.
[WOS 15 pkt, 2 cytowania]

17. **Karczmarczyk, A.**, Bortko, K., Bartków, P., Pazura, P., Jankowski, J. (2018, August). Influencing information spreading processes in complex networks with probability spraying. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1038-1046). IEEE.
[WOS 15 pkt, 2 cytowania]

18. Wątróbski, J., **Karczmarczyk, A.**, Jankowski, J., Ziemba, P., Wolski, W. (2017). Hierarchical Representation of Website Evaluation Model Using Survey and Perceptual Based Criteria. In Information Technology for Management. Ongoing Research and Development (pp. 229-248). Springer, Cham.
[WOS 15 pkt, 2 cytowania]

19. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2019). Multi-Criteria Approach to Planning of Information Spreading Processes Focused on Their Initialization With the Use of Sequential Seeding. In Information Technology for Management: Current Research and Future Directions (pp. 116-134). Springer, Cham.
[1 cytowanie]

20. **Karczmarczyk, A.**, Jankowsk, J., Wątróbski, J. (2019, September). Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 663-673). IEEE.
[1 cytowanie]

21. **Karczmarczyk, A.**, Jankowski, J., Watrobski, J. (2019). Parametrization of spreading processes within complex networks with the use of knowledge acquired from network samples. Procedia Computer Science, 159, 2279-2293.
[70pkt, 1 cytowanie]

22. Jankowski, J., Michalski, R., Bródka, P., **Karczmarczyk, A.** (2017, July). Increasing Coverage of Information Diffusion Processes by Reducing the Number of Initial Seeds.

In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 713-720).
[WOS 15 pkt, 1 cytowanie]

23. Sałabun, W., **Karczmarczyk, A.**, Mejsner, P. (2017). Experimental Study of Color Contrast Influence in Internet Advertisements with Eye Tracker Usage. In Neuroeconomic and Behavioral Aspects of Decision Making (pp. 365-375). Springer, Cham.
[WOS 15 pkt, 1 cytowanie]

24. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). OONIS—Object-Oriented Network Infection Simulator. SoftwareX, 14, 100675.
[200 pkt]

25. **Karczmarczyk, A.**, Jankowski, J., Wątrobski, J. (2021). Multi-Criteria Seed Selection for Targeting Multi-Attribute Nodes in Complex Networks. Symmetry, 13(4), 731.
[IF: 2.645, 70pkt]

26. Rymaszewski, S., Wątróbski, J., **Karczmarczyk, A.** (2020). Identification of reference multi criteria domain model-Production line optimization case study. Procedia Computer Science, 176, 3794-3801.
[70 pkt]

27. Wątróbski, J., **Karczmarczyk, A.**, Rymaszewski, S. (2020). Multi-criteria decision making approach to production line optimization. Procedia Computer Science, 176, 3820-3830.
[70 pkt]

28. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2018). Comparative Study of Different MCDA-Based Approaches in Sustainable Supplier Selection Problem. In Information Technology for Management: Emerging Research and Applications (pp. 176-193). Springer, Cham.

29. **Karczmarczyk, A.** (2011). Zastosowanie sprawiedliwej wymiany sekretów w dystrybucji faktur elektronicznych. Studia i Materialy Polskiego Stowarzyszenia Zarzadzania Wiedza/Studies Proceedings Polish Association for Knowledge Management, (57).

30. **Karczmarczyk, A.**, Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeted Influence Maximization within Social Networks – in proceedings of International Conference on Computational Science: ICCS 2021
[w druku, planowany czas publikacji - druga połowa 2021]

31. **Karczmarczyk, A.**, Wątróbski, J., Jankowski, J. (2021). Seeding for Complementary Campaign Objectives in Social Networks - in proceedings of The Americas Conference on Information Systems: AMCIS 2021
[w druku, planowany czas publikacji - druga połowa 2021]

7.1.3. Charakterystyka pozostałego dorobku naukowego

Pozostały dorobek naukowy spoza cyklu **A1-A10** obejmuje 21 publikacji naukowych. Są to publikacje o zasięgu krajowym i światowym. Dużą ich część stanowią publikacje w

międzynarodowych czasopismach oraz materiały z konferencji międzynarodowych indeksowane w bazach Scopus oraz WebOfScience. Tematyka zrealizowanych prac koncentruje się wokół zagadnień metodycznych i praktycznych związanych z wykorzystaniem oraz rozwojem metod MCDA. Doświadczenie pozyskane w realizacji prac poza głównym cyklem wspomogło późniejsze opracowanie rozwiązań MCDA w tematyce propagacji informacji w sieciach społecznych.

Prace (4)[3], (12), (13), (18) koncentrują się na aplikacji i doskonaleniu metod MCDA w obszarze cyfrowego zrównoważenia, a w szczególności: poszukiwania zrównoważonych interfejsów, zrównoważoną ocenę jakości systemów informatycznych oraz witryn internetowych. Wzrost roli informacji i technologii informacyjno-komunikacyjnych (ICT) doprowadził do wykształcenia nowych pojęć, a w tym pojęcia społeczeństwa informacyjnego (ang. information society). Pojęcie to stanowi adaptację pryncypiów zrównoważonego rozwoju na grunt społeczeństwa informacyjnego (ang. sustainable information society, SIS). Problem badawczy podjęty w pracy (4) stanowiło opracowanie metodyki pomiaru zrównoważonego społeczeństwa informacyjnego. Uwzględniono przy tym literaturę przedmiotu, gdzie do podstawowych zadań realizowanych podczas pomiaru SIS zalicza się: możliwość elastycznej definicji kryteriów i celów pomiaru SIS oraz ocenę stopnia realizacji zakładanych celów. W pracy zaproponowano autorski indeks pomiaru SIS oparty na wybranych technikach MCDA.

Cykl publikacji (12), (13), (18) stanowi próbę budowy nowej metody oceny jakości serwisów internetowych. Stosowane w literaturze metody oceny serwisów internetowych jak eQual, Ahn, czy SiteQual oparte są na trywialnym aparacie matematycznym. Najczęściej stosowana jest w nich średnia arytmetyczna, służąca do agregacji ocen. Warto zauważyć, że zastosowanie metod MCDA do oceny serwisów internetowych niesie większy potencjał niż tylko konstrukcja rankingu. Przedstawiona w pracach (12), (13), (18) autorska metoda Pequal oparta jest na funkcji dostosowania jakości (Quality Function Deployment) oraz algorytmach wybranych metod MCDA. Funkcja dostosowania jakości jest ustrukturalizowanym procesem zapewniającym środki identyfikacji i dostarczającym opinii użytkowników o jakości produktu na kolejnych etapach jego tworzenia. Metodycznie, w pracach wykorzystano metody Promethee II, TOPSIS, AHP oraz COMET, również w wersjach rozmytych. Podczas badań, oprócz starannie zgromadzonych danych ankietowych, dążąc do obiektywizacji eksperckiej oceny, rozszerzono zbiór danych wejściowych o dane pomiarowe z urządzeń typu eye-tracker.

Prace (2), (3), (6), (8), (16), (28) koncentrują się na poszukiwaniu rozwiązań metodycznych i algorytmicznych wybranych problemów zrównoważonego rozwoju. Prace posadowione są w ważnych i aktualnych problematykach odnawialnych źródeł energii oraz zrównoważonego transportu i logistyki. Obszary te są szeroko dyskutowane w literaturze przedmiotu, a konkluzje autorów wyraźnie wskazują na potrzebę doskonalenia warsztatu istniejących modeli wspomagania tego procesu decyzyjnego. W tym zakresie publikacje te stanowią odpowiedź na aktualne wyzwania badawcze.

W pracy (3) opracowano nową metodę wielokryterialną PROSA. Przy jednoczesnym zachowaniu w niej uniwersalnych właściwości metody Promethee II, ograniczono w niej efekt liniowej kompensacji czynników. W wyniku przeprowadzonych badań dowiedziono, że metoda PROSA preferuje bardziej zrównoważone alternatywy decyzyjne spełniając paradygmat strong sustainability.

[3]W sekcji 7.1.3 numeracja literatury odnosi się do pozycji wykazanych w sekcji 7.1.2.

Grupa prac (2), (6), (8), (16), (28) posadowiona jest w obszarze zrównoważonego transportu i logistyki. W pracy (2) wykorzystano metody wielokryterialne w ocenie możliwości wdrożenia frachtowych, elektrycznych samochodów w obszarze logistyki miejskiej. Kontynuacją i dopełnieniem tej pracy jest publikacja (6), gdzie używając wielokryterialnej metody COMET dokonano identyfikacji pełnego modelu domenowego (obszar samochodów elektrycznych). W obszarze zrównoważonego zarządzania transportem posadowiona jest praca (8). Tutaj również dokonano porównania efektywności szeregu modeli wielokryterialnych. W pracach (16) oraz (28) podjęto się wielokryterialnego modelowania ważnego i aktualnego problemu oceny i doboru zrównoważonego dostawcy. Dokonano tutaj oceny efektywności wybranych metod MCDA.

W kolejny nurt badań wpisują się prace (7), (9) oraz (10). Obejmują one obszar rozwoju nowych metod MCDA. W szczególności prace dotyczą adaptacji arytmetyki przedziałowej czy kolejnych rozwinięć arytmetyki rozmytej (hesistant fuzzy sets) w wybranych metodach wielokryterialnych.

Dodatkowo, prowadzone były prace powiązane z dystrybucją faktur elektronicznych (14), (29), oceną jakości wybranych wyszukiwarek obrazów (15), oceną intensywności interaktywnego przekazu reklamowego (23), jak też budową wielokryterialnych modeli optymalizacji linii produkcyjnej (26), (27).

7.1.4. Pozostałe

— Nagroda Rektora ZUT w Szczecinie za wybitne osiągnięcia naukowe stopnia drugiego za rok 2018.

— Nagroda Best Paper Award za publikację [41] na konferencji FedCSIS 2017.

— Publikacja [42] została wyróżniona spośród wszystkich 248 pozycji poprzez umieszczenie na okładce wydania czasopisma *Energies, Volume 10, Issue 11 (November 2017)*.

— Publikacja A2 od 2019 roku znajduje się na liście najczęściej pobieranych artykułów otwartych czasopisma Omega The International Journal of Management Science wydawnictwa Elsevier o wskaźniku IF 5.324. Ponadto została oznaczona jako "Highly Cited Paper"i "Hot Paper"w bazie WoS.

— Uczestnictwo i wygłoszenie referatów na międzynarodowych konferencjach ASONAM 2018 (The 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining) w Barcelonie i FedCSIS (14th Federated Conference on Computer Science and Information Systems) w Lipsku.

— Funkcja Guest Editor w Special Issue Exploring of Sustainable Supplier Selection w 2018 roku w czasopiśmie Sustainability wydawnictwa MDPI – IF 2.576.

— Zaproszenie do pełnienia funkcji Guest Editor w Special Issue Production Line Optimization and Sustainability w 2021 roku w czasopiśmie Sustainability wydawnictwa MDPI – IF 2.576.

— Trzykrotny finalista Ogólnopolskiej Olimpiady Języka Angielskiego Dla Studentów Wyższych Uczelni Technicznych w latach 2006, 2007 i 2009.

— W roku 2007 uczestnictwo w finale regionalnym konkursu ACM Central European Programming Contest w Pradze.

— W roku 2011 I miejsce w konkursie e-point na najlepsze prace dyplomowe z zakresu rozwiązań internetowych realizowanych w technologii JAVA.

7.2. Dorobek dydaktyczny

7.2.1. Kursy i sylabusy

W ramach dotychczasowej pracy dydaktycznej prowadziłem zajęcia na poziomie S1 zgodnie z wykazem:

1. Programowanie systemów i aplikacji internetowych
   – *cykl 15 wykładów i 15 laboratoriów*
2. Systemy i platformy biznesu elektronicznego
   – *cykl 8 wykładów i 15 laboratoriów*
3. Systemy i platformy biznesu cyfrowego
   – *cykl 15 wykładów i 15 laboratoriów*
4. Programowanie dokumentów dynamicznych
   – *cykl 8 wykładów*
5. Aplikacje internetowe 1
   – *cykl 15 wykładów i 15 laboratoriów*
6. Pracownia dyplomowa
   – *cykl 8 seminariów*
7. Inżynierski projekt zespołowy
   – *cykl 4 wykładów i 15 spotkań projektowych*
8. Wprowadzenie do informatyki
   – *wygłoszenie 3 wykładów w ramach większego cyklu prowadzonego przez wielu wykładowców*

Ponadto współtworzyłem sylabusy do następujących kursów na poziomie S1:

1. Aplikacje internetowe 1
2. Aplikacje internetowe 2
3. Aplikacje internetowe

7.2.2. Prace dyplomowe

Podczas dotychczasowej pracy dydaktycznej byłem promotorem następujących prac inżynierskich:

1. Mobilny system informacji dla pasażerów komunikacji miejskiej
   praca inżynierska, obrona w 2018 r.
2. System rejestracji i zarządzania wizytami w klinikach weterynaryjnych
   praca inżynierska, obrona w 2019 r.

3. Aplikacja webowa wspierająca użycie fiszek w procesie nauki i zapamiętywania
   praca inżynierska, obrona w 2019 r.
4. Wyszukiwarka gier wideo
   praca inżynierska, obrona w 2020 r.
5. Internetowy system zarządzania i wynajmu mieszkań
   praca inżynierska, obrona w 2021 r.
6. System do zarządzania i współdzielenia ulubionych witryn WWW
   praca inżynierska, obrona w 2021 r.
7. Aplikacja mobilna do wyszukiwania najkorzystniejszej stacji benzynowej w okolicy
   praca inżynierska, obrona planowana w II kwartale 2021 r.
8. Mobilna aplikacja informująca pasażerów komunikacji miejskiej o wydarzeniach na drodze
   praca inżynierska, obrona planowana w II kwartale 2021 r.
9. Projekt i implementacja w wersji mobilnej dynamicznego modelu wielokryterialnego do wspomagania decyzji wyboru produktów konsumenckich na przykładzie samochodów
   praca magisterska, obrona planowana w II kwartale 2021 r.
10. Projekt i implementacja systemu wspomagającego trwałe zapamiętywanie obcojęzycznego słownictwa z wykorzystaniem teorii krzywej zapominania Hermanna Ebbinghausa
    praca inżynierska, obrona planowana w III kwartale 2021 r.
11. Projekt i implementacja dwuwymiarowej gry online Okręty z wykorzystaniem technologii webowych i API
    praca inżynierska, obrona planowana w 2022 r.
12. System do webowej i mobilnej nauki języków obcych
    praca inżynierska, obrona planowana w 2022 r.
13. Projekt i implementacja trójwymiarowej gry strategicznej
    praca inżynierska, obrona planowana w 2022 r.
14. System do ewidencjonowania czasu pracy oraz zarządzania pracownikami w przedsiębiorstwie
    praca inżynierska, obrona planowana w 2022 r.
15. Projekt i implementacja systemu gromadzenia i przetwarzania statystyk z gier komputerowych
    praca inżynierska, obrona planowana w 2022 r.
16. Internetowy system do agregacji i analizy treningów sportowych
    praca inżynierska, obrona planowana w 2022 r.

7.3. Dorobek organizacyjny

— Jestem aktywnym członkiem Association for Computing Machinery (ACM) od 2008 roku.

— W latach 2007-2009 byłem prezesem koła naukowego TWIPS działającego na Wydziale Informatyki ZUT w Szczecinie.

— W ramach prowadzonej działalności gospodarczej prowadzonej pod nazwą IdeaSpot Artur Karczmarczyk przeprowadziłem praktyki studenckie dla 5 osób oraz praktyki absolwenckie dla 8 osób w latach 2017-2021.

— Jestem członkiem Kolegium Wydziału Informatyki ZUT w Szczecinie.

— Jestem przewodniczącym sesji *WS01: 4th Symposium on Information Systems and Technologies for Management and Economy* na konferencji *25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems 2021* z ponad 20 zgłoszonymi i recenzowanymi publikacjami.

— Uczestniczyłem we współpracy pomiędzy Technoparkiem Pomerania i Wydziałem Informatyki ZUT w Szczecinie w tworzeniu bloku przedmiotów obieralnych Komercjalizacja.

7.4. Dorobek zawodowy

7.4.1. Historia zatrudnienia

2006-2014 Unizeto Technologies S.A.
ścieżka kariery obejmująca stanowiska kwalifikowanego podwykonawcy, wdrożeniowca, programistę, skończywszy na inżynierze oproramowania w dziale kierowania projektami

2014-obecnie IdeaSpot Artur Karczmarzcyk
właściciel jednoosobowej działalności gospodarczej, w ramach której w szczytowym momencie zatrudnione było 12 osób

2017-obecnie Wydział Informatyki ZUT w Szczecinie
pracownik naukowo-dydaktyczny na stanowisku aystent

7.4.2. Najciekawsze projekty

— Współtworzylem zaawansowany sklep internetowy wraz z czytelnią online norm dla Polskiego Komitetu Normalizacyjnego.

— Uczestniczyłem w tworzeniu prototypu Internetowego Konta Pacjenta – program pilotażowy dla grupy pacjentów diabetycznych z południa Polski.

— Współtworzę system anonimowej eskalacji problemów związanych z mobbingiem wykorzystywany m.in. w brytyjskim NHS czy francuskim Deezer.

— Kieruję wytwarzaniem oprogramowania dla platformy esport wyposażonej między innymi w systemy zaproszeń znajomych czy komunikator tekstowy.

7.4.3. Najciekawsze szkolenia i certyfikaty zawodowe

— Zend Certified PHP Engineer

— Szkolenie Programowanie Zespołowe w ramach Microsoft IT Academy Programme

— Certyfikat REQB - Certyfikowany Profesjonalista Inżynierii Wymagań

— Szkolenie w ramach Uniwersystetu Ernst&Young

— CS169.1x: Software as a Service (BerkeleyX, The University of California, Berkeley through edX)

— Introducing Agile Software Development

— Managing Agile Software Development

— Planning Agile Software Development

— Create Work Breakdown Structure (PMBOK Guide Fourth Edition)

— Estimating Activity Resources and Durations (PMBOK Guide Fourth Edition)

— Managing Projects within Organisations (PMBOK Guide Fourth Edition)

— Project Management Overview (PMBOK Guide Fifth Edition)

— Project Management Process Groups

Spis rysunków

Spis tablic

Bibliografia

[1] R. Hanna, A. Rohm, and V. L. Crittenden, "We're all connected: The power of the social media ecosystem," *Business horizons*, vol. 54, no. 3, pp. 265–273, 2011.

[2] D. J. Watts, J. Peretti, and M. Frumin, *Viral marketing for the real world*. Harvard Business School Pub., 2007.

[3] J. L. Iribarren and E. Moro, "Impact of human activity patterns on the dynamics of information diffusion," *Physical review letters*, vol. 103, no. 3, p. 038702, 2009.

[4] J. Berger and K. L. Milkman, "What makes online content viral?," *Journal of marketing research*, vol. 49, no. 2, pp. 192–205, 2012.

[5] J. Y. Ho and M. Dempsey, "Viral marketing: Motivations to forward online content," *Journal of Business research*, vol. 63, no. 9-10, pp. 1000–1006, 2010.

[6] X. Zhang, D.-D. Han, R. Yang, and Z. Zhang, "Users' participation and social influence during information spreading on twitter," *PloS one*, vol. 12, no. 9, p. e0183290, 2017.

[7] O. Hinz, B. Skiera, C. Barrot, and J. U. Becker, "Seeding strategies for viral marketing: An empirical comparison," *Journal of Marketing*, vol. 75, no. 6, pp. 55–71, 2011.

[8] Y. Liu-Thompkins, "Seeding viral content: The role of message and network factors," *Journal of Advertising Research*, vol. 52, no. 4, pp. 465–478, 2012.

[9] C. Kiss and M. Bichler, "Identification of influencers: measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.

[10] M. Bampo, M. T. Ewing, D. R. Mather, D. Stewart, and M. Wallace, "The effects of the social structure of digital networks on viral marketing performance," *Information systems research*, vol. 19, no. 3, pp. 273–290, 2008.

[11] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media: sentiment of microblogs and sharing behavior," *Journal of management information systems*, vol. 29, no. 4, pp. 217–248, 2013.

[12] A. Dobele, A. Lindgreen, M. Beverland, J. Vanhamme, and R. Van Wijk, "Why pass on viral messages? because they connect emotionally," *Business Horizons*, vol. 50, no. 4, pp. 291–304, 2007.

[13] C. Camarero and R. San José, "Social and attitudinal determinants of viral marketing dynamics," *Computers in Human Behavior*, vol. 27, no. 6, pp. 2292–2300, 2011.

[14] S. Helm, "Viral marketing-establishing customer relationships by'word-of-mouse'," *Electronic markets*, vol. 10, no. 3, pp. 158–161, 2000.

[15] K. Kandhway and J. Kuri, "How to run a campaign: Optimal control of sis and sir information epidemics," *Applied Mathematics and Computation*, vol. 231, pp. 79–92, 2014.

[16] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, ACM, 2003.

[17] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[18] D. Cruz and C. Fill, "Evaluating viral marketing: isolating the key criteria," *Marketing Intelligence & Planning*, vol. 26, no. 7, pp. 743–758, 2008.

[19] A. Mochalova and A. Nanopoulos, "A targeted approach to viral marketing," *Electronic Commerce Research and Applications*, vol. 13, no. 4, pp. 283–294, 2014.

[20] H. T. Nguyen, T. N. Dinh, and M. T. Thai, "Cost-aware targeted viral marketing in billion-scale networks," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, 2016.

[21] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[22] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[23] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[24] P. ERDdS and A. R&WI, "On random graphs i," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[25] J.-P. Brans and B. Mareschal, "Promethee methods," in *Multiple criteria decision analysis: state of the art surveys*, pp. 163–186, Springer, 2005.

[26] J. Wątróbski, J. Jankowski, P. Ziemba, A. Karczmarczyk, and M. Zioło, "Generalised framework for multi-criteria method selection: Rule set database and exemplary decision support system implementation blueprints," *Data in brief*, vol. 22, p. 639, 2019.

[27] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 03 1951.

[28] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the national academy of sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[29] M. Ripeanu, I. Foster, and A. Iamnitchi, "Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design," *arXiv:cs/0209028*, Sept. 2002. arXiv: cs/0209028.

[30] C.-L. Hwang and K. Yoon, "Methods for multiple attribute decision making," in *Multiple attribute decision making*, pp. 58–191, Springer, 1981.

[31] "Snowball Sampling Function - R Documentation."

[32] J. Jankowski, P. Bródka, P. Kazienko, B. K. Szymanski, R. Michalski, and T. Kajdanowicz, "Balancing speed and coverage by sequential seeding in complex networks," *Scientific reports*, vol. 7, no. 1, p. 891, 2017.

[33] K. E. Read, "Cultures of the central highlands, new guinea," *Southwestern Journal of Anthropology*, pp. 1–43, 1954.

[34] G. Voss, A. Godfrey, and K. Seiders, "Do satisfied customers always buy more? the roles of satiation and habituation in customer repurchase," *Marketing Science Institute Working Paper Series 2010*, pp. 10–101, 2010.

[35] R. Likert, "A technique for the measurement of attitudes.," *Archives of psychology*, 1932.

[36] T. L. Saaty and L. G. Vargas, "The legitimacy of rank reversal," *Omega*, vol. 12, no. 5, pp. 513–516, 1984.

[37] M. Zdrowia, "Program profilaktyki raka piersi (mammografia) - ministerstwo zdrowia - portal gov.pl," Apr 2018.

[38] Gus, "Ludność. stan i struktura ludności oraz ruch naturalny w przekroju terytorialnym. stan w dniu 30 vi 2015 r.," Feb 2016.

[39] B. J. Cox, "Object-oriented programming: an evolutionary approach," 1986.

[40] A. Karczmarczyk, J. Jankowski, and J. Wątróbski, "Oonis—object-oriented network infection simulator," *SoftwareX*, vol. 14, p. 100675, 2021.

[41] P. Ziemba, J. Wątróbski, A. Karczmarczyk, J. Jankowski, and W. Wolski, "Integrated approach to e-commerce websites evaluation with the use of surveys and eye tracking based experiments," in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 1019–1030, IEEE, 2017.

[42] P. Ziemba, J. Wątróbski, M. Zioło, and A. Karczmarczyk, "Using the prosa method in offshore wind farm location problems," *Energies*, vol. 10, no. 11, p. 1755, 2017.

A1.

Karczmarczyk, A., Jankowski, J., Wątróbski, J. (2018). Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one, 13(12), e0209372.

PLOS ONE

RESEARCH ARTICLE

Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks

Artur Karczmarczyk[1], **Jarosław Jankowski**[1]*, **Jarosław Wątróbski**[2]

1 Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, ul. Żołnierska 49, 71-210 Szczecin, Poland, **2** Faculty of Economics and Management, University of Szczecin, Mickiewicza 64, 71-101, Szczecin, Poland

* jjankowski@wi.zut.edu.pl

Citation: Karczmarczyk A, Jankowski J, Wątróbski J (2018) Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PLoS ONE 13(12): e0209372. https://doi.org/10.1371/journal.pone.0209372

Editor: Lidia Adriana Braunstein, Universidad Nacional de Mar del Plata, ARGENTINA

Received: July 27, 2018

Accepted: December 4, 2018

Published: December 27, 2018

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The current marketing landscape, apart from conventional approaches, consists of campaigns designed especially for launching information diffusion processes within online networks. Associated research is focused on information propagation models, campaign initialization strategies and factors affecting campaign dynamics. In terms of algorithms and performance evaluation, the final coverage represented by the fraction of activated nodes within a target network is usually used. It is not necessarily consistent with the real marketing campaigns using various characteristics and parameters related to coverage, costs, behavioral patterns and time factors for overall evaluation. This paper presents assumptions for a decision support system for multi-criteria campaign planning and evaluation with inputs from agent-based simulations. The results, which are delivered from a simulation model based on synthetic networks in a form of decision scenarios, are verified within a real network. Last, but not least, the study proposes a multi-objective campaign evaluation framework with several campaign evaluation metrics integrated. The results showed that the recommendations generated with the use of synthetic networks applied to real networks delivered results according to the decision makers' expectation in terms of the used evaluation criteria. Apart from practical applications, the proposed multi-objective approach creates new evaluation possibilities for theoretical studies focused on information spreading processes within complex networks.

1 Introduction

The evolution of social networking platforms has led to a crucial need to understand how millions of online users behave, including their online and real life behaviors, patterns and predispositions [1]. Apart from studying social relations and online activity, information spreading processes are among the phenomenas with high attention from both researchers and practitioners. In a number of cases, as a result of information spreading, viral marketing seems to produce better results than traditional advertising campaigns [2]. There is an increase in the

number of online marketers using this opportunity to place even greater efforts in the engagement of potential consumers to benefit from their services and products by propagating information. Due to an increased trustworthiness of communications within a social network that has ties that are particularly strong, recommendations that are socially oriented have a greater impact on the targeted consumers than traditional commercial messages [3]. The research that is related to diffusion of marketing content takes into consideration the factors that lead to campaigns that are successful [4] [5], factors affecting usersÔÇÖ participation during information spreading [6], the initial seed sets that are selected for the initialization of the campaign [7] [8], as well as using epidemic models to analyze diffusion processes [9]. Other studies emphasize the role of different centrality measures used for the selection of initial influencers [10], the impact of homophily for successful selection of the initial network nodes [11], the role of the content and network structures [12], user motivation to forward content [5], the role of emotions [13] [14] and other factors [15]. Apart from static networks and single layer structures, multilayer networks [16] and the spreading of information in temporal networks have been studied in the more recent research [17].

Many earlier studies were focused on theoretical and empirical approaches increasing the number of reached customers within a network [7] [18]. While it is an important metric of the campaign success, several other factors should also be taken into account [19]. Apart from coverage, they include the campaign's costs and duration, number of initial seeds and their selection strategies [20] [21].

This study proposes and examines the framework for a multi-objective evaluation of information spreading processes. The presented framework can be used for strategic planning of information spreading processes in order to help selecting the appropriate strategy for selection of the initial nodes within the network and adjusting the number of activated nodes in the seeding process. While viral marketing processes can be based on increasing the motivation of content forwarding, the evaluation of the potential of available approaches creates another areas of applications of the multi-objective methods. The main contribution of the presented study is the framework for multi-objective selection of methods influencing campaign dynamics and coverage with the use of several evaluation criteria. In practical terms, an evaluation model was created with the use of the PROMETHEE II method and agent-based simulations were performed with sensitivity analysis used.

The remainder of this paper provides the methodological background and the conceptual framework in Section 2. This is followed by the example planning process in Section 3, data evaluation and searching for alternatives in Section 4 with conclusions in Section 5.

2 Materials and methods

2.1 Information spreading in complex networks

Social network marketing strategies are geared to motivating users to pass the advertised product information to their friends and contacts within their social networks. With its interdisciplinary approach, the research that has been done in this field attracts sociologists, physicists, computer scientists and marketers with a wide range of approaches and research goals [7] [9] [3]. The prior research in this field implemented a macroscopic approaches to analyze the quantity of customers acquired using the diffusion of innovations' mechanics [22] [23]. The processes at the level of social networks, as well as their participants, are monitored at a detailed level, offering a microscopic view. The identification and assessment of those who send and receive messages make the detailed monitoring of the processes involved in the distribution of information possible [24].

The methodological background on network structures evolved simultaneously, yet separately, on various disciplines [25]. In this paper, network G is defined as a set of nodes (vertices) $V(G)$ interconnected with the set of edges $E(G)$, which can be represented with the following mathematical notation: $G(V, E)$. A path in graph G is a set of edges $\{\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_{n-1}, v_n\}\}$, where the end of the $\{v_i, v_{i+1}\}$ edge is the beginning of the $\{v_{i+1}, v_{i+2}\}$ edge for every $i = 0, \ldots, n - 2$, and where every node and edge are unique. The length of a path is the number of edges it comprises of. The distance $d(i, j)$ from node i to node j is the length of the shortest path from i to j.

The current research done in the field can be identified as taking different directions. The dedicated solutions like linear threshold [22] and independent cascades model [26], as well as epidemic research models are implemented to model the way information is spread [9]. A large number of studies relate to the initiation of the information distribution processes and network node selection [7]. The most of the seeding strategies use network centrality measures for obtaining the nodes' ranking and initiating the seeding process, assuming an increased potential to distribute information resulting from the top nodes, having a vital role in the network structures. One of the most fundamental characteristics of a graph's node is its degree, i.e. the number of edges incident to this node, denoted $deg(v)$, where $v \in V$. In case of information spreading, the higher the node's degree, the more nodes the information can be potentially propagated to. The degree distribution $P(k)$ of a network represents the fraction of the nodes in the network which have degree equal to k. Other measures based on closeness, betweenness or eigenvector centralities are used as well. The closeness of a node is a measure of its centrality in a network, calculated as the sum of the lengths of the shortest paths between the node and all other nodes in the network:

$$C(i) = \frac{1}{\sum_j d(j, i)} \tag{1}$$

The smaller the closeness value, the more central position in the network the node has, thus allowing to reach every other node in fewer steps. For every pair of vertices (v_i, v_j) in graph G, there exists at least one shortest path between v_i and v_j with the number of edges on the path minimized. The betweenness of a vertex v_k is the ratio of the number of such shortest paths that pass through v_k to all such shortest paths. The higher the betweenness value of a node, the more nodes can be accessed through that node.

Eventually, the eigenvector centrality is a measure of influence of a node in a network. Each node in a network obtains a relative score based on a concept that connections from the high-scored nodes contribute to the node's score more than connections from the low-scoring ones. Therefore, a high eigenvector centrality value means that a node has more influence on the other nodes in the network.

These approaches tend to be used despite the fact that they require computational resources that are limited and that they fail in the delivery of an optimal seed set. Better results are obtained from solutions that are more sophisticated, like greedy-based selection, along with extensions it may have, however, the computational costs are substantial and it is not easily implemented on networks that are very complex [26].

Structural measures are used to improve optimization, so that nodes in the same network segments are not selected to allow a better seed allocation. These types of solutions are based more on better use of processes of natural diffusion and use sequential seeding [27], avoid nodes from within the same communities with intra connections that are close by using target communities [28], use dynamic rankings with sequential seeding [29] and use mechanisms for voting that have lower weights once activated nodes have been detected [30]. Apart from basic

centrality measures, the central nodes in networks can be detected using a k-shell based approach [31]. Alternative approaches use bio-inspired algorithms to select the initial set of nodes [32].

The majority of earlier approaches are based on networks that are static, while the more recent studies account for networks that are dynamic. They have temporal characteristics and more reality-based specifics, as opposed to static snapshots [17]. The other research paths took base on the multi-layer networks and processes of information spreading that are intra layer [16]. There have been attempts to use other knowledge about the on-going processes to obtain better results by using adaptive seeding, even though the majority of solutions are geared to the seed set initiated processes [33]. Other approaches account for numerous campaigns that are on-going, as well as how they interrelate [34].

Unfortunately, the knowledge about the social network in which the campaign is going to be performed is often limited to some basic characteristics. Moreover, the number of nodes and edges in a real social network is often immense, which makes advanced simulations infeasible. The research takes into account simulations within synthetic networks to investigate phenomena within different network structures. While collecting information about real networks is difficult, synthetic networks based on theoretical models can be used. Moreover, the structure of synthetic networks can be adjusted during the generation process, thus allowing the researchers to perform a more profound analysis of the processes in complex networks. Simulation studies often use networks based on the free-scale model proposed by Barabasi-Albert (BA) [35], small world model proposed by Watts-Strogatz (WS) [36] and random graph model introduced by Erdos-Renyi (ER) [37]. For example [38] used WS, ER and BA synthetic networks for modeling interacting processes, [12] analyzed the role of structures of ER, BA, and WS networks on campaign performance, [39] proposed a framework to analyze multiple spreading processes and verified it with the use of BA networks, [40] used WS networks for cooperative epidemic modeling.

The characteristics BA and WS theoretical models are close to real systems. The Barabasi-Albert network model was first created in 1999, as a result of a study of the at-the-time structure of the WWW. The construction of BA networks is based on two complementary mechanisms: network growth and the mechanism of preferential attachment. The BA model is similar to several natural and human-made systems, such as the Internet, WWW, citation networks or social networks, to name just a few, where several selected nodes (hubs) have unusually high degree compared to the remaining nodes of the network. Fig 1a presents an example of a Barabasi-Albert network and the chart on Fig 1d depicts the degree distribution of a sample BA model. The Erdos-Renyi network model was first described in 1959 and is constructed on the assumptions that at first the number N of nodes is defined and, subsequently, from all $\binom{N}{2}$ pairs of nodes, random E pairs are selected between which the edges are created. A sample ER model and degree distribution of a sample ER model is presented on Fig 1b and 1e respectively. The ER model offers a simple and powerful model with many applications, but might be inappropriate for modeling some real-life phenomena, due to the fact it does not generate local clustering of nodes. Therefore, in 1998 the Watts-Strogatz model was created to addresses this issue. The WS model accounts for clustering, but keeps the short average path lengths from the ER model. Fig 1c presents an example of a WS network and the chart on Fig 1f depicts the degree distribution of a sample WS model.

2.2 MCDA foundations of the proposed approach

In case of a viral marketing campaign in a social network, the ordering party might be interested not only in maximizing the coverage of the campaign, but also in affecting its dynamics,

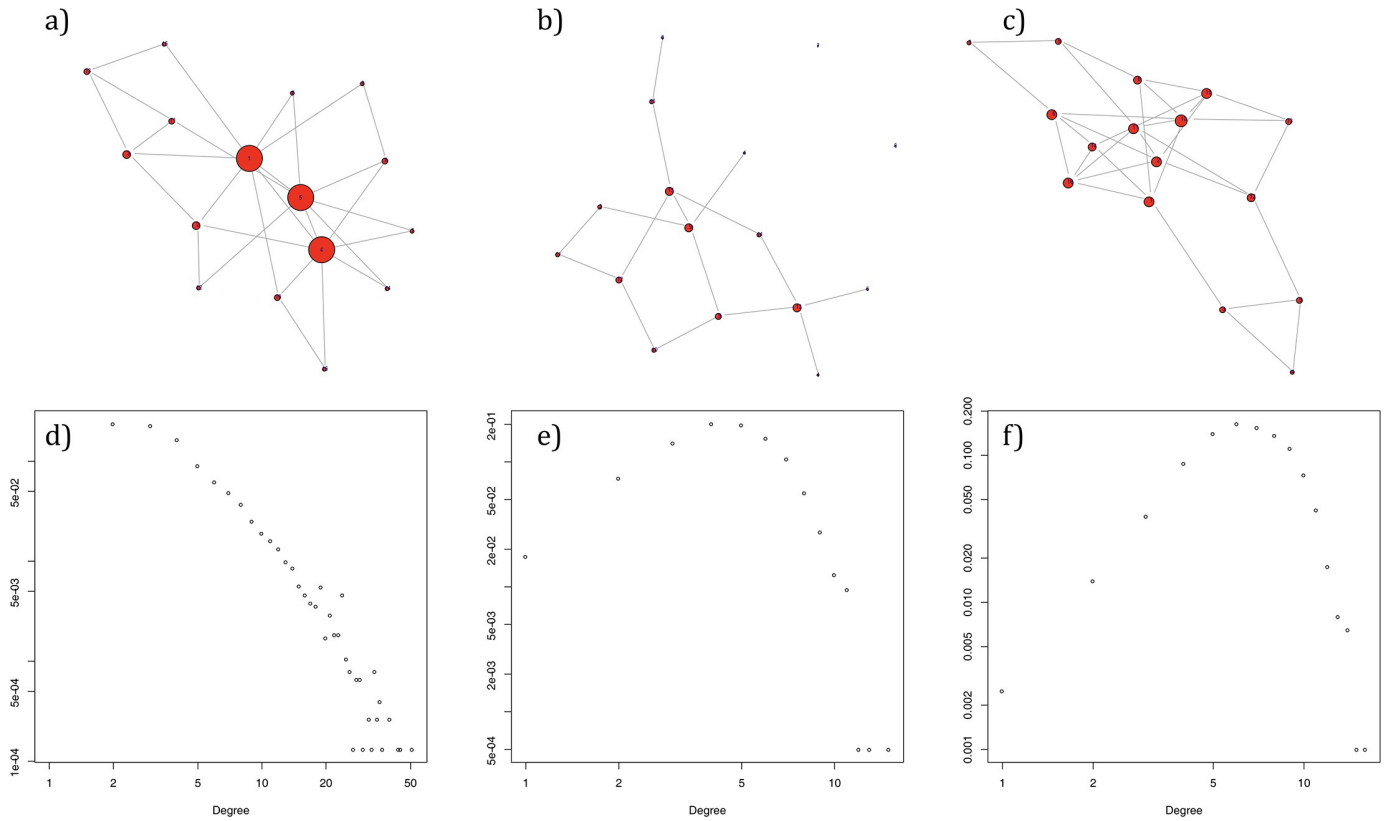


Fig 1. Graph representation of example 16-node synthetic networks: a) BA, b) ER, c) WS; and degree distribution charts of example 2000-node synthetic networks: d) BA, e) ER, f) WS.

https://doi.org/10.1371/journal.pone.0209372.g001

as well as keeping the campaign cost within a reasonable budget. All these aspects need to be considered before launching the campaign. Therefore, planning a viral marketing campaign in social network is a multi-criteria problem, which can be presented as (2) [41]:

$$max\{c_1(a), c_2(a), \ldots, c_k(a) | a \in A\}, \tag{2}$$

where A is a finite set of possible campaign strategies $\{a_1, a_2, \ldots, a_n\}$, whereas $\{c_1(\cdot), c_2(\cdot), \ldots, c_k(\cdot)\}$ is a set of evaluation criteria. Some of the criteria might be maximized and others minimized. The criterial performance of each strategy regarding each criterion can be expressed in a form of a performance table. Intuitively, it is expected from the decision maker (DM) to identify the strategy that optimizes all criteria. However, usually there exists no alternative that optimizes all criteria simultaneously.

Let us consider an example viral marketing campaign, for which multiple alternative strategies were prepared. The strategies are characterized by three criteria: seeding fraction, propagation probability and the potential coverage that can be obtained. The coverage is a very important criterion, however, generally a strategy that obtains 100% coverage is not always chosen, as it would require infecting a massive number of initial seeds in the network or providing multiple incentives to increase the propagation probability in the network. On the other hand, if a strategy with minimal seeding fraction and minimal propagation probability is chosen, it cannot be expected to cover the complete network. Therefore, a compromise solution between the strategies should be chosen.

It is important to note that the solution to a multi-criteria problem depends not only on the criterial performance of each alternative, but also on the campaign ordering party itself. There is no absolute best strategy for all campaigns and the best compromise strategy depends on the preferences of the DM.

Three natural dominance relations can be associated to a decision problem of the multi-criteria nature presented in (2): indifference, preference and incomparability. Let us consider two alternatives a and b. If for every criterion c_i a is as good as b, then the two strategies are indifferent (aIb). If for every criterion c_j a is as good or equal to b and there exists at least a single criterion c_k for which a is better than b, then a is preferred to b (aPb).

Finally, if there is a criterion c_m for which a is better than b, but there also exists a criterion c_n for which b is better than a, then the two strategies are incomparable (aRb). Strategies which are best at each criterion are rare and, therefore, usually most strategies are incomparable without additional information from the campaign ordering party. This information can include inter alia the weights expressing the relative importance of each criterion or preferences associated to each pairwise comparison of strategies when each criterion is considered on its own [41].

Multiple multi-criteria decision analysis methods have been invented in order to reduce the number of incomparabilities (R) in the decision graph between the considered viral marketing campaign strategies. The MCDA attempts to handle this task can be generally divided into two approaches, the so-called American and European MCDA schools. The former is based on aggregating all the decision-making problem criteria into a single criterion—a utility function. Such approach has the benefit of providing the possibility to produce a complete ranking of strategies with a precise score given to each one. However, such approach largely transforms the structure of the decision problem. On the other hand, Roy [42, 43] proposed to construct outranking relations by enriching the dominance relation between the alternatives where possible. In such an approach (European MCDA school), not all incomparabilities are eliminated, however, a reliable selection of the best alternative is possible.

Presently, the literature review allows to observe a number of approaches (MCDA methods) based on the above American and European approaches [44]. Discussions about the up-to-date MCDA methods can be found inter alia in [45, 46]. The AHP, TOPSIS and Electre methods are often indicated as popular and widely used in the problems of evaluation and ranking creation [47, 48]. The selection of the aggregation technique (the utilized MCDA method) may influence the quality of the obtained modeling results [49–52] and requires justification in the context of the modeling aspects adapted in the paper [53, 54].

When analyzing the characteristics of the data and the environment of the constructed MCDA model, it should be noted that the input data of the model has a quantitative character and is expressed on the cardinal scale. It was decided that the weights of the individual criteria should be taken into account and that they should be expressed using explicitly specified numerical values [55]. Thus, the result is expected to be expressed on a quantitative scale [53]. The modeling process also assumed the natural imprecision of the preference information, which in practice, in MCDA methods, takes the form of complex preferential functions (e.g. pseudo-criteria) [42]. Additionally, the construction and usage (decision problematic) of the model, should result in a ranking of variants (strategies) [43]. The obtained ranking is expected to provide a complete order of the strategies [56]. None of the popular MCDA methods (AHP, TOPSIS, Electre) meet all of the indicated requirements at once. Based on a set of 56 MCDA methods discussed in [45] and the MCDA taxonomy contained there, it is easy to show that these assumptions are fully implemented only by the PROMETHEE II method. Therefore, it was decided to use this method in the next steps of the data analysis.

PROMETHEE is a family of MCDA methods that use pairwise comparison and outranking flows to produce a ranking of best decision variants [57]. The weights expressing the relative importance of each criterion need to be specified by the decision maker. This is a complicated process based on the DM's priorities and perceptions. The actual values of the criteria weights can be freely selected by the campaign ordering party. Fortunately, multi-criteria decision analysis provides tools such as sensitivity and robustness analyses, which allow to verify the effects the chosen values have on the resulting rankings and to sequentially adjust the weights.

When the PROMETHEE methods are used for viral marketing campaign strategy selection, all strategies are compared pairwise. The preference for one alternative over another is studied under each criterion. For a small difference d in evaluations of the two compared strategies, a small preference P would be assigned to the better one, or no preference at all, if the difference is negligible. On the other hand, the larger the difference d between alternatives, the higher the preference P. The preference P between strategies a and b under criterion c_j is expressed as real numbers $P_j(a, b)$ and is in the range between 0 and 1. The actual preference value assigned depends on the preference function in the DM's brain. The authors of the PROMETHEE methods propose six preference functions to express the preference function of the DM: usual criterion, U-shape criterion, V-shape criterion, level criterion, V-shape with indifference criterion, Gaussian criterion (see Fig 2 and Table 1) [41, 43, 58]. The three variables presented in Table 1, i.e. p, q and s require an explanation. The p preference threshold is the smallest difference between two alternatives that would result in a full preference of one alternative over the other. The q indifference threshold is the largest difference between two alternatives that the DM considers negligible. Parameter s denotes the inflection point of the Gaussian preference function and should be selected as a value between q and p.

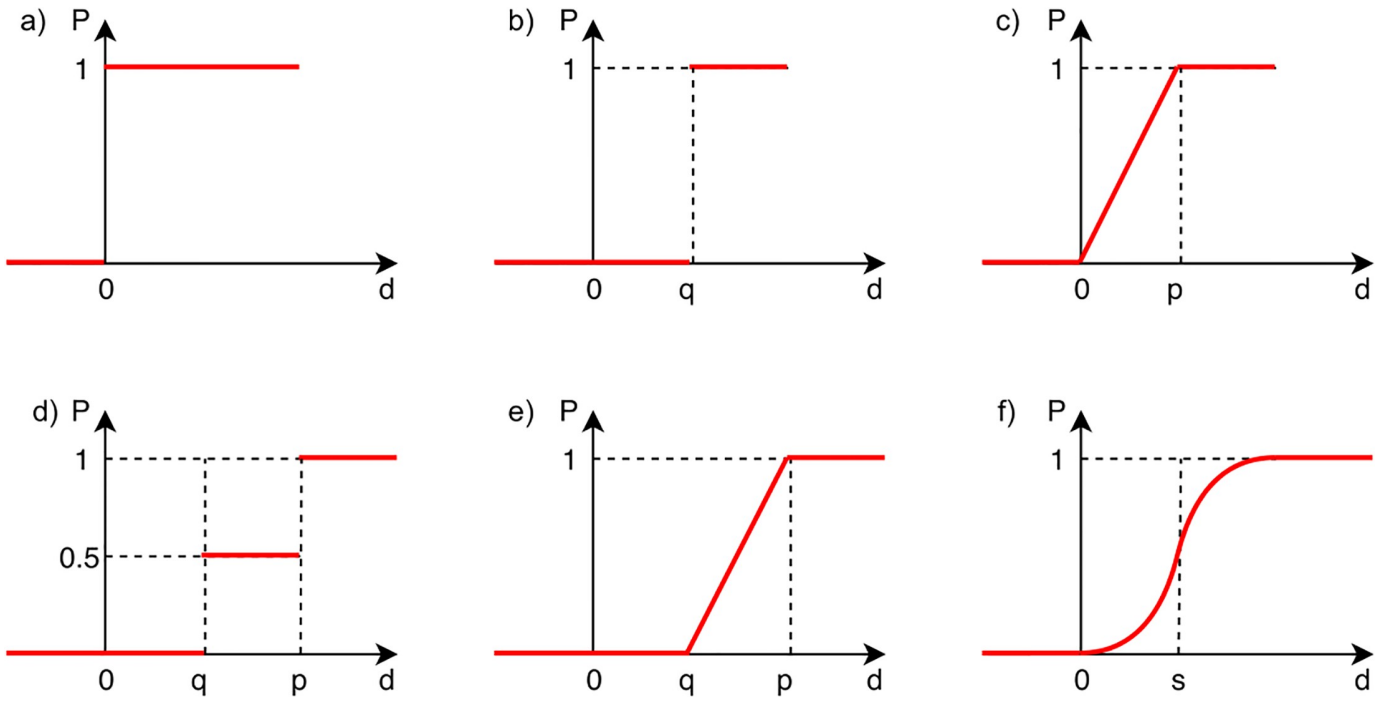


Fig 2. Visual representation of the six preference functions used in the PROMETHEE methods: a) usual criterion, b) U-shape criterion, c) V-shape criterion, d) level criterion, e) V-shape with indifference criterion, f) Gaussian criterion.

https://doi.org/10.1371/journal.pone.0209372.g002

Table 1. Formulae for the six preference functions used in the PROMETHEE methods [41].

a) usual	b) U-shape	c) V-shape
$P(d) = \begin{cases} 0 & d \leq 0 \\ 1 & d > 0 \end{cases}$	$P(d) = \begin{cases} 0 & d \leq q \\ 1 & d > q \end{cases}$	$P(d) = \begin{cases} 0 & d \leq 0 \\ \frac{d}{p} & 0 \leq d \leq p \\ 1 & d > p \end{cases}$
d) level	**e) V-shape with *q***	**f) Gaussian**
$P(d) = \begin{cases} 0 & d \leq 0 \\ \frac{1}{2} & q \leq d \leq p \\ 1 & d > p \end{cases}$	$P(d) = \begin{cases} 0 & d \leq 0 \\ \frac{d-q}{p-q} & q \leq d \leq p \\ 1 & d > p \end{cases}$	$P(d) = \begin{cases} 0 & d \leq 0 \\ 1 - e^{-\frac{d^2}{2s^2}} & d > 0 \end{cases}$

https://doi.org/10.1371/journal.pone.0209372.t001

It was decided in the process of preference modeling to use two of the six preference functions: V-shape and Gaussian. Their choice was dictated by the possibility of including the natural imprecision of the preference information of the decision maker into the modeling process [41, 43]. The form of the V-shape function, being the most complex structure of the preference function, is based on the concepts of strong and weak preferences, as well as indifference [43]. It is directly related to the concept of a pseudo-criterion, i.e. criterial function with the thresholds of indifference and weak and strong preference. This, in contrast to pre-criterion, quasi-criterion and real criterion, allows to effectively model the areas of information uncertainty of the decision-maker and, consequently, also to examine the robustness of the model in a wider scope, instead of simply generating a ranking [53]. Complementary, for comparative purposes, the Gaussian form was used as the second preferential function. In this research, contrary to the classic tasks of the MCDA methods, where only a small number of variants is being ordered [58], the obtained sample is fairly complex. It was assumed that the distribution of preferences is based on the Gauss function and, thus, reflects the normal distribution. Such representation allows to build, based on the given sample, a softer form of the preferential function (when compared to the linear form of the V-shape function). It is worth noting that this form of the preference function is based exclusively on the concepts of weak preference and indifference.

For each viral marketing campaign strategy, an aggregated preference index can be computed with the formula (3):

$$\begin{cases} \pi(a,b) = \sum_{j=1}^{k} P_j(a,b) w_j \\ \pi(b,a) = \sum_{j=1}^{k} P_j(b,a) w_j \end{cases} \tag{3}$$

where w_j denotes the weight assigned to the C_j criterion. $\pi(a, b) \sim 0$ implies a weak global preference, whereas $\pi(a, b) \sim 1$ implies a strong global preference of a over b.

The obtained indices are then used to calculate the positive and negative outranking flows with (4) and (5) [41]:

$$\phi^{+}(a) = \frac{1}{n-1} \sum_{x \in A} \pi(a,x) \tag{4}$$

$$\phi^{-}(a) = \frac{1}{n-1} \sum_{x \in A} \pi(x,a) \tag{5}$$

The $\phi^{+}(a)$ value indicates the *strength* of alternative a, i.e. how well it is outranking other alternatives. On the other hand, the $\phi^{-}(a)$ value represents how the alternative a is outranked by other alternatives, thus showing its *weakness*. The PROMETHEE I method uses the ϕ^{+} and

ϕ^- values to produce a partial ranking of the alternatives [41]. The usage of the PROMEHTEE II method, in turn, would allow to obtain the complete ranking of the campaign strategies based on the net outranking flow (6):

$$\phi(a) = \phi^+(a) - \phi^-(a) \tag{6}$$

For two strategies a and b, if $\phi(a) > \phi(b)$ then aPb. Contrarily, if $\phi(a) = \phi(b)$ then aIb.

If the criterion c_j is given the weight of 100% while the rest of the criteria is given the weight of 0%, a single criterion net flow for each strategy a is obtained: $\phi_j(a)$. When all single criterion net flows for all k criteria and n strategies are known, then all strategies can be represented as points in a k-dimensional space. Since the decision problems usually consist of more than two criteria, the n points from the k-dimensional space need to be projected to a plane.

The family of PROMETHEE methods comes with the GAIA visual modeling. A GAIA plane is a plane on which the alternatives and the criteria unit vectors are projected, for which as much information as possible is preserved after projection. The quantity of information that was preserved by the projection is denoted as δ. The GAIA plane can successfully support the decision problem analysis if $\delta >= 50\%$. The GAIA plane allows the analyst to learn the information about the criteria and alternatives in a decision problem [41]:

- the length of the criterion vector on the plane represents how discriminating the criterion is. The longer the vector, the more effect the criterion has on the final decision;
- the vectors of criteria similar in terms of preference are pointed in similar directions;
- the vectors of criteria conflicting in terms of preference are pointed in opposite directions;
- the vectors of criteria not related to each other in terms of preference are pointed orthogonally;
- alternatives with similar performance are grouped closely together;
- alternatives supported by a particular criterion are located in the direction pointed by this criterion's vector.

An example GAIA plane for a sample viral marketing strategy campaign selection problem with four possible strategies and three criteria (c_1—maximization of coverage, c_2—minimization of the number of iterations, c_3—minimization of the seeding fraction) is presented on Fig 3. The analysis of the example allows to observe that, since the c_1 and c_3 vectors are pointing in opposite directions, the preference for maximizing the coverage is in conflict with the preference for minimizing the seeding fraction. On the other hand, the c_2's orthogonal direction compared to c_1 and c_3 implies that the preference for minimization the number of iterations is not related to the preference for maximizing the coverage or minimizing the seeding fraction. The lengths of the criteria's vectors suggest that the maximization of coverage (c_1) is most discriminating in this decision problem. The alternatives a_1 and a_4 are grouped together at the location pointed by the vector c_1, therefore, it can be assumed that they both are similar to each other and mostly supported by the coverage maximization criterion. On the other hand, they are located in a direction opposite to the direction of the c_3 vector, which implies that they fail to use a small seeding fraction.

2.3 Conceptual framework and evaluation criteria

The selection of the best strategy for a viral marketing campaign in social networks is a complex decision-making problem based on multiple criteria. Moreover, running simulations on a real network is most often time consuming and sometimes impossible. Therefore, in the

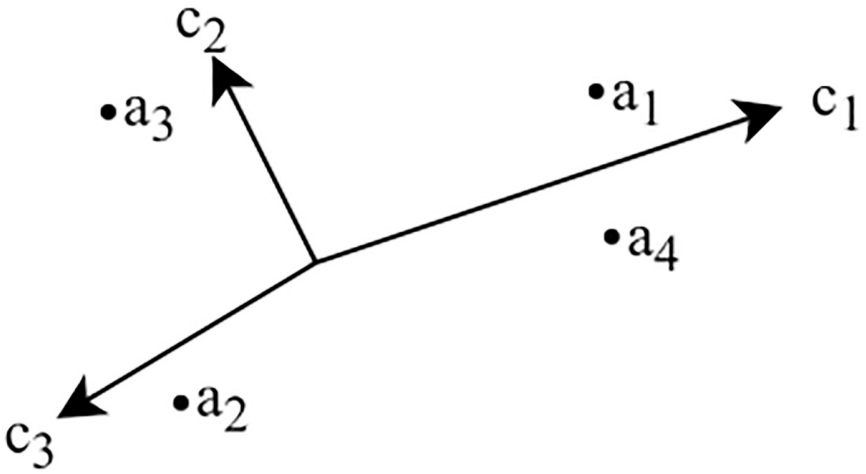


Fig 3. Example GAIA plane representing a viral marketing campaign strategy selection with four possible strategies and three criteria: c_1—Maximization of coverage, c_2—Minimization of the number of iterations, c_3—Minimization of the seeding fraction.

https://doi.org/10.1371/journal.pone.0209372.g003

presented approach (see Fig 4) the authors propose to run the planning process on a synthetic model, which has similar properties to the target real network, yet allows to perform multiple simulations resulting in data for the performance table as an input to the strategy evaluation process.

In order to obtain such a synthetic model resembling the target real network, the authors propose to generate a set of BA, ER and WS networks with various parameters and number of nodes equal to 10%, 20%, . . . 100% of the real network. Subsequently, KullbackÔÇôLeibler divergence [59] can be used to learn which of the generated networks is closest to the real one. Criteria affecting the complexity of simulations, such as number of nodes and edges, can also be considered. Additional criteria can also be added to the decision process, depending on the analyst's needs. When the performance table is created for all synthetic networks and all criteria, the most preferred one should be selected based on the gathered data with the use of MCDA methods.

The next element of the proposed framework is the decision model structuring process, during which the decision criteria for evaluating the possible campaign strategies are chosen. In the authors' approach, the criteria can be divided into two groups. The first group contains the input criteria for constructing a strategy—Par1, Par2, . . ., Par$m$. The second group contains the strategy performance evaluation criteria Eff1, Eff2, . . ., Eff$n$, whose values are based on achieved effects and can be obtained by simulating each strategy. Nevertheless, the proposed framework assumes the decision-maker's freedom in selection of the decision criteria and in grouping them into clusters, depending on the campaign ordering party requirements.

After the criteria have been selected and the decision model has been structured, the chosen synthetic network should be used to perform a complete set of simulations required to obtain the performance table containing the evaluations of each strategy. For simulations independent cascades model (IC) [26] was implemented. Information spreading process is initiated by a set of nodes activated by seeding. Spreading is based on propagation probability $PP(a, b)$ that node a activates node b in the step $t + 1$ under condition that node a was activated at time t by other node or was selected as a seed [60]. The main reason for selecting this model was a relatively small number of seeds needed to induce diffusion what can be important for small

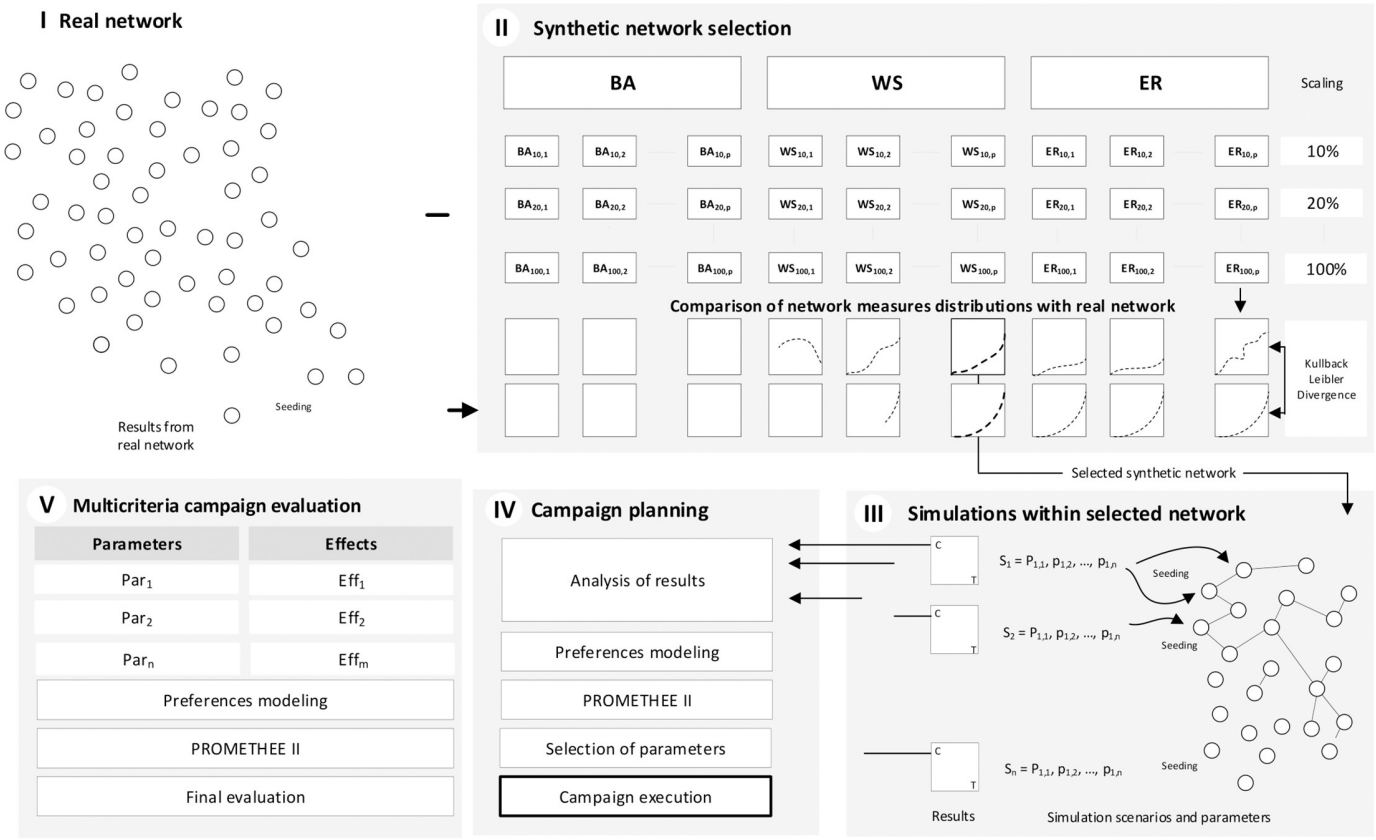


Fig 4. Proposed framework based on five stages: Analysis of real network (I), synthetic network selection process (II), simulations within synthetic network (III), campaign planing (IV) and multi-criteria campaign evaluation(V).

https://doi.org/10.1371/journal.pone.0209372.g004

networks. In linear threshold model (LT) small number of activated nodes would have no effect [26].

Subsequently, the obtained performance table is used to perform MCDA analysis of the possible strategies with the use of PROMETHEE II. The analysis includes the following aspects:

- generation of a complete ranking of the viral marketing campaign strategies, based on various preference functions;
- usage of the GAIA plane to verify how each criterion affects the strategy selection;
- performing sensitivity analysis to verify the stability intervals of the rankings of the leading campaign strategies.

It should be noted that during the analysis, the preference modeling step is repeated multiple times. The initial preference weights of the criteria can be subsequently modified to verify the robustness of the obtained strategy selection problem solution. Eventually, the analyst provides the recommendation which strategy, i.e. campaign parameters, should be used to run the campaign on the target real network. Last, but not least, the authors' proposed framework can also be used to monitor the results of the executed campaign, as well as to perform a multi-criteria evaluation of the campaign strategies on a real network.

Table 2. Mapping viral campaign characteristics into simulation model parameters and outputs.

Criteria	Type	Viral campaign	Model parameter	Symbol
Par1	Input	Number of initial customers	Percentage of network nodes activated during seeding process (seeding fraction)	SF
Par2	Input	Motivation to spread the content	Propagation probability	PP
Par3	Input	Initial customers selection	Computing node rankings and selection of nodes with highest propagation potential	R
Eff4	Output	Time required to reach assumed number of customers	Number of simulation steps	S
Eff5	Output	Number of reached customers	Number of activated nodes within the network	C

https://doi.org/10.1371/journal.pone.0209372.t002

The conceptual framework proposes a model in a generalized form with Par1,Par2,. . ., Par*m* criteria related to campaign settings treated as input variables and Eff1,Eff2,. . .,Eff*n* evaluation metrics related to campaigns results understood as output variables. The generalized model can be parametrized for campaigns with different mechanics, used strategies and goals. Like it was discussed in [19], criteria can vary across different campaigns, sectors, strategies and available resources. Therefore, a set of multi-criteria decisions needs to be made at the planning stage of the viral marketing campaign, based on the campaign objectives and available budget. It leads us to framework verification with the use of parameters mapping initial viral campaigns parameters and settings into the parameters of the simulation model. For the model validation we propose a set of criteria discussed in earlier studies and the specifics of the used simulation with the independent cascade model and the agent based approach presented in Table 2.

2.3.1 The number of initial customers and seeding fraction (Par1). The process of information spreading in viral marketing campaign in social networks is initialized with seeding the advertising content to a group of people (initial set of nodes). The fraction of nodes that are selected from the network for seeding can be adjusted according to the campaign objectives and it is affecting the dynamics and coverage of the process. The earlier research usually uses fixed ranges of seeding percentage as a parameter [7]. The activation of the initial seeds is recognized as the main cost of viral marketing campaigns [61]. The cost can be growing for highly influential nodes, while they attract high attention from marketers and the users from their direct and indirect connections. Other research focuses on minimization of the seed set to reduce the initial costs with probabilistic coverage guarantee [62]. The cost-effectiveness can decrease when more nodes are added to the seeding [63]. If too many users are targeted, an overexposure effect takes place [64]. While the activation of large fraction of network nodes as the seed set can result in high number of nodes reached within the network, it requires high activation costs. The goal can be to use the smallest possible seed set delivering satisfactory results [65] [62]. To include the above factors, we define criterion Par1 representing the fraction of nodes used as the initial seeds denoted as seeding fraction (SF) within the simulation model.

2.3.2 Spreading the content and propagation probability (Par2). In order to motivate the network members to pass the information further, some financial investments need to be made. As a result, the propagation probability is directly related to the campaign costs. From the practical point of view, the propagation probability can be increased with coupons and other incentives [66]. Authors discuss the role of incentives for increasing the camping dynamics and the costs of incentives is related to the degree of the target nodes [67]. The proposed approach minimizes the cost while guarantees the number of reached users. One of the strategies is enforcing the propagation dynamics without the use of additional seeds and users with high centrality measures which are expensive to reach [61]. Activation of early adopters and

increasing their propagation probabilities may require higher incentives [68]. Multi-scale incentives can be used for users from different target groups to further boost the diffusion rate [69]. The top influential nodes, such as a popular user, may require more incentives to be recruited as a Seed [70]. To generalize the aforementioned factors, we use Par2 as the main result of the increased motivation and propagation probability (PP) during the simulations.

2.3.3 Selection of initial customers and nodes ranking method (Par3). The nodes for seeding are selected based on their ranks computed from various centrality measures, such as degree, betweenness, closeness or eigenvector centrality. Each centrality measure requires some level of effort, indirectly related to a third kind of cost. Intuitively, if the seeding fraction is high and the network members are motivated to increase their propagation probability, the process of information diffusion should execute dynamically and achieve high network coverage. However, the budget for the campaign can be limited. Moreover, the aim of the campaign might not be to achieve high coverage very fast, but to keep the campaign slowly crawling for a longer amount of time. Computational cost of choosing seeds was analyzed earlier in relation to greedy algorithm [71]. Another study discusses computational costs and propose upper-bound estimation based algorithm to accelerate the computing speed [72]. Authors with the same approximation ratio like greedy algorithm [26]. The authors of [73] emphasize that the earlier approaches use impractical assumptions that any seed user can be acquired with the same cost and the same is the benefit obtained when influencing each user. Study [73] proposes cost-aware targeted viral marketing focused mainly on of selecting a node. Costs may represent the degree of difficulty with which people accept specific information [74]. From the perspective of network analysis, the centrality measures like page rank can represent costs because they are usually proportional to social influence. They can be used for mapping the corresponding cost values to all users in a given social network. Positive correlation between degree centrality and the success of viral marketing is observed [7]. To include above factors within the model, we assume different costs for different rankings methods. Nodes selection costs are represented by parameter Par3 within the model. For example they are lower for degree and higher for betweenness computations during simulations.

2.3.4 Campaign duration and number of simulation steps (Eff4). While Par1-Par3 are related to the model inputs and key factors affecting campaign performance like number of initial customers, budges, incentives and other forms of customer motivation, the proposed approach assumes monitoring of the campaign effects and assigning to them the preferences of the decision maker. The campaign cognitive goals can be based on reach, awareness and knowledge, behavioral goals are represented by number of actions and rate at which creatives are transferred [19]. From the perspective of the decision maker, the time when the assumed number of messages is received can be crucial. One of the goals can be minimizing the time in which assumed coverage is achieved [75]. Other authors emphasize the velocity and the speed of transmission, persistence and mental barriers [20]. Another study minimizes the complete influence time with cost represented by a fuzzy variable [21]. The role of time was emphasized in terms of campaigns with limited time (eg. political campaigns) [76]. In the proposed model, the duration of the campaign is represented by evaluation criteria Eff4 and (in simulations) as the number of simulation steps until the process is finished.

2.3.5 Campaign coverage and the total number of activated nodes (Eff5). Another measured result is related to the network coverage and is represented by criterion five (Eff5). It is the most common effect taken into consideration. Most of research focuses on maximizing reach and number of infected nodes and is treated as the main goal of a campaign [18]. The ability to reach a large number of customers with limited advertising budget is the key feature of the viral marketing [2]. From the perspective of algorithms, total coverage is the key evaluation factor used for influence maximization problem and seed selection algorithms evaluation

[26]. It is represented by the number of activated nodes within the network used during simulations.

3 Planning a viral marketing campaign with the use of synthetic networks

The empirical study has been divided into two subsections—planning and evaluation. In the former, a substantially smaller synthetic network was chosen to facilitate the planning of a viral marketing campaign. In the latter, an evaluation of marketing strategies in a real network [77] was presented.

3.1 Synthetic network selection

In the empirical research, the authors used the proposed framework to plan a viral marketing campaign for a real network [77]. The real network is based on 7610 nodes, 15751 edges with average values of main metrics: total degree D = 4.14, closeness C = 0.0004, PageRank PR = 0.0001, EigenVector EV = 0.003, clustering coefficient CC = 0.49 and betweenness B = 13478.93. The degree distribution of the network is presented on the chart on Fig 5a. In order to approximate the real network, a set of 150 synthetic networks was generated. This set was built by combining three network models (BA, ER and WS) with the following parameters: percentage of nodes of the real network—10%, 20%, . . ., 100%, i.e. 761, 1522, . . ., 7610 nodes; out-degree parameters with values 1,2,3,4 and 5 for the BA networks; number of edges in graph equal to 100%, 200%, 300%, 400% and 500% for the ER networks and neighborhood within which the vertices of the lattice will be connected with values 1,2,3,4,5 with rewiring probability 0.5 for the WS networks.

Kullback-Leibler divergence (KLD) [59] was used to evaluate the similarity of the synthetic networks to the real network. The results are visually presented on Fig 6. The analysis of Fig 6 allows to observe that the ER and WS synthetic networks are moderately similar to the real

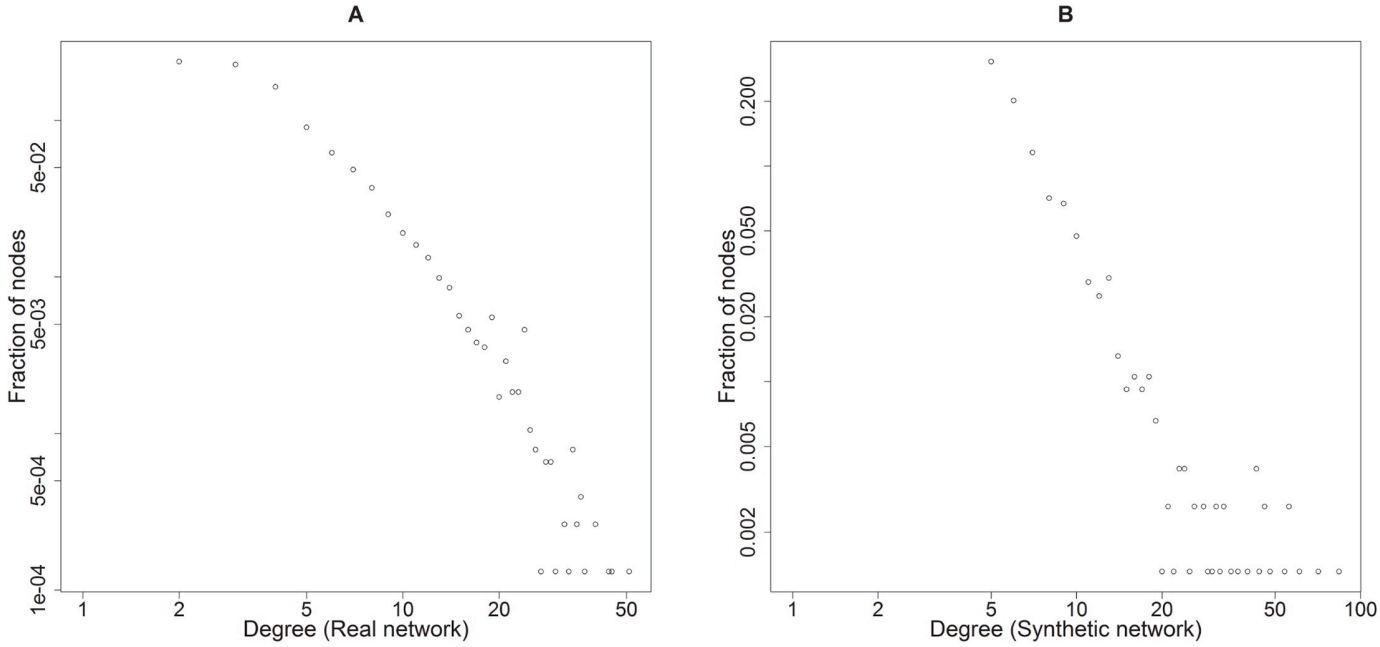


Fig 5. Degree distribution chart of A) the real network, B) the selected synthetic network.

https://doi.org/10.1371/journal.pone.0209372.g005

BA											
		1		2		3		4		5	
Nodes %	Nodes	Edges	KLD	Edges	KLD	Edges	KLD	Edges	KLD	Edges	KLD
10%	761	760	5.32E-03	1519	3.58E-03	2277	2.53E-03	3034	2.05E-03	3790	1.68E-03
20%	1522	1521	5.54E-03	3041	3.43E-03	4560	2.01E-03	6078	1.29E-03	7595	1.61E-03
30%	2283	2282	6.40E-03	4563	2.74E-03	6843	2.39E-03	9122	1.42E-03	11400	1.11E-03
40%	3044	3043	5.26E-03	6085	2.47E-03	9126	1.56E-03	12166	1.22E-03	15205	9.70E-04
50%	3805	3804	5.84E-03	7607	2.57E-03	11409	1.36E-03	15210	1.27E-03	19010	8.79E-04
60%	4566	4565	3.59E-03	9129	2.01E-03	13692	1.52E-03	18254	1.09E-03	22815	5.92E-04
70%	5327	5326	3.82E-03	10651	2.24E-03	15975	1.27E-03	21298	8.09E-04	26620	5.38E-04
80%	6088	6087	5.57E-03	12173	1.97E-03	18258	9.13E-04	24342	8.08E-04	30425	5.74E-04
90%	6849	6848	5.02E-03	13695	2.55E-03	20541	1.19E-03	27386	8.45E-04	34230	6.39E-04
100%	7610	7609	4.65E-03	15217	1.28E-03	22824	9.89E-04	30430	6.49E-04	38035	5.07E-04
ER											
		1		2		3		4		5	
Nodes %	Nodes	Edges	KLD	Edges	KLD	Edges	KLD	Edges	KLD	Edges	KLD
10%	761	761	3.84E-03	1522	3.13E-03	2283	2.87E-03	3044	2.63E-03	3805	2.50E-03
20%	1522	1522	3.91E-03	3044	3.17E-03	4566	2.75E-03	6088	2.62E-03	7610	2.49E-03
30%	2283	2283	3.81E-03	4566	3.13E-03	6849	2.79E-03	9132	2.61E-03	11415	2.48E-03
40%	3044	3044	3.87E-03	6088	3.12E-03	9132	2.80E-03	12176	2.60E-03	15220	2.49E-03
50%	3805	3805	3.79E-03	7610	3.13E-03	11415	2.80E-03	15220	2.59E-03	19025	2.49E-03
60%	4566	4566	3.86E-03	9132	3.09E-03	13698	2.81E-03	18264	2.61E-03	22830	2.48E-03
70%	5327	5327	3.86E-03	10654	3.13E-03	15981	2.79E-03	21308	2.62E-03	26635	2.48E-03
80%	6088	6088	3.83E-03	12176	3.11E-03	18264	2.80E-03	24352	2.61E-03	30440	2.49E-03
90%	6849	6849	3.82E-03	13698	3.07E-03	20547	2.79E-03	27396	2.62E-03	34245	2.48E-03
100%	7610	7610	3.85E-03	15220	3.12E-03	22830	2.80E-03	30440	2.60E-03	38050	2.49E-03
WS											
		1		2		3		4		5	
Nodes %	Nodes	Edges	KLD	Edges	KLD	Edges	KLD	Edges	KLD	Edges	KLD
10%	761	761	3.99E-03	1522	3.17E-03	2283	2.77E-03	3044	2.63E-03	3805	2.53E-03
20%	1522	1522	3.80E-03	3044	3.07E-03	4566	2.79E-03	6088	2.63E-03	7610	2.49E-03
30%	2283	2283	3.84E-03	4566	3.07E-03	6849	2.81E-03	9132	2.61E-03	11415	2.50E-03
40%	3044	3044	3.86E-03	6088	3.10E-03	9132	2.79E-03	12176	2.60E-03	15220	2.50E-03
50%	3805	3805	3.83E-03	7610	3.13E-03	11415	2.82E-03	15220	2.61E-03	19025	2.51E-03
60%	4566	4566	3.83E-03	9132	3.17E-03	13698	2.79E-03	18264	2.61E-03	22830	2.48E-03
70%	5327	5327	3.83E-03	10654	3.12E-03	15981	2.79E-03	21308	2.60E-03	26635	2.49E-03
80%	6088	6088	3.84E-03	12176	3.12E-03	18264	2.76E-03	24352	2.61E-03	30440	2.49E-03
90%	6849	6849	3.81E-03	13698	3.12E-03	20547	2.78E-03	27396	2.59E-03	34245	2.48E-03
100%	7610	7610	3.81E-03	15220	3.10E-03	22830	2.77E-03	30440	2.62E-03	38050	2.49E-03

Fig 6. Visual representation of the 150 synthetic networks used to approximate the [77] real network.

https://doi.org/10.1371/journal.pone.0209372.g006

network, regardless of the size of the network or parameters selected. On the other hand, in case of the BA networks the similarity to the real network depends on the number of nodes and parameters chosen. The more nodes and edges, the closer its degree distribution is to the degree distribution of the real network. The lowest KLD value was observed for the BA network with 7610 nodes and 38035 edges. However, selection of such a network provides little computational benefit compared to the original real network. Therefore, the actual network for the campaign planning was selected with the use of MCDA analysis of all the 150 potential synthetic networks based on the following criteria: K1—number of nodes, K2—number of edges and K3—KLD value; with the preference of the minimum values for all criteria and equal weights of all the criteria. As the result of the analysis, the BA network containing 10% nodes of the original network (761) and 3034 edges with network metrics degree D = 7.97, closeness C = 0.3211, PageRank PR = 0.0013, EigenVector EV = 0.086, clustering coefficient CC = 0.040 and betweenness B = 812.03 was selected. The degree distribution of the selected synthetic network is presented on the chart on Fig 5b.

3.2 Overview of the simulations in the synthetic network

In order to ascertain repeatability of the results regardless of the simulated parameters, ten simulation scenarios were generated, in which for each node a random value from the range of $< 0, 1 >$ was assigned. This value was later used in the simulations to decide if the particular node passes the information through (the drawn value was smaller than the simulated propagation probability) or if the propagation stops (the drawn value was higher than the simulated propagation probability).

During the simulation stage, a total of 400 sets of parameters was tested, as a Cartesian product of the following parameter values:

- Par1—0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10
- Par2—0.01, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90
- Par3—degree [0.0060], betweenness [0.0110], closeness [0.0085], eigenvector centrality [0.0090]—the values assigned to each measure were obtained based on the actual time of ranking generation based on each measure.

Each simulation was repeated for all 10 scenarios, thus resulting in 4000 simulation runs. After each simulation run, the iteration of the last infection, as well as the achieved coverage was registered. Their averaged values were then saved as the empirically measured performance values of the Eff4 and Eff5 criteria.

3.3 PROMETHEE II analysis

After the simulations finished, the output from subsection 3.2 was used to create performance tables for the PROMETHEE II analysis. Initially, a V-shape preference function was used to model the comparison preferences, with the indifference threshold $q = 0$ (no uncertainty taken into consideration) and preference threshold p equal to the standard deviation value for each criterion Par1-Eff5. The preference direction for the cost criteria Par1-Par3 were minimized and for the dynamics and coverage criteria Eff4-Eff5 maximized. Initially, every criterion was assumed to be equally important and, therefore, the weights of all criteria were set to 1 (see Table 3a).

The first 10 strategies from the ranking obtained with the PROMETHEE II method are presented in Table 4 and on Fig 7a. It can be observed, that all strategies from the list were based on the fastest (and therefore cheapest) degree measure. The leading alternatives A9 and A13,

Table 3. PROMETHEE II parameters for the synthetic network.

	Par1	Par2	Par3	Eff4	Eff5
	Statistics				
Minimum	0.01	0.01	0.0010	1	1.05%
Maximum	0.1	0.9	0.0110	10.40	100%
Average	0.06	0.45	0.0086	5.19	75.91%
Standard Dev.	0.03	0.29	0.0019	2.09	32.76%
a)	**V-shape, q = 0**				
Q: indifference	0	0	0	0	0
P: preference	0.03	0.29	0.0019	2.09	32.76%
weights	1	1	1	1	1
b)	**V-shape, q = 50% SD**				
Q: indifference	0.015	0.145	0.0009	1.045	16.38%
P: preference	0.03	0.29	0.0018	2.090	32.76%
weights	1	1	1	1	1
c)	**Gaussian**				
S: Gaussian	0.06	0.45	0.0086	5.19	75.91%
weights	1	1	1	1	1

https://doi.org/10.1371/journal.pone.0209372.t003

having a difference between their ϕ values equal to only 0.003 are very similar. Both use the smallest possible seeding factor of 0.01, whereas the propagation probability is equal to 0.2 and 0.3 for A9 and A13 respectively. As a result of the strategy A9, the campaign took averagely 10.4 iterations and covered 58.37% of the network, whereas for the strategy A13, the campaign was more dynamic (averagely 7.9 iterations) and covered more network (averagely 81.13%). The A17 strategy, ranked third, also uses the seeding factor of 0.01, but the propagation probability was increased to 0.4. It can be observed, that while the process averagely took 7 iterations, the coverage increased intensely to the level of 91.83%. The A5 strategy, ranked 4, uses the computationally cheapest parameters—seeding factor and average propagation probability set to 0.01 and the degree measure used. While the propagation process averagely lasted long, i.e. 9.7 iterations, the obtained coverage was merely 16.89% on average. It can be observed, that the alternatives A17 and A57, ranked 3 and 7 respectively, obtained the same coverage, with equal propagation probability and very similar dynamics of the process. However, in case of A17 the seeding fraction was equal to 0.01 and in case of A57 it was twice as much, i.e. 0.02, for which fact the latter was penalized in the overall ranking. The observation of Fig 7a allows to observe that for the best 10 strategies, the coverage grew along with the propagation probability, but that caused shortening of the process due to its high dynamics.

3.4 GAIA analysis

The basic PROMETHEE II analysis was followed by the GAIA analysis, which allows to study the relations between criteria, as well as shows which criteria support which strategies. A set of GAIA planes for the PROMETHEE decision problem specified in subsection 3.3 is presented on Fig 8. Fig 8a represents the decision problem with individual criteria and all strategies visible, whereas on Fig 8b the strategies were hidden for better visibility of the criteria vectors. A $\delta = 61.9\%$ quality of the projection was obtained for this GAIA plane. The analysis of Fig 8a allows to observe a very unnatural distribution of the points on the chart. As a matter of fact, this synthetic arrangement of points results from the strategies' simulative origin and creates a comprehensive grid of possible evaluations of the strategies.

Table 4. Results of the PROMETHEE II method analysis on the synthetic network: a) V-shape preference, b) V-shape preference with indifference threshold, c) Gaussian preference.

Action	ϕ	ϕ^+	ϕ^-	Rank	SF	PP	Measure	Last Iter.	Coverage
a)	**V-shape, q = 0**								
A9	0.5106	0.6656	0.155	1	0.01	0.2	degree	10.4	58.37%
A13	0.5103	0.6205	0.1102	2	0.01	0.3	degree	7.9	81.13%
A17	0.4833	0.5832	0.0999	3	0.01	0.4	degree	7	91.83%
A5	0.4762	0.6489	0.1727	4	0.01	0.1	degree	9.7	16.89%
A53	0.4758	0.5943	0.1185	5	0.02	0.3	degree	7.7	81.13%
A49	0.4736	0.6365	0.1629	6	0.02	0.2	degree	9.3	58.52%
A57	0.447	0.5566	0.1096	7	0.02	0.4	degree	6.8	91.83%
A89	0.4361	0.613	0.1768	8	0.03	0.2	degree	9	58.83%
A21	0.4226	0.5433	0.1207	9	0.01	0.5	degree	6.2	96.95%
A45	0.4225	0.6054	0.183	10	0.02	0.1	degree	8.2	20.18%
b)	**V-shape, q = 50% SD**								
A13	0.5266	0.5686	0.042	1	0.01	0.3	degree	7.9	81.13%
A9	0.5005	0.6349	0.1344	2	0.01	0.2	degree	10.4	58.37%
A53	0.4982	0.5415	0.0433	3	0.02	0.3	degree	7.7	81.13%
A17	0.4745	0.5324	0.0579	4	0.01	0.4	degree	7	91.83%
A49	0.4686	0.6029	0.1343	5	0.02	0.2	degree	9.3	58.52%
A5	0.4508	0.6112	0.1604	6	0.01	0.1	degree	9.7	16.89%
A93	0.4482	0.5019	0.0538	7	0.03	0.3	degree	7.1	81.13%
A57	0.4473	0.5072	0.0599	8	0.02	0.4	degree	6.8	91.83%
A89	0.437	0.5782	0.1411	9	0.03	0.2	degree	9	58.83%
A133	0.4081	0.4819	0.0738	10	0.04	0.3	degree	7.1	81.13%
c)	**Gaussian**								
A9	0.1919	0.2103	0.0184	1	0.01	0.2	degree	10.4	58.37%
A11	0.1792	0.1996	0.0204	2	0.01	0.2	closeness	10.4	58.37%
A12	0.1734	0.1948	0.0214	3	0.01	0.2	ev	10.2	58.37%
A10	0.1679	0.1975	0.0296	4	0.01	0.2	betweenness	10.4	58.37%
A49	0.1554	0.174	0.0186	5	0.02	0.2	degree	9.3	58.52%
A52	0.1503	0.1721	0.0217	6	0.02	0.2	ev	9.8	58.37%
A13	0.1486	0.1581	0.0096	7	0.01	0.3	degree	7.9	81.13%
A5	0.1447	0.2064	0.0617	8	0.01	0.1	degree	9.7	16.89%
A51	0.1426	0.1633	0.0207	9	0.02	0.2	closeness	9.3	58.52%
A16	0.1372	0.1497	0.0125	10	0.01	0.3	ev	8.1	81.13%

https://doi.org/10.1371/journal.pone.0209372.t004

It can be noticed from Fig 8 that the lengths of all criteria vectors is similar, which confirms their similar importance in the evaluation. If one of the vectors was significantly longer, it would mean that the related criterion is more discriminating. The layout of the vectors' directions allow to note that the cost criterion of the average propagation probability (Par2) is in strong conflict with the average coverage criterion (Eff5) in terms of preference. This confirms an intuitive thesis that the preference for reducing the cost of motivating the members of the social network for passing the seeded information further is in conflict with the preference for maximizing the achieved network coverage. On the other hand, all the cost criteria Par1-Par3 are perpendicular to each other, which means they are generally not related in terms of preference. Last, but not least, the vector representing criterion Eff4 (average last infection iteration) is slightly angled in the direction of both the the vectors representing criterion Par1 (seeding fraction cost) and Par2 (average propagation probability). This means, that the preference for

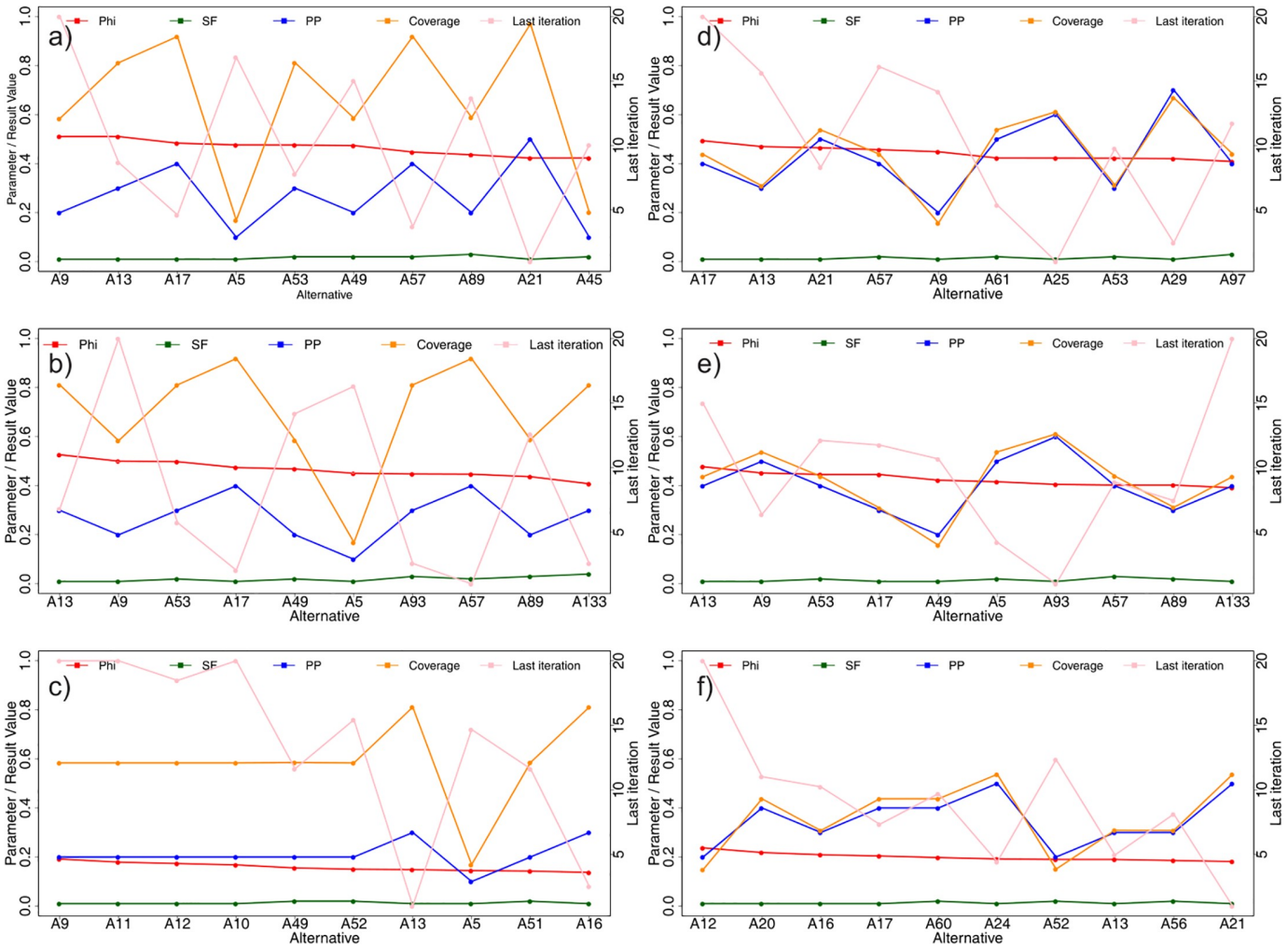


Fig 7. Visual representation of the 10 best strategies in PROMETHEE II rankings for the synthetic (a-c) and real (d-f) networks, based on V-shape preference function with no indifference threshold (a,d), V-shape preference function with indifference threshold (b,e) and Gaussian preference function (c-f).

https://doi.org/10.1371/journal.pone.0209372.g007

increasing the duration of the campaign may result in selecting similar strategies to the ones in case of the preference for minimizing the costs related to the seeding fraction or the average propagation probability.

One of the advantages of the PROMETHEE methods is their ability to aggregate data into groups and clusters. Fig 8c provides the GAIA analysis for a scenario where the three cost criteria Par1-Par3 were aggregated into a single Cost group, Eff4 into Dynamics group and Eff5 into Coverage group. The δ for Fig 8c is very high (90.9%), proving this projection to be very reliable. The analysis of this figure allows to confirm the strong conflict between the Coverage maximization group preference with the Cost minimization preference, however, there is no clear relation between the preferences for the Dynamics and Coverage maximization groups.

3.5 Sensitivity analysis

In the scenario analyzed in subsection 3.3, the weights of all criteria were equally set to 1. However, with such a dense grid of alternatives as presented on Fig 8a, it is easy to anticipate that if

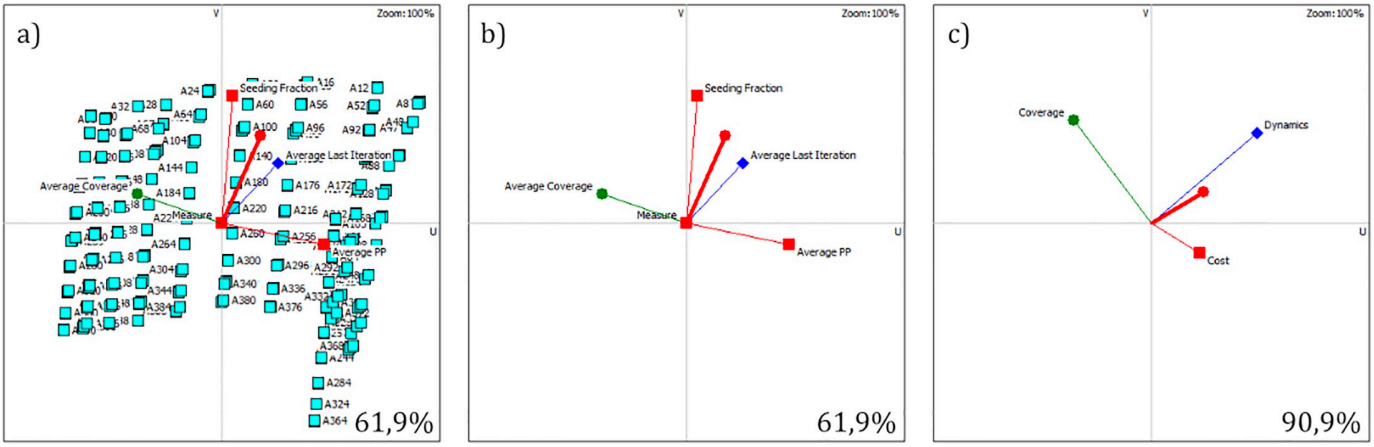


Fig 8. Synthetic network's GAIA analysis of individual criteria with visible (a) and hidden (b) strategies. GAIA analysis of grouped criteria (c).

https://doi.org/10.1371/journal.pone.0209372.g008

these preference weights were to change, the ranking of the best strategies would change. For this very reason, MCDA methods provide a tool called sensitivity analysis, which allows to verify the stability of the ranking and to learn how the positions of the alternatives would change if the change in preferences would occur. The results of the performed sensitivity analysis are presented in Tables 5 and 6 for grouped and individual criteria respectively.

In case of the grouped criteria (Table 5), the initial weights were 60% for the Cost group and 20% each for the Dynamics and Coverage groups. A very wide stability interval, equal to 80.10% can be observed for the Dynamics group, which means that the weight of this criterion can be largely increased and the leading alternative would not change its position. On the other hand, if the weight of the Dynamics group dropped by as little as 0.10%, a change in the ranking leader would occur. Much narrower stability interval is observed for the Cost and

Table 5. Stability intervals for criteria groups in the PROMETHEE II ranking with V-shape preference function with no indifference threshold for the synthetic network.

Group	Min Weight	Max Weight	Interval
Cost	59.82%	75.15%	15.33%
Dynamics	19.90%	100.00%	80.10%
Coverage	10.14%	20.06%	9.92%

https://doi.org/10.1371/journal.pone.0209372.t005

Table 6. Stability intervals for individual criteria in the PROMETHEE II ranking with V-shape preference function with no indifference threshold for the synthetic network.

Ranks	1			2			3		
Criterion	Min	Max	Interval	Min	Max	Interval	Min	Max	Interval
Par1	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	15.24%	100.00%	84.76%
Par2	19.88%	33.51%	13.63%	19.88%	26.70%	6.82%	19.88%	20.99%	1.11%
Par3	1.08%	100.00%	98.92%	1.12%	100.00%	98.88%	4.99%	100.00%	95.01%
Eff4	19.90%	100.00%	80.10%	19.90%	31.17%	11.27%	19.90%	21.39%	1.49%
Eff5	10.14%	20.06%	9.92%	16.21%	20.06%	3.85%	19.43%	20.06%	0.63%

https://doi.org/10.1371/journal.pone.0209372.t006

Coverage criteria groups, equal to 15.33% and 9.92% respectively. There is possibility to increase the Cost's weight or decrease the Coverage's weight slightly without the change of the ranking leader.

If the individual criteria Par1-Eff5 are taken into consideration, wide stability intervals can be observed for the first rank, and, therefore, the ranking stability was also performed for ranks 2 and 3. No changes in the weight of Par1 can cause a change in the rank of the first and second alternative. Only if the Par1 weight drops to below 15.24%, the strategy on rank 3 would be replaced. Similarly, the stability interval for the first three ranks of the Par3 criterion is equal to 95.01%. In case of the Eff4 and Eff5 criteria, the stability intervals for the first rank are the same as in Table 5. For the two leading ranks, the stability interval drops from 80.10% and 9.92% to 11.27% and 3.85% for Eff4 and Eff5 respectively. To sum up, the sensitivity analysis allows to notice that the criteria Par2, Eff4 and Eff5 are most discriminating to the final rankings of the viral advertising strategies for the selected synthetic network.

3.6 Uncertainty analysis

The strategies' evaluations in subsections 3.3 and 3.5 were based on certain data, i.e. on situations that the analyst was always able to specify their preference of one strategy over another regarding to individual criteria. However, such differentiation might not always be possible, especially if the difference between the criteria evaluation values are negligibly small. Therefore, in the subsequent section of the analysis, an uncertainty analysis was performed with the use of the PROMETHEE II method. The evaluation model from subsection 3.3 was modified to use the V-shape preference function with indifference area. Therefore, apart from the p threshold, which remained unchanged, the values of the q indifference threshold were set to 50% of the standard deviation value for each criterion Par1-Eff5.

The results of the PROMETHEE II analysis with uncertainty taken into account are presented in Table 4b. The analysis of the results allow to observe that strategy A13 overranked the previously leading strategy A9, whereas strategy A17, previously ranked 3, obtained position 4, while position 3 was taken by the strategy 57, previously ranked 5. Strategy A49 advanced from rank 6 to rank 5, whereas the rank of strategy A5 was reduced from 4 to 6. The rank of the strategies A57 and A89 was reduced by one, i.e. from 7 and 8 to 8 and 9 respectively. Two strategies A93 and A133, previously outside the set of the top ten strategies, advanced to ranks 7 and 10 respectively, when uncertainty was taken into consideration. The 10 leading strategies are presented on the chart on Fig 7b.

3.7 Gaussian preference function

In the final step of the PROMETHEE II analysis, the sharp V-shaped preference function was replaced with a much softer Gaussian preference function, with its s parameter equal to the mean value of each criterion Par1-Eff5. The obtained ranking varies substantially from the ones obtained in subsections 3.3 and 3.6. The first visible difference is that only four of the strategies from the original ranking remained in the top ten positions of the new ranking: A9, A13, A5 and A49 on positions 1, 7, 8 and 5 respectively (see Table 4c and Fig 7c). The remaining ranks were were distributed between strategies based on different centrality measures: closeness (A11, A51), eigenvector centrality (A12, A52, A16) and betweenness (A10). The sensitivity analysis of the newly obtained ranking is presented in Table 7. The analysis of the table allows to observe that Par1 is the least and Eff5 is the most discriminating criterion to the final ranking.

Table 7. Stability intervals for individual criteria in the PROMETHEE II ranking with Gaussian preference function.

Ranks	1			2			3		
Criterion	Min	Max	Interval	Min	Max	Interval	Min	Max	Interval
Par1	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%
Par2	0.00%	47.86%	47.86%	0.00%	42.45%	42.45%	0.00%	39.60%	39.60%
Par3	0.00%	100.00%	100.00%	0.00%	41.76%	41.76%	9.55%	35.88%	26.33%
Eff4	1.33%	100.00%	98.67%	7.66%	100.00%	92.34%	9.23%	37.17%	27.94%
Eff5	1.82%	43.84%	42.02%	7.49%	38.45%	30.96%	9.86%	35.64%	25.78%

https://doi.org/10.1371/journal.pone.0209372.t007

4 Evaluation of viral marketing campaign strategies in a real network

4.1 Overview of the simulations in the real network

In the second stage of the empirical research, the proposed framework was used to evaluate the viral marketing campaign strategies within a real network [77]. Similarly to the study in subsection 3.2, ten simulation scenarios were generated for the network, in order to ascertain the repeatability of the results regardless of the simulated parameters, as well as 400 various sets of parameters for the criteria Par1-Par3 as in subsection 3.2 were tested in the simulations. In case of the real network, the values of the Par3 criterion were as follows: degree [0.0200], betweenness [3.6900], closeness [2.1200], eigenvector centrality [0.030]. The output of the simulations was used to obtain the average performance values for Eff4 and Eff5 criteria for each of the 400 sets of parameters.

4.2 PROMETHEE II analysis

The results of simulations from subsection 4.1 were used to build the performance table for the PROMETHEE II analysis. As in section 3.3, V-shape preference function was used for modeling the comparison preferences, with the indifference threshold $q = 0$ and the preference threshold p equal to the standard deviation value for each criterion Par1-Eff5. The direction of the preference functions were as in subsection 3.3, to allow comparison of the results on the synthetic network and the real network (see Table 8a). The ten strategies which ranked highest are presented in Table 9a and Fig 7d.

The analysis of Table 9a allows to observe that similarly to the strategies obtained on the synthetic network, all 10 best strategies are based on the cheapest ranking measure, i.e. degree. The leading strategy A17 is based on a low seeding fraction (0.1), and mediocre propagation probability (0.4), which leads to very long process (16.2 iterations), but mediocre coverage (43.74%). The strategies A13 and A21 obtained a very close ϕ values, 0.4694 and 0.4645 respectively. The analysis of 7 shows that whilst having the same seeding fraction (0.1), they differ in the propagation probability and obtained duration and coverage. The higher-ranked strategy A13 uses lower propagation probability (0.3 compared to 0.5) and lasts longer (averagely 15.3 iterations compared to 13.8 iterations), but results in much lower coverage (30.92% compared to 53.73%). A further analysis of Fig 7d allows to observe that for the 10 leading strategies, increasing the propagation probability results in the increase of the coverage and in the reduction of the count of the propagation process iterations.

It can be observed that six of the strategies, i.e. A17, A13, A21, A57, A9 and A53 also occurred on the top 10 positions of the ranking obtained from the synthetic network in subsection 3.3. A further comparison of the results allows to note that the both rankings of strategies are highly correlated, with the Pearson correlation coefficient equal to 0.7589. The high

Table 8. PROMETHEE II parameters for the real network.

	Par1	Par2	Par3	Eff4	Eff5
	Statistics				
Minimum	0.01	0.01	0.02	1	1.00%
Maximum	0.1	0.9	3.69	21	74.79%
Average	0.06	0.45	1.4650	9.48	44.53%
Standard Dev.	0.03	0.29	1.5433	3.80	24.14%
a)	**V-shape, q = 0**				
Q: indifference	0	0	0	0	0
P: preference	0.03	0.29	1.5433	3.80	24.14%
weights	1	1	1	1	1
b)	**V-shape, q = 50% SD**				
Q: indifference	0.015	0.145	0.0009	1.90	12.07%
P: preference	0.03	0.29	0.0018	3.80	24.14%
weights	1	1	1	1	1
c)	**Gaussian**				
S: Gaussian	0.06	0.45	1.4650	9.48	44.53%
weights	1	1	1	1	1

https://doi.org/10.1371/journal.pone.0209372.t008

correlation can be visually confirmed on the chart on Fig 9. The chart shows the positions of each strategy obtained in the ranking based on the synthetic network (x axis) and on the real network (y axis). The closer the strategy is plotted to the diagonal line on the chart, the more similar was the rank of the strategy in each ranking. The analysis of the figure clearly shows a similarity of the evaluations of strategies in both rankings.

4.3 GAIA analysis

The basic PROMETHEE II analysis was followed by the GAIA analysis. A set of GAIA planes for the PROMETHEE decision problem specified in subsection 4.2 is presented on Fig 8. Fig 10a represents the decision problem with individual criteria and all strategies visible, whereas on Fig 10b the strategies were hidden for better clarity of the criteria vectors' analysis. A δ = 69.1% quality of the projection was obtained for this GAIA plane. The analysis of Fig 10a allows to observe that although the grid structure of the strategies similar to the one from subsection 3.4 is still noticeable, much more randomness can be observed, especially in the II and III quadrants, i.e. where the Eff4 and Eff5 preference vectors point to. This higher randomness level in the grid results from the fact that here the values of the the Eff4 and Eff5 criteria are taken from the empirical measurement based on a real network, as opposed to the synthetic network in subsection 3.4.

The analysis of Fig 10b shows that again the preference for the Par2 criterion (average propagation probability) is in strong conflict with the preference for the Eff5 criterion (average coverage). On the other hand, the vectors for criteria Par1, Par3 and Eff4 are pointing similar directions which indicates similarity in preference of these three criteria. In contrast to what was observed in subsection 3.4, the vectors of Par1 and Par3 criteria (seeding fraction and ranking generation measure) are no longer perpendicular to each other. Instead, they point the same direction, thus demonstrating a similarity in preference of reducing the cost of both these criteria. This can be caused by a considerably higher value of the standard deviation for the values of the criterion Par3 for the real network compared to the synthetic network. Moreover, it is significant to keep in mind that the action of generating the rankings of network nodes before seeding is strongly related to the action of seeding limited fraction of the best nodes from such obtained ranking.

Table 9. Results of the PROMETHEE II method analysis on the real network: a) V-shape preference, b) V-shape preference with indifference threshold, c) Gaussian preference.

Action	ϕ	ϕ^+	ϕ^-	Rank	SF	PP	Measure	Last Iter.	Coverage
a)	V-shape, q = 0								
A17	0.493	0.6396	0.1466	1	0.01	0.4	degree	16.2	43.74%
A13	0.4694	0.6273	0.1579	2	0.01	0.3	degree	15.3	30.92%
A21	0.4645	0.6033	0.1387	3	0.01	0.5	degree	13.8	53.73%
A57	0.4567	0.6109	0.1542	4	0.02	0.4	degree	15.4	43.84%
A9	0.4483	0.6123	0.1641	5	0.01	0.2	degree	15	15.75%
A61	0.4227	0.5715	0.1488	6	0.02	0.5	degree	13.2	53.82%
A25	0.4224	0.5671	0.1446	7	0.01	0.6	degree	12.3	61.16%
A53	0.4218	0.5897	0.1679	8	0.02	0.3	degree	14.1	31.21%
A29	0.4202	0.5678	0.1476	9	0.01	0.7	degree	12.6	66.82%
A97	0.4085	0.5783	0.1699	10	0.03	0.4	degree	14.5	43.94%
b)	V-shape, q = 50% SD								
A17	0.478	0.5942	0.1162	1	0.01	0.4	degree	16.2	43.74%
A21	0.452	0.5481	0.0961	2	0.01	0.5	degree	13.8	53.73%
A57	0.4463	0.5625	0.1162	3	0.02	0.4	degree	15.4	43.84%
A13	0.446	0.5774	0.1314	4	0.01	0.3	degree	15.3	30.92%
A9	0.4227	0.5607	0.138	5	0.01	0.2	degree	15	15.75%
A61	0.4164	0.5142	0.0978	6	0.02	0.5	degree	13.2	53.82%
A25	0.4057	0.5055	0.0999	7	0.01	0.6	degree	12.3	61.16%
A97	0.403	0.5267	0.1237	8	0.03	0.4	degree	14.5	43.94%
A53	0.4029	0.5355	0.1325	9	0.02	0.3	degree	14.1	31.21%
A20	0.3922	0.5585	0.1663	10	0.01	0.4	ev	17.6	43.73%
c)	Gaussian								
A12	0.2372	0.2934	0.0562	1	0.01	0.2	ev	21	14.87%
A20	0.2182	0.2432	0.025	2	0.01	0.4	ev	17.6	43.73%
A16	0.2092	0.2433	0.0341	3	0.01	0.3	ev	17.3	30.78%
A17	0.2041	0.2291	0.025	4	0.01	0.4	degree	16.2	43.74%
A60	0.1978	0.2231	0.0253	5	0.02	0.4	ev	17.1	43.74%
A24	0.1922	0.2188	0.0267	6	0.01	0.5	ev	15.1	53.73%
A52	0.1902	0.2464	0.0562	7	0.02	0.2	ev	18.1	15.08%
A13	0.19	0.2241	0.0341	8	0.01	0.3	degree	15.3	30.92%
A56	0.1859	0.2204	0.0345	9	0.02	0.3	ev	16.5	30.81%
A21	0.1816	0.2086	0.027	10	0.01	0.5	degree	13.8	53.73%

https://doi.org/10.1371/journal.pone.0209372.t009

When the criteria Par1-Eff5 are aggregated into three groups again, i.e. Cost, Dynamics and Coverage (see Fig 10c with $\delta = 94.0\%$), a very similar relation of the groups to the one obtained in subsection 3.4 can be observed (compare with Fig 8c). There is a strong conflict between the Cost minimization and Coverage maximization criteria, but the Coverage and Dynamics maximization criteria are not highly related in terms of preference. This time, however, a very minute similarity between the preference for the Dynamics maximization and Cost minimization criteria can be observed.

4.4 Sensitivity analysis

Similar to subsection 3.5, a sensitivity analysis was performed also for the real network. The results of the performed analysis are presented in Tables 10 and 11 for grouped and individual criteria respectively.

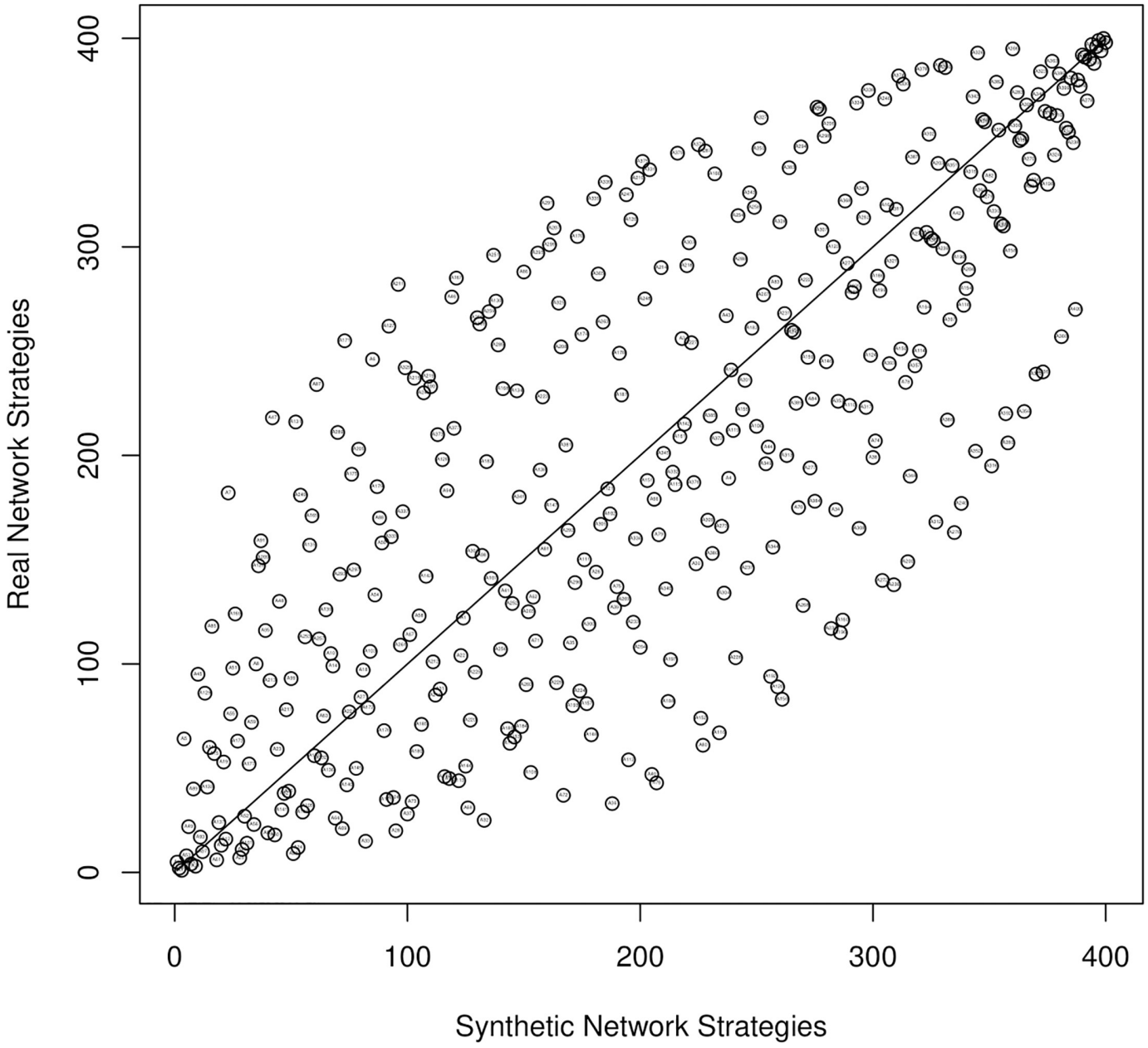


Fig 9. Comparison of the ranks of strategies obtained based on a synthetic and a real network [77].

https://doi.org/10.1371/journal.pone.0209372.g009

In case of the grouped criteria (Table 10), the initial weights were 60% for the Cost group and 20% each for the Dynamics and Coverage groups. The widest stability interval, equal to 55.99% can be observed for the Dynamics group, which means that the weight of this criterion can be decreased by 13% or increased by 42.99% and the leading alternative would not change its position. A narrower stability interval can be observed for the Cost criteria group, equal to 32.80%. The weight of this group can be increased by 9.97% or decreased by 22.83% without the change of the leader strategy. For the Coverage criteria groups, the stability interval is

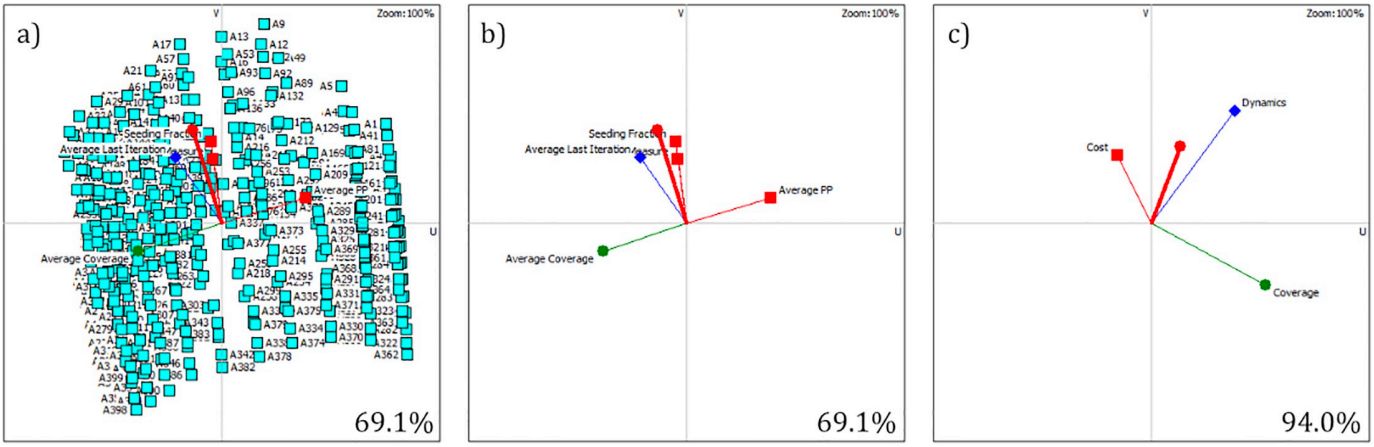


Fig 10. Real network's GAIA analysis of individual criteria with visible (a) and hidden (b) strategies. GAIA analysis of grouped criteria (c).

https://doi.org/10.1371/journal.pone.0209372.g010

equal to 14.99% and its weight can be reduced to 12.86% or increased to 27.85% without a change of the leader strategy. When compared to the results obtained for the strategies based on the synthetic network (see Table 5), a 0.9292 correlation coefficient is obtained for the intervals. However, the Dynamics cluster interval is narrower, yet the Cost and Coverage cluster intervals are wider.

If the individual criteria Par1-Eff5 are taken into consideration, wide stability intervals can be observed for the first rank, and, therefore, the ranking stability was also performed for ranks 2 and 3. No changes in the weight of Par1 can cause a change in the rank of the first strategy, however, if the Par1 weight drops to below 11.58%, the strategy on rank 2 would be replaced. Similarly, the stability interval for the first three ranks of the Par3 criterion is equal to 91.01%. In case of the Eff4 and Eff5 criteria, the stability intervals for the first rank are the

Table 10. Stability intervals for criteria groups in the PROMETHEE II ranking with V-shape preference function with no indifference threshold for the real network.

Group	Min Weight	Max Weight	Interval
Cost	37.17%	69.97%	32.80%
Dynamics	7.00%	62.99%	55.99%
Coverage	12.86%	27.85%	14.99%

https://doi.org/10.1371/journal.pone.0209372.t010

Table 11. Stability intervals for individual criteria in the PROMETHEE II ranking with V-shape preference function with no indifference threshold for the real network.

Ranks	1			2			3		
Criterion	Min	Max	Interval	Min	Max	Interval	Min	Max	Interval
Par1	0.00%	100.00%	100.00%	11.58%	100.00%	88.42%	15.03%	100.00%	84.97%
Par2	8.98%	27.97%	18.99%	19.01%	27.52%	8.51%	19.01%	22.09%	3.08%
Par3	2.69%	100.00%	97.31%	7.96%	100.00%	92.04%	8.99%	100.00%	91.01%
Eff4	7.00%	62.99%	55.99%	17.22%	49.18%	31.96%	17.22%	23.88%	6.66%
Eff5	12.86%	27.85%	14.99%	13.60%	20.74%	7.14%	18.36%	20.74%	2.38%

https://doi.org/10.1371/journal.pone.0209372.t011

same as in Table 10. If two leading ranks were considered instead of a single leading rank, the stability interval would drop from 55.99% and 14.99% to 49.18% and 20.74% for Eff4 and Eff5 respectively. Again, as in subsection 3.5, the sensitivity analysis allows to notice that the criteria Par2, Eff4 and Eff5 are most discriminating to the final rankings of the viral advertising strategies. When the results for the real network are compared with the results for the synthetic network, correlation indexes of 0.9217, 0.4928 and 0.3546 are obtained respectively for the stability intervals for up to 1, up to 2 and up to 3 leading strategies.

4.5 Uncertainty analysis

An uncertainty analysis similar to the one in subsection 3.6 was also performed for the real network. Again, the q indifference threshold for all criteria was set to 50% of their standard deviation values (see Table 8b). The results of the uncertainty analysis are presented in Table 9b.

The analysis of the results allow to observe that even when the uncertainty is taken into account, strategy A17 remains the leading one for the real network. The order of the subsequent strategies changes from A13, A21 and A57 to A21, A56 and A13 on ranks 2 to 4. Strategies A9, A61 and A25 remained unchanged on positions 5 to 7. Strategy A53 was degraded from rank 8 to rank 9, whereas the rank 8 was given to the previously 10th strategy A97. When the uncertainty was taken into consideration, strategy A20, previously outside of the set of the top ten strategies, obtained rank 10. The ten leading strategies are presented on the chart on Fig 7e.

4.6 Gaussian preference function

In the final step of the analysis, the sharp V-shaped preference function was replaced with the Gaussian preference function, with its s parameter equal to the mean value of each criterion Par1-Eff5. In contrast to the rankings from subsections 4.2 and 4.5, the obtained ranking contains mostly strategies based on the eigenvector centrality measure. The only three strategies based on the degree measure are A17, A13 and A21 (ranks 4, 8 and 10 respectively). The analysis of Fig 7f shows that the highest-appraised strategy A12 is based on a small value of seeding fraction (0.01 for Par1) and long duration (21 iterations in Eff4), however the obtained coverage is very small (14.87% for Eff5).

The results of the sensitivity analysis for the ranking is presented in Table 12. It can be observed, that the Par3 criterion is the least discriminating one when the Gaussian preference function is used and its weight can be vastly modified (up to 86.12%) without considerable changes in ranking. However, if the weight of the Eff5 criterion grew by 5.17%, the leading strategy in the ranking would change.

Table 12. Stability intervals for individual criteria in the PROMETHEE II ranking with Gaussian preference function for the real network.

Ranks	1			2			3		
Criterion	Min	Max	Interval	Min	Max	Interval	Min	Max	Interval
Par1	0.00%	100.00%	100.00%	0.00%	100.00%	100.00%	5.85%	100.00%	94.15%
Par2	10.63%	59.54%	48.91%	10.63%	27.19%	16.56%	15.23%	27.19%	11.96%
Par3	0.00%	97.48%	97.48%	0.00%	94.34%	94.34%	0.00%	86.12%	86.12%
Eff4	11.05%	100.00%	98.95%	11.05%	60.75%	49.70%	11.92%	44.57%	32.65%
Eff5	0.00%	25.17%	25.17%	13.53%	25.17%	11.64%	13.53%	23.27%	9.74%

https://doi.org/10.1371/journal.pone.0209372.t012

5 Conclusions

Recently, marketers have put more and more efforts to perceive the positive user experience within online platforms. While intrusive marketing techniques resulted in increased interest in software focused on blocking advertising content marketers, a new demand to focus on searching for more sustainable solutions appeared. The overall number of contacted customers can be less important than the real interest in products and a proper specification of the campaign intensity. In the area of viral marketing and information spreading processes, the highest attention was put on increasing campaign coverage with the use of seeding methods and techniques increasing the propagation probability like incentives and other ways to motivate customers to spread the content.

The approach presented in this paper shows a framework based on multi-criteria decision support, targeted on planning and evaluation of marketing campaigns with different preferences and criteria taken into account. The results showed how multi-criteria evaluation of results can affect strategies, campaign parameters and allocated budgets. The presented approach makes it possible to perform an evaluation of different scenarios within simulated environment before the campaign within a real environment begins. The empirical study showed that the characteristics of information spreading processes within the network sample selected according to network measures' distributions are similar to those observed within a real network. Various scenarios can be tested without interaction with real environments.

During the empirical study, an example viral marketing campaign was planned for an actual real network. Based on the real network parameters, a corresponding synthetic model was selected. Preference modeling and a profound multi-objective decision analysis were performed, which resulted in the selection of the best strategy in the context of the previously modeled preferences.

The research has identified possible areas of improvement and future works. First of all, the decision support system utilized in the presented framework was based on a set of five criteria. This set can be expanded to provide more precise evaluations. Secondly, future works include a more detailed evaluation of the relations between the processes within real networks and theoretical structures generated with different parameters. Another direction can be a sampling of real networks and performing simulations on samples of real networks instead of theoretical network models.

Supporting information

S1 File. Microsoft Excel spreadsheet containing partial results used during the research.
(XLSX)

Acknowledgments

This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562.

Author Contributions

Conceptualization: Artur Karczmarczyk, Jarosław Jankowski, Jarosław Wątróbski.

Formal analysis: Artur Karczmarczyk, Jarosław Jankowski.

Investigation: Artur Karczmarczyk, Jarosław Jankowski, Jarosław Wątróbski.

Methodology: Jarosław Jankowski, Jarosław Wątróbski.

Supervision: Jarosław Jankowski.

Validation: Jarosław Jankowski.

Visualization: Artur Karczmarczyk.

Writing – original draft: Artur Karczmarczyk, Jarosław Jankowski.

Writing – review & editing: Jarosław Wątróbski.

References

1. Hanna R, Rohm A, Crittenden VL. WeÔÇÖre all connected: The power of the social media ecosystem. Business horizons. 2011; 54(3):265–273. https://doi.org/10.1016/j.bushor.2011.01.007
2. Watts DJ, Peretti J, Frumin M. Viral marketing for the real world. Harvard Business School Pub.; 2007.
3. Iribarren JL, Moro E. Impact of human activity patterns on the dynamics of information diffusion. Physical review letters. 2009; 103(3):038702. https://doi.org/10.1103/PhysRevLett.103.038702 PMID: 19659326
4. Berger J, Milkman KL. What makes online content viral? Journal of marketing research. 2012; 49(2):192–205. https://doi.org/10.1509/jmr.10.0353
5. Ho JY, Dempsey M. Viral marketing: Motivations to forward online content. Journal of Business research. 2010; 63(9-10):1000–1006. https://doi.org/10.1016/j.jbusres.2008.08.010
6. Zhang X, Han DD, Yang R, Zhang Z. UsersÔÇÖ participation and social influence during information spreading on Twitter. PloS one. 2017; 12(9):e0183290. https://doi.org/10.1371/journal.pone.0183290 PMID: 28902906
7. Hinz O, Skiera B, Barrot C, Becker JU. Seeding strategies for viral marketing: An empirical comparison. Journal of Marketing. 2011; 75(6):55–71. https://doi.org/10.1509/jm.10.0088
8. Liu-Thompkins Y. Seeding viral content: The role of message and network factors. Journal of Advertising Research. 2012; 52(4):465–478. https://doi.org/10.2501/JAR-52-4-465-478
9. Kandhway K, Kuri J. How to run a campaign: Optimal control of SIS and SIR information epidemics. Applied Mathematics and Computation. 2014; 231:79–92. https://doi.org/10.1016/j.amc.2013.12.164
10. Kiss C, Bichler M. Identification of influencers: measuring influence in customer networks. Decision Support Systems. 2008; 46(1):233–253. https://doi.org/10.1016/j.dss.2008.06.007
11. Nejad MG, Amini M, Babakus E. Success factors in product seeding: The role of homophily. Journal of Retailing. 2015; 91(1):68–88. https://doi.org/10.1016/j.jretai.2014.11.002
12. Bampo M, Ewing MT, Mather DR, Stewart D, Wallace M. The effects of the social structure of digital networks on viral marketing performance. Information systems research. 2008; 19(3):273–290. https://doi.org/10.1287/isre.1070.0152
13. Stieglitz S, Dang-Xuan L. Emotions and information diffusion in social media: sentiment of microblogs and sharing behavior. Journal of management information systems. 2013; 29(4):217–248. https://doi.org/10.2753/MIS0742-1222290408
14. Dobele A, Lindgreen A, Beverland M, Vanhamme J, Van Wijk R. Why pass on viral messages? Because they connect emotionally. Business Horizons. 2007; 50(4):291–304. https://doi.org/10.1016/j.bushor.2007.01.004
15. Camarero C, San José R. Social and attitudinal determinants of viral marketing dynamics. Computers in Human Behavior. 2011; 27(6):2292–2300. https://doi.org/10.1016/j.chb.2011.07.008
16. Salehi M, Sharma R, Marzolla M, Magnani M, Siyari P, Montesi D. Spreading processes in multilayer networks. IEEE Transactions on Network Science and Engineering. 2015; 2(2):65–83. https://doi.org/10.1109/TNSE.2015.2425961
17. Michalski R, Kajdanowicz T, Bródka P, Kazienko P. Seed selection for spread of influence in social networks: Temporal vs. static approach. New Generation Computing. 2014; 32(3-4):213–235. https://doi.org/10.1007/s00354-014-0402-9
18. Helm S. Viral marketing-establishing customer relationships by'word-of-mouse'. Electronic markets. 2000; 10(3):158–161.
19. Cruz D, Fill C. Evaluating viral marketing: isolating the key criteria. Marketing Intelligence & Planning. 2008; 26(7):743–758. https://doi.org/10.1108/02634500810916690
20. Welker CB. The paradigm of viral communication. Information Services & Use. 2002; 22(1):3–8. https://doi.org/10.3233/ISU-2002-22102

21. Ni Y, Shi Q, Wei Z. Optimizing influence diffusion in a social network with fuzzy costs for targeting nodes. Journal of Ambient Intelligence and Humanized Computing. 2017; 8(5):819–826. https://doi.org/10.1007/s12652-017-0552-y
22. Rogers EM. Diffusion of innovations. Simon and Schuster; 2010.
23. Pfitzner R, Garas A, Schweitzer F. Emotional Divergence Influences Information Spreading in Twitter. ICWSM. 2012; 12:2–5.
24. Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2010. p. 1029–1038.
25. Wasserman S, Faust K. Social network analysis: Methods and applications. vol. 8. Cambridge university press; 1994.
26. Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2003. p. 137–146.
27. Jankowski J, Bródka P, Kazienko P, Szymanski B, Michalski R, Kajdanowicz T. Balancing Speed and Coverage by Sequential Seeding in Complex Networks. arXiv preprint arXiv:160907526. 2016;.
28. He JL, Fu Y, Chen DB. A novel top-k strategy for influence maximization in complex networks with community structure. PloS one. 2015; 10(12):e0145283. https://doi.org/10.1371/journal.pone.0145283 PMID: 26682706
29. Jankowski J. Dynamic rankings for seed selection in complex networks: Balancing costs and coverage. Entropy. 2017; 19(4):170. https://doi.org/10.3390/e19040170
30. Zhang JX, Chen DB, Dong Q, Zhao ZD. Identifying a set of influential spreaders in complex networks. Scientific reports. 2016; 6:27823. https://doi.org/10.1038/srep27823 PMID: 27296252
31. Kitsak M, Gallos LK, Havlin S, Liljeros F, Muchnik L, Stanley HE, et al. Identification of influential spreaders in complex networks. Nature physics. 2010; 6(11):888. https://doi.org/10.1038/nphys1746
32. Sankar CP, Asharaf S, Kumar KS. Learning from bees: An approach for influence maximization on viral campaigns. PloS one. 2016; 11(12):e0168125. https://doi.org/10.1371/journal.pone.0168125 PMID: 27992472
33. Seeman L, Singer Y. Adaptive seeding in social networks. In: Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on. IEEE; 2013. p. 459–468.
34. Granell C, Gómez S, Arenas A. Competing spreading processes on multiplex networks: awareness and epidemics. Physical review E. 2014; 90(1):012808. https://doi.org/10.1103/PhysRevE.90.012808
35. Barabási AL, Albert R. Emergence of scaling in random networks. science. 1999; 286(5439):509–512. https://doi.org/10.1126/science.286.5439.509 PMID: 10521342
36. ERDdS P, R&WI A. On random graphs I. Publ Math Debrecen. 1959; 6:290–297.
37. Watts DJ, Strogatz SH. Collective dynamics of ÔÇśsmall-worldÔÇÖnetworks. nature. 1998; 393 (6684):440. https://doi.org/10.1038/30918 PMID: 9623998
38. Massaro E, Bagnoli F. Epidemic spreading and risk perception in multiplex networks: a self-organized percolation method. Physical Review E. 2014; 90(5):052817. https://doi.org/10.1103/PhysRevE.90.052817
39. Wei X, Chen S, Wu X, Feng J, Lu Ja. A unified framework of interplay between two spreading processes in multiplex networks. EPL (Europhysics Letters). 2016; 114(2):26006. https://doi.org/10.1209/0295-5075/114/26006
40. Wei X, Wu X, Chen S, Lu Ja, Chen G. Cooperative epidemic spreading on a two-layered interconnected network. SIAM Journal on Applied Dynamical Systems. 2018; 17(2):1503–1520. https://doi.org/10.1137/17M1134202
41. Brans JP, Mareschal B. PROMETHEE methods. In: Multiple criteria decision analysis: state of the art surveys. Springer; 2005. p. 163–186.
42. Roy B, Słowiński R. Questions guiding the choice of a multicriteria decision aiding method. EURO Journal on Decision Processes. 2013; 1(1-2):69–97. https://doi.org/10.1007/s40070-013-0004-7
43. Roy B. Paradigms and challenges. In: Multiple criteria decision analysis: state of the art surveys. Springer; 2005. p. 3–24.
44. Ceballos B, Lamata MT, Pelta DA. A comparative analysis of multi-criteria decision-making methods. Progress in Artificial Intelligence. 2016; 5(4):315–322. https://doi.org/10.1007/s13748-016-0093-1
45. Wątróbski J, Jankowski J, Ziemba P, Karczmarczyk A, Zioło M. Generalised framework for multi-criteria method selection. Omega. 2018;.

46. Mardani A, Jusoh A, Zavadskas EK. Fuzzy multiple criteria decision-making techniques and applications–Two decades review from 1994 to 2014. Expert systems with Applications. 2015; 42(8):4126–4148. https://doi.org/10.1016/j.eswa.2015.01.003
47. Celik M, Er ID. Fuzzy axiomatic design extension for managing model selection paradigm in decision science. Expert Systems with Applications. 2009; 36(3):6477–6484. https://doi.org/10.1016/j.eswa.2008.07.038
48. Kurka T, Blackwood D. Selection of MCA methods to support decision making for renewable energy developments. Renewable and Sustainable Energy Reviews. 2013; 27:225–233. https://doi.org/10.1016/j.rser.2013.07.001
49. Wang X, Triantaphyllou E. Ranking irregularities when evaluating alternatives by using some ELECTRE methods. Omega. 2008; 36(1):45–63. https://doi.org/10.1016/j.omega.2005.12.003
50. Peng Y, Wang G, Wang H. User preferences based software defect detection algorithms selection using MCDM. Information Sciences. 2012; 191:3–13. https://doi.org/10.1016/j.ins.2010.04.019
51. Chang YH, Yeh CH, Chang YW. A new method selection approach for fuzzy group multicriteria decision making. Applied Soft Computing. 2013; 13(4):2179–2187. https://doi.org/10.1016/j.asoc.2012.12.009
52. Kolios A, Mytilinou V, Lozano-Minguez E, Salonitis K. A comparative study of multiple-criteria decision-making methods under stochastic inputs. Energies. 2016; 9(7):566. https://doi.org/10.3390/en9070566
53. Guitouni A, Martel JM. Tentative guidelines to help choosing an appropriate MCDA method. European Journal of Operational Research. 1998; 109(2):501–521. https://doi.org/10.1016/S0377-2217(98)00073-3
54. Ulengin F, Topcu YI, Sahin SO. An artificial neural network approach to multicriteria model selection. In: Multiple Criteria Decision Making in the New Millennium. Springer; 2001. p. 101–110.
55. Vansnick JC. On the problem of weights in multiple criteria decision making (the noncompensatory approach). European Journal of Operational Research. 1986; 24(2):288–294. https://doi.org/10.1016/0377-2217(86)90051-2
56. Corrente S, Greco S, Słowiński R. Multiple criteria hierarchy process with ELECTRE and PROMETHEE. Omega. 2013; 41(5):820–846. https://doi.org/10.1016/j.omega.2012.10.009
57. Wątróbski J, Ziemba E, Karczmarczyk A, Jankowski J. An Index to Measure the Sustainable Information Society: The Polish Households Case. Sustainability. 2018; 10(9).
58. Roy B, Bouyssou D. Aide multicritère à la décision: méthodes et cas. Economica Paris; 1993.
59. Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Statist. 1951; 22(1):79–86. https://doi.org/10.1214/aoms/1177729694
60. Wang C, Chen W, Wang Y. Scalable influence maximization for independent cascade model in large-scale social networks. Data Mining and Knowledge Discovery. 2012; 25(3):545–576. https://doi.org/10.1007/s10618-012-0262-1
61. Mochalova A, Nanopoulos A. Non-intrusive Viral Marketing Based on Percolation Centrality. In: ECIS; 2015.
62. Zhang P, Chen W, Sun X, Wang Y, Zhang J. Minimizing seed set selection with probabilistic coverage guarantee in a social network. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2014. p. 1306–1315.
63. Dinh TN, Zhang H, Nguyen DT, Thai MT. Cost-effective viral marketing for time-critical campaigns in large-scale social networks. IEEE/ACM Transactions on Networking (ToN). 2014; 22(6):2001–2011. https://doi.org/10.1109/TNET.2013.2290714
64. Abebe R, Adamic L, Kleinberg J. Mitigating overexposure in viral marketing. arXiv preprint arXiv:170904123. 2017;.
65. Shakarian P, Paulo D. Large social networks can be targeted for viral marketing with small seed sets. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012). IEEE Computer Society; 2012. p. 1–8.
66. Wu W, Du DZ, et al. Coupon Advertising in Online Social Systems: Algorithms and Sampling Techniques. arXiv preprint arXiv:180206946. 2018;.
67. Kotnis B, Sunny A, Kuri J. Incentivized Campaigning in Social Networks. IEEE/ACM Transactions on Networking. 2017; 25(3):1621–1634. https://doi.org/10.1109/TNET.2016.2645281
68. Zhang B, Qian Z, Li W, Lu S. Pricing strategies for maximizing viral advertising in social networks. In: International conference on database systems for advanced applications. Springer; 2015. p. 418–434.
69. Doo M, Liu L. Probabilistic diffusion of social influence with incentives. IEEE Transactions on Services Computing. 2014; 7(3):387–400. https://doi.org/10.1109/TSC.2014.2310216
70. Tang J, Tang X, Yuan J. Profit maximization for viral marketing in online social networks. In: Network Protocols (ICNP), 2016 IEEE 24th International Conference on. IEEE; 2016. p. 1–10.

71. Long C, Wong RCW. Viral marketing for dedicated customers. Information Systems. 2014; 46:1–23. https://doi.org/10.1016/j.is.2014.05.003

72. Hu H, Wen Y, Feng S. Budget-efficient viral video distribution over online social networks: Mining topic-aware influential users. IEEE Transactions on Circuits and Systems for Video Technology. 2018; 28 (3):759–771. https://doi.org/10.1109/TCSVT.2016.2620152

73. Nguyen HT, Dinh TN, Thai MT. Cost-aware targeted viral marketing in billion-scale networks. In: INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, IEEE. IEEE; 2016. p. 1–9.

74. Yue W, WeiJing H, Lang Z, TengJiao W, DongQing Y. Influence maximization with limit cost in social network. SCIENCE CHINA-INFORMATION SCIENCES. 2013; 56(7).

75. Goyal A, Bonchi F, Lakshmanan LV, Venkatasubramanian S. On minimizing budget and time in influence propagation over social networks. Social network analysis and mining. 2013; 3(2):179–192. https://doi.org/10.1007/s13278-012-0062-z

76. Jankowski J, Szymanski BK, Kazienko P, Michalski R, Bródka P. Probing Limits of Information Spread with Sequential Seeding. Scientific reports. 2018; 8(1):13996. https://doi.org/10.1038/s41598-018-32081-2 PMID: 30228338

77. Newman ME. The structure of scientific collaboration networks. Proceedings of the national academy of sciences. 2001; 98(2):404–409. https://doi.org/10.1073/pnas.98.2.404

A2.

Wątróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., Zioło, M. (2019). Generalised framework for multi-criteria method selection. Omega, 86, 107-124.

Omega 86 (2019) 107–124

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Generalised framework for multi-criteria method selection☆

Jarosław Wątróbski[a,*], Jarosław Jankowski[b], Paweł Ziemba[a], Artur Karczmarczyk[b], Magdalena Zioło[a]

[a] Faculty of Economics and Management, University of Szczecin, Mickiewicza 64, 71-101 Szczecin, Poland
[b] Faculty of Computer Science and Information Systems, West Pomeranian University of Technology, Zolnierska 49, 71-210 Szczecin, Poland

ARTICLE INFO

Article history:
Received 10 September 2017
Accepted 12 July 2018
Available online 21 July 2018

Keywords:
Multi-criteria decision analysis
MCDA
Multi-criteria method selection
Incomplete decision problem description

ABSTRACT

Multi-Criteria Decision Analysis (MCDA) methods are widely used in various fields and disciplines. While most of the research has been focused on the development and improvement of new MCDA methods, relatively limited attention has been paid to their appropriate selection for the given decision problem. Their improper application decreases the quality of recommendations, as different MCDA methods deliver inconsistent results. The current paper presents a methodological and practical framework for selecting suitable MCDA methods for a particular decision situation. A set of 56 available MCDA methods was analysed and, based on that, a hierarchical set of methods' characteristics and the rule base were obtained. This analysis, rules and modelling of the uncertainty in the decision problem description allowed to build a framework supporting the selection of a MCDA method for a given decision-making situation. The practical studies indicate consistency between the methods recommended with the proposed approach and those used by the experts in reference cases. The results of the research also showed that the proposed approach can be used as a general framework for selecting an appropriate MCDA method for a given area of decision support, even in cases of data gaps in the decision-making problem description. The proposed framework was implemented within a web platform available for public use at www.mcda.it.

1. Introduction

The increasing complexity of economic and social systems results in an increase in the complexity of the related decision problems [1]. They concern, among others, political decisions [2], organization management [3,4], financial management [5,6] and marketing [7]. Many decision problems are characterized by large dimensionality [8], the occurrence of sources of uncertainty and risk factors [9]. It is important to reconcile the contradictory goals, make decisions with many criteria and strive for compromise solutions [10]. Policy makers face the complexity of decision situations and they require methods and systems that support the decision-making [11]. In response to these needs, many solutions dedicated to selected areas, as well as general-purpose methods have been developed [12]. In this context, multi-criteria decision analysis (MCDA) methods are widely used. Apart from the formal foundations, these methods are characterized by the possibilities of handling a multitude of conflicting goals, as well as different stakeholders within decision making process [13]. In the recent years, a dynamic development of MCDA methods has been observed [14]. However, they significantly differ in many dimensions such as complexity, the way in which preferences and evaluation criteria are represented, the type of data aggregation, the possibility of including uncertain data, and the availability of implementations in decision support systems or criteria compensation [15–17]. The extensive number of possible MCDA methods results in a problem with their proper selection and application in specific decision situations.

While various MCDA methods can be used for improving the quality of decisions, they often produce conflicting results when compared [18–22]. It is worth noticing that a decision-maker (DM) may reach different decisions even when applying the same weights of criteria and the criterial evaluations of variants. This fact has been confirmed in a number of publications, in which rankings of decision variants were examined with the use of different MCDA methods [1,19,23–26]. Such analytical methods also often fail to provide guidelines [16]. The question is, which

☆ This manuscript was processed by Associate Editor Prof. B. Lev.

* Corresponding author.
E-mail addresses: jaroslaw.watrobski@wneiz.pl (J. Wątróbski), jjankowski@wi.zut.edu.pl (J. Jankowski), pziemba@ajp.edu.pl (P. Ziemba), akarczmarczyk@wi.zut.edu.pl (A. Karczmarczyk), magdalena.ziolo@usz.edu.pl (M. Zioło).

https://doi.org/10.1016/j.omega.2018.07.004

decision support method should be used and which characteristics of the decision problems are affecting the selection of a method?

A significant research issue, which is not entirely solved yet, is to determine a method suitable for a given problem, since only a method which is correctly chosen allows to obtain a solution that is most satisfying for the DM [20] in the context of a given decision-making situation. This problem emerges when the decision maker is unable to obtain a detailed description of the decision-making situation [18,27]. The complexity, uniqueness, or the fact that the decision-making situations can occur simultaneously in a short period makes analysis of them challenging [20,28]. In such conditions, the DM faces a dilemma of either making the decision based on incomplete information, or not making it on time [29]. In consequence, it becomes necessary to use formal procedures and guidelines for selecting MCDA methods also in cases of partial lack of knowledge about the decision-making situation.

The literature review provides a vast range of works dealing with the MCDA method selection problem for a given decision-making problem. However, the range of these solutions is often limited to the few of the best-known MCDA methods [30,31] or to a single, arbitrarily selected, field of application [22,32]. The studied approaches often also require that a decision-maker knows in advance certain formal aspects of the problems. In reality, a decision-maker may find it difficult to define a priori all relevant details of a given decision situation. Unfortunately, there is a lack of approaches addressing this uncertainty.

As a result, the motivation of the current research was:

- to build a formal guideline for MCDA method selection, which is independent of the problem domain,
- to use an extensive set of available MCDA methods and their characteristics,
- to obtain high accuracy of recommendation of particular MCDA methods for a given decision-making situation,
- to address the lack of knowledge issue in the descriptions of the decision-making situations.

In this paper, a new approach for selecting an MCDA method is proposed. As the authors aimed to develop an approach independent of the area of usage, the proposed framework is based on determining a set of characteristics of the available MCDA methods. Furthermore, the authors endeavoured to address the knowledge gaps in decision-making situation description and, additionally, to analyse their influence on the process of the MCDA method selection. The authors' technical contribution is also provided in a form of a useful website-based tool for supporting the process of MCDA method selection.

According to the authors' best knowledge, this research is the first successful attempt to handle uncertainty in the decision-making situation description during MCDA methods selection process. The entire solution space was examined. Surprisingly, the results clearly show that even partial uncertainty in a selected aspect of the decision-making situation description does not significantly affect the contents of the recommended set of methods.

The practical confirmation of the proposed framework was based on scientific literature as a reliable source of expert knowledge and the fact that usually the decision makers, who are experts, have knowledge of which method should provide a sufficient solution to the problem [33]. Practical examples are positioned in the field of sustainable logistics and transport as the field with wide usage of MCDA methods [30,34]. Research confirmed that the recommendations for MCDA methods' usage delivered by the proposed framework are consistent with the methods used by the experts for solving specific problems.

The paper is organized as follows: Section 2 provides the MCDA methodology foundations followed by the definition of the research gap. In Section 3, a framework of multi-criteria method selection is provided. A discussion of the range in which uncertainty of the decision-making problem description affects the framework is also presented. An outline of an expert system supporting MCDA methods selection is also provided. In Section 4, an exemplary confirmation of the proposed framework in the area of sustainable transport and logistics is presented. The article concludes with a discussion of the achieved results and areas of further research.

2. Literature review

2.1. MCDA foundations

The MCDA methods' task is to support a decision-maker in choosing the most preferable variant from many possible options, taking into account a multitude of criteria characterizing acceptability of individual decision variants. The criteria can also grade the quality of the variants when all options are permissible and the problem is to choose the best one subjectively. In this case, subjectivity refers to the importance of individual criteria, as for each decision-maker some factors are typically more significant than others. Furthermore, the uncertainty and inaccuracy of data describing alternatives influence the subjectivity of evaluation [35].

Multi-criteria problems can be divided into continuous ones, such as multiple-criteria linear programming, and discrete ones, such as those solved by methods based on utility or value function and outranking methods [36]. The utility/value theory-based approach determines two types of relationships between variants: indifference ($a_i\ I\ a_j$) and preference ($a_i\ P\ a_j$) of one variant over another. Methods in this group leave out non-comparability of the decision variants and assume transitivity and completeness of preference [29]. Methods based on outranking relations often expand a set of basic preferential situations with the result that contains indifference of decision variants ($a_i\ I\ a_j$), weak preference - one variant over another ($a_i\ Q\ a_j$), the strict preference - one variant over another ($a_i\ P\ a_j$), and incomparability between data variations ($a_i\ R\ a_j$) [29]. The preferential situations can be combined in an "outranking" relation, which contains the situations of indifference as well as strict and weak preference ($a_i\ S\ a_j$) [37].

The preference scenarios in the outranking methods are related to the thresholds used in them (outranking methods). Indifference (q), preference (p) and veto (v) are the three kinds of thresholds [27]. The thresholds allow the recognition of the uncertainty of the evaluations by the preferences' gradation. Furthermore, in many outranking methods (e.g. ELECTRE III), the weak preference has the form of a linear function whose values, from the interval *[0, 1]*, increase when approaching the threshold p, and, as a result, the preferences are subject to a characteristic fuzzification. Moreover, the preference thresholds' usage determines the form of the preference criterion used in the MCDA method. When no thresholds are used, the MCDA method uses a so-called true-criterion. However, application of the indifference threshold only determines the use of a semi-criterion by the method, and application of the indifference and preference thresholds means that the method uses a pseudo-criterion [38].

Two basic operational approaches may be distinguished to aggregate performance of variants: (1) aggregation to a single criterion (American school), (2) aggregation by using the outranking relationship (European school) [37]. Moreover, mixed (indirect) approaches, which combine elements of American and European decision-making schools, are applied. The approach can be exemplified by a group of PCCA (Pairwise Criterion Comparison Approach) methods [39].

MCDA methods are also different depending on the nature and characteristics of the used data [28]. The nature of data is closely connected to the measurement scale. Data can be quantitative or qualitative and can be expressed in the cardinal (quantitative) or

ordinal (qualitative) scale [40]. What is more, the cardinal scale can be of interval or ratio (relative) type [35]. In case of a relative scale, the data is presented in relations to other data. For example, the weight of criterion g_1 can be expressed in relations to criterion g_2 (g_1 is 3 times more important than g_2) [57]. The characteristics of the data used refer to whether the data is certain or uncertain [41]. The certain data, which is also called deterministic, is expressed in a crisp form, whereas uncertain data (non-deterministic) is represented by some kind of distribution (continuous or discrete) [20,28]. New methods based on the fuzzy set theory make it possible to express uncertain data in a fuzzy form [41]. The data type refers to both the scale on which the criterion performance of the variants is presented, as well as to the weights of the criteria. A summary of individual MCDA methods and their basic properties is presented in Table 1 and Supplementary material – Section 1.

According to Roy, there are four stages in the decision-making process [37]: (I) defining an object of the decision and the set of potential decision variants A as well as the determination of the reference problematics on A; (II) analysing consequences and developing the consistent set of criteria C; (III) modelling comprehensive preferences and operationally aggregating performances; (IV) investigating and developing the recommendation, based on the results of stage III and the problem defined in Stage I. Roy argues that the stages are not serial. For instance, some elements of Stage I can require performing elements of Stage II. Similarly, the decision-making process cannot be simplified by eliminating individual stages. In Stage I, Roy [35] distinguishes four decision problematics: α - selection, β - sorting, γ - ranking, δ - description with formal representation presented in Supplementary material – Section 2.

In Stage III, the operational approach for a given decision problem should be selected. Stage IV, in particular, requires selecting the computational procedure (the MCDA method), depending on the decision issue and the decision-maker's operational approach [37]. Roy's model indicates that the selection of the MCDA method is a vital element of solving a decision problem [17]. Furthermore, to obtain a "good" solution to the problem, one needs to apply a properly selected method.

2.2. The problem of selection of a proper MCDA method

Even in the early study of [18], it was found that "the great diversity of MCDA procedures may be seen as a strong point, it can also be a weakness. Up to now, there has been no possibility of deciding whether one method makes more sense than another in a specific problem situation. A systematic axiomatic analysis of decision procedures and algorithms is yet to be carried out."

Roy [37] also indicated that the selection of the MCDA method is a vital element of solving a decision problem. When defining the operational approach, the author paid attention to the method selection problem within four stages in the decision-making process. Furthermore, to obtain a "good" solution to the problem, a decision-maker needs to apply an adequately selected method [17]. However, selecting a multi-criteria method only on the basis of the decision issue and operational approach seems to be too general, as the decision-maker can choose many methods to solve a given decision problem on such a basis. The issue is the multitude of MCDA methods and their diversity [33,42].

Decision-makers are often unable to fully justify their choice of the method which was applied to solve their decision situation [21]. The selection of a multi-criteria method is usually carried out arbitrarily and is motivated by the decision-maker's knowledge of a given method or availability of software supporting the method [10,43]. Similar issues are also levelled in relation to MCDA software selection. Decision-makers usually choose decision support software, which, they are familiar with [44]. On this account, it is not an MCDA method that is selected for a decision problem, but the decision problem is adjusted to a chosen multi-criteria method [20]. It is difficult to answer a question which method is most suitable to solve a given kind of a problem [18,19]. The selection of a proper MCDA method for a given decision situation is salient, since various methods can yield different results for the same problem [18–26]. The difference in results when applying various calculating procedures can be influenced by the following factors [19,45]: (a) various techniques use weights differently in their calculations; (b) algorithms differ in their approach to selecting the "best" solution; (c) many algorithms attempt to scale the objectives, which affects the weights already chosen; (d) some algorithms introduce additional parameters affecting the final recommendations.

The literature analysis shows several works dealing with the subject of multi-criteria method selection for a given decision problem. They can be categorized into those which, when selecting an MCDA method, were based on: benchmarking [1,19,25,26,46,47], multi-criteria methods (it was recognized that the issue of selecting an MCDA method is a multi-criteria problem) [48] as well as the informal [16,30,49] or formal [21,22,31,32,50,51,52] structuring of a problem or a decision situation. A summary of the up to date approaches to MCDA method selection is presented in Supplementary material – Section 3. The presented approaches are not without shortcomings. The benchmark-based approaches ignore that the solutions considered for decision-making are usually optimal in Pareto term. In fact, they do not allow them to choose the optimal MCDA method, but only compare the compliance of the solutions of each method. The multi-criteria approach places the problem of methods selection in loop as it requires the use of MCDA method [20,47]. In turn, the informal approach does not give clear and unambiguous guidance on the choice of the method of MCDA, applicable to the particular class of decision problem. The formal approaches are characterized by an accurate selection of the methods oscillating on the border of acceptability. The IDEA approach [21,22,32] achieves an accuracy of 63–73%, depending on the matched MCDA method. The range of discipline and methodical approaches used so far, often limiting them only to the analysed domain of MCDA methods usage, is also problematic.

The guidelines for the selection of the MCDA method may be redundant for some classes of decision problems. The degree of criteria compensation is essential for the problems in the field of sustainability [30], but for other classes of problems, such a guideline is unnecessary. When analysing the coverage of individual methodical approaches, it should be noted that the previously mentioned works on the MCDA method selection considered a comparatively limited set of methods. The highest number of 29 methods was examined in [32,51] considered the 24 methods, [21] - 22, [22] - 21, [25] - 18, [48] - 16, [19] - 8, [50] - 6, [30] - 5, [31] - 4, [52] - 3, [26] - 3, [1] - 3. The publications often failed to include the relatively new methods such as ANP, Vikor and fuzzy extensions of the top classical methods. As a result the motivation of the current research was to build a formal guideline and framework for MCDA method selection independent from the problem domain with the use of a complete set of available MCDA methods and their characteristics.

3. The proposed framework for MCDA method selection

3.1. Main assumptions

In this section, we propose a generalized framework for the selection of a suitable MCDA method for a particular decision situation. The conceptual framework is shown in Fig. 1. There are two separate components of the framework, the methodological and the practical elements. The first one is based on methodological

Table 1
Taxonomy of MCDA methods.

Method name	Available binary relations					Linear compensation effect			Type of aggregation			Type of preferential information				
	I	*P*	*Q*	*R*	*S*	No	Total	Partial	Single criterion	Outranking	Mixed	Deterministic	Cardinal	Non-deterministic	Ordinal	Fuzzy
AHP	1	1	0	0	0	0	0	1	1	0	0	1	1	1	0	0
ANP	1	1	0	0	0	0	0	1	1	0	0	1	1	1	0	0
ARGUS	1	0	0	1	1	0	0	1	0	1	0	1	0	1	1	0
COMET	1	1	0	0	0	0	1	0	1	0	0	1	1	1	1	1
ELECTRE I	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0
ELECTRE II	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0
ELECTRE III	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0
ELECTRE IS	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0
ELECTRE IV	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0
ELECTRE TRI	0	0	0	1	1	0	0	1	0	1	0	1	1	0	1	0
EVAMIX	1	1	0	0	0	0	0	1	1	0	0	1	1	0	1	0
Fuzzy AHP	1	1	0	0	0	0	0	1	1	0	0	1	1	1	0	1
Fuzzy ANP	1	1	0	0	0	0	0	1	1	0	0	1	1	1	0	1
Fuzzy methods of extracting the minimum and maximum values of the attribute	1	1	1	0	0	1	0	0	0	0	1	0	1	1	1	1
Fuzzy PROMETHEE I	1	1	0	1	0	0	0	1	0	1	0	1	1	1	1	1
Fuzzy PROMETHEE II	1	1	0	0	0	0	0	1	0	1	0	1	1	1	1	1
Fuzzy SAW	1	1	1	0	0	0	1	0	1	0	0	0	1	1	1	1
Fuzzy TOPSIS	1	1	0	0	0	0	1	0	1	0	0	1	1	1	0	1
Fuzzy VIKOR	1	1	0	0	0	0	1	0	1	0	0	1	1	1	0	1
IDRA	1	1	0	0	0	0	0	1	0	0	1	1	1	1	0	0
Lexicographic method	1	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0
MACBETH	1	1	0	0	0	0	0	1	1	0	0	1	1	1	1	0
MAPPAC	1	1	1	1	0	0	0	1	0	0	1	1	1	0	0	0
MAUT	1	1	0	0	0	0	0	1	1	0	0	0	1	1	0	0
MAVT	1	1	0	0	0	0	0	1	1	0	0	1	1	0	0	0
Maximax	1	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0
Maximin	1	1	0	0	0	1	0	0	1	0	0	1	1	0	1	0
Maximin fuzzy method	1	1	1	0	0	1	0	0	1	0	0	1	1	1	1	1
MELCHIOR	0	0	0	1	1	0	0	1	0	1	0	1	0	0	1	0
Methods of extracting the minimum and maximum values of the attribute	1	1	0	0	0	1	0	0	0	0	1	1	1	0	1	0
NAIADE I	0	0	0	1	1	0	0	1	0	1	0	1	1	1	1	1
NAIADE II	0	0	0	0	1	0	0	1	0	1	0	1	1	1	1	1
ORESTE	1	1	0	1	0	0	0	1	0	1	0	1	0	0	1	0
PACMAN	1	1	1	1	0	0	0	1	0	0	1	1	1	0	1	0
PAMSSEM I	0	0	0	1	1	0	0	1	0	1	0	1	1	1	1	1
PAMSSEM II	0	0	0	0	1	0	0	1	0	1	0	1	1	1	1	1
PRAGMA	1	1	0	1	0	0	0	1	0	0	1	1	1	0	0	0
PROMETHEE I	1	1	0	1	0	0	0	1	0	1	0	1	1	0	1	0
PROMETHEE II	1	1	0	0	0	0	0	1	0	1	0	1	1	0	1	0
QUALIFLEX	0	0	0	1	1	0	0	1	0	0	1	1	0	0	1	0
REGIME	0	0	0	1	1	0	0	1	0	1	0	1	0	0	1	0
Simple Additive Weighting (SAW)	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0
SMART	1	1	0	0	0	0	0	1	1	0	0	1	1	0	0	0
TACTIC	1	1	0	1	0	0	0	1	0	1	0	1	1	1	0	0
TOPSIS	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0
UTA	1	1	0	0	0	0	0	1	1	0	0	1	0	0	1	0
VIKOR	1	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0
DEMATEL	1	1	0	0	0	0	1	0	1	0	0	1	0	1	1	0
REMBRANDT	1	1	0	0	0	0	0	1	1	0	0	1	1	1	0	0

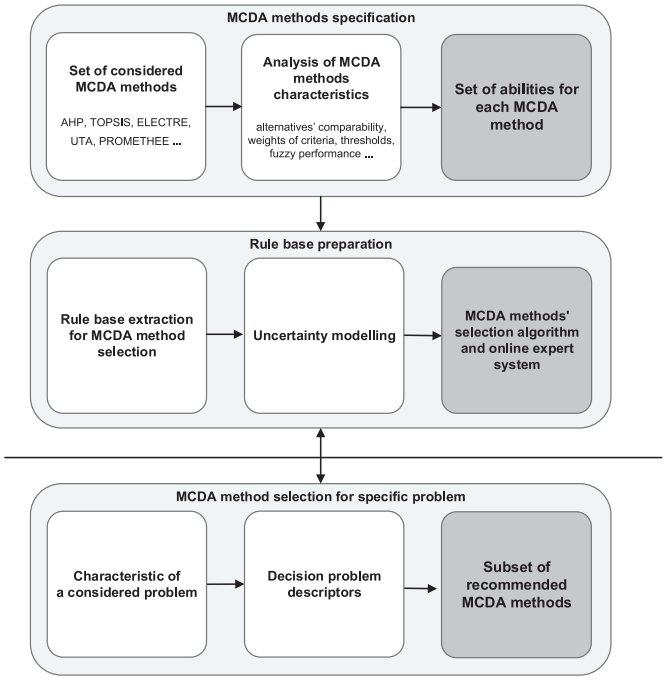


Fig. 1. Research procedure.

aspects and the rules' database generation. Both elements are required for the second part, i.e. practical verification when the descriptors of the considered problem are gathered and the subset of recommended methods is presented. The methodological aspects include creating a set of the considered MCDA methods and an analysis of their properties. The set of characteristics for each method is obtained and presented in Table 2. The subset of the recommended MCDA methods is obtained on the basis of the rule base (Table 3) and the characteristics of the problem considered. The validity of the proposed framework is reported in the following section.

Let *DP* be a specific multi-criteria decision problem. The classic approach to the decision problem allows presenting it in the form of a three-element set (*A, G, E*), where *A* defines a set of decision variants; *G* is a set of criteria, and *E* represents the efficiency of criterial performance, wherein $E=G(A)$ [20,35,51]. If sets *A* and *G* are presented as vectors, then set *E* is given as a matrix $E=A^T \cdot G$. Additionally, an aggregated performance of variants can be presented in short as *E(A)*. The weights of criteria W can be defined absolutely, e.g. g_1, or in respect to other criteria, e.g. g_1/g_2. In a similar way, the criteria performances of variants can be expressed, e.g. $g_1(a_1)$ or $g_1(a_1)/g_1(a_2)$. New MCDA methods often use the fuzzy set theory [14], which allows using uncertain data as trapezoidal fuzzy numbers, i.e. $\tilde{N}=(n_l, n_u, \alpha_F, \beta_F)$, or triangular fuzzy numbers, i.e. $\tilde{N}=(n, \alpha_F, \beta_F)$ [53-55].

The basis of the proposed framework is a set of 56 MCDA methods (or their combinations) and their characteristics containing nine descriptive properties of MCDA methods organized in a hierarchical form. It is worth noting that the decision-maker does not always have full knowledge of the given decision problem. Therefore, in certain situations, the DM is not able to fully define the descriptors (c) of the decision problem and, therefore, also his needs regarding the characteristics (m) of the MCDA method. For this reason, we propose a hierarchical structure of descriptors and characteristics, adapted to various levels of the definition of the DM's needs. Each level of hierarchy is deepening the accuracy of the description from the preceding level.

3.2. Proposed decision problem descriptors and MCDA methods' properties

An *i*-element set *M* of MCDA methods and a vector *m* of their properties with a dimension $\dim(m)$ are given. Therefore, there exists a matrix describing the properties of individual methods, in a form of *TAB* with dimensions $i \times \dim(m)$. There is also a decision-making problem (DP) described by decision problem descriptors in the form of a vector *c*, with $\dim(c) \leq \dim(m)$. For a given decision-making problem, from the set (vector) *c*, a subset of descriptors $\tilde{c}$ constituting a description of the decision-making situation is determined: $f(DP, c) = \tilde{c}$, where $\tilde{c} = [\tilde{c}_1, \ldots, \tilde{c}_j]$. Having a subset $\tilde{c}$ in the form of a vector and having a matrix *TAB* of the MCDA methods properties, a subset $\tilde{M}_{TAB}$ of methods is constructed according to the formula:

$$\tilde{M}_{TAB} = \left\{ M_k : \ \forall_{l \in [1,j]} \ \ \exists_{x \in [1, \dim(m)]} \ \tilde{c}_l = TAB[k,x] \ \ for \ 1 \leq k \leq i \right\}$$

where *i* denotes the number of MCDA methods in the set *M*, *j* denotes the number of descriptors in the subset $\tilde{c}$ (the length of the vector $\tilde{c}$).

The description of the decision-making problem, expressed with the *c* descriptors, is in accordance with the subset of the vector *m* of the properties of the MCDA methods belonging to the set *M*. Although formally the decision-making situation descriptors and the MCDA method properties are different sets, the problem descriptors are accurately reflected by the properties of particular methods.

3.2.1. Decision problem descriptors

In the proposed framework, we show that, each decision-making problem can be described by the DM using the maximum of nine descriptors belonging to the set $\tilde{c} \subseteq c$.

At the first level of the hierarchy, the DM only defines the general descriptors of the decision problem:

- c1 – whether different weights of the individual criteria will be taken into account in the decision problem; possible values are: 0 – no, 1 – yes;
- c2 – on what scale the criterial performance of the variants will be compared; possible values are: 1 – qualitative, 2 – quantitative, 3 – relative;
- c3 – whether the decision problem is characterized by uncertainty; possible values are: 0 – no, 1 – yes;
- c4 – what the decision problematic is; possible values are: 1 – selection, 2 – classification, 3 – ranking+selection,[1] 4 – classification+selection.

Of course, the knowledge about the decision problem can be clarified by the DM. While we can assume that c_2 is fully defined, the rest of the descriptors of the decision problem on the second level of the proposed hierarchy are presented as follows:

- c1.1 – if weights are used, what their type will be; possible values are: 1 – qualitative, 2 – quantitative, 3 – relative;
- c3.1 – if the problem is characterized by uncertainty, which uncertainty aspect it concerns; possible values are: 1 – input data uncertainty, 2 – DM's preference uncertainty, 3 – both;
- c4.1 – if the problematic of ranking is considered, what kind of variants' ranking is expected; possible values are: 1 – partial ranking, 2 – complete ranking.

The third level of the descriptors' hierarchy refers only $c_{3.1}$ and addresses data or preference uncertainty in the decision problem:

[1] The MCDA methods which deal with the ranking problematic are also efficient when considering the issue of choice

Table 2
The set of properties of the considered MCDA methods.

M_i	MCDA method	Abbr.	m_{i1}	$m_{i1.1}$	m_{i2}	m_{i3}	$m_{i3.1}$	$m_{i3.1.1}$	$m_{i3.1.2}$	m_{i4}	$m_{i4.1}$	Reference
M_1	AHP	A_H	1	3	3	0	0	0	0	3	2	[56]
M_2	AHP + TOPSIS	$A_H + T_P$	1	3	2	0	0	0	0	3	2	[56]
M_3	ANP	A_N	1	3	3	0	0	0	0	3	2	[57]
M_4	ARGUS	A_G	1	1	1	0	0	0	0	1	0	[58]
M_5	COMET	C_T	0	0	2	1	1	2	0	3	2	[59]
M_6	ELECTRE I	E_1	1	2	1	0	0	0	0	1	0	[27]
M_7	ELECTRE II	E_2	1	2	1	0	0	0	0	3	1	[27]
M_8	ELECTRE III	E_3	1	2	2	1	2	0	3	3	1	[60]
M_9	ELECTRE IS	E_S	1	2	2	1	2	0	3	1	0	[27]
M_{10}	ELECTRE IV	E_4	0	0	1	1	2	0	3	3	1	[27]
M_{11}	ELECTRE TRI	E_T	1	2	2	1	2	0	3	2	0	[27]
M_{12}	EVAMIX	E_V	1	2	2	0	0	0	0	3	2	[61]
M_{13}	Fuzzy AHP	A_F	1	3	3	1	1	3	0	3	2	[62]
M_{14}	Fuzzy AHP + fuzzy TOPSIS	$A_F + T_F$	1	3	2	1	1	3	0	3	2	[63]
M_{15}	Fuzzy ANP	A_{NF}	1	3	3	1	1	3	0	3	2	[64]
M_{16}	Fuzzy ANP + fuzzy TOPSIS	$A_{NF} + T_F$	1	3	2	1	1	3	0	3	2	[56]
M_{17}	Fuzzy MIN_MAX[1]	E_F	0	0	1	1	1	2	0	4	0	[65]
M_{18}	Fuzzy PROMETHEE I	P_{1F}	1	2	2	1	3	3	3	3	1	[66]
M_{19}	Fuzzy PROMETHEE II	P_{2F}	1	2	2	1	3	3	3	3	2	[66]
M_{20}	Fuzzy SAW	S_F	1	2	2	1	1	3	0	3	2	[67]
M_{21}	Fuzzy TOPSIS	T_F	1	2	2	1	1	3	0	3	2	[53]
M_{22}	Fuzzy VIKOR	V_F	1	2	2	1	1	3	0	3	2	[68]
M_{23}	Goal Programming	G_P	0	0	2	0	0	0	0	1	0	[69]
M_{24}	IDRA	I_D	1	2	2	0	0	0	0	3	1	[13]
M_{25}	Lexicographic method	L_M	1	1	1	0	0	0	0	1	0	[70]
M_{26}	MACBETH	M_B	1	3	3	0	0	0	0	3	2	[71]
M_{27}	MAPPAC	M_P	1	2	2	0	0	0	0	3	1	[72]
M_{28}	MAUT	M_U	1	2	2	0	0	0	0	3	2	[73]
M_{29}	MAVT	M_V	1	2	2	0	0	0	0	3	2	[73]
M_{30}	Maximax	M_X	0	0	1	0	0	0	0	1	0	[74]
M_{31}	Maximin	M_N	0	0	1	0	0	0	0	1	0	[74]
M_{32}	Maximin fuzzy method	M_F	1	2	2	1	1	2	0	1	0	[54]
M_{33}	MELCHIOR	M_C	1	1	2	1	2	0	3	3	1	[75]
M_{34}	MIN_MAX[1]	E_M	0	0	1	0	0	0	0	1	0	[74]
M_{35}	NAIADE I	N_1	0	0	2	1	1	2	0	3	1	[76]
M_{36}	NAIADE II	N_2	0	0	2	1	1	2	0	3	2	[76]
M_{37}	ORESTE	O_R	1	1	2	1	2	0	1	3	1	[77]
M_{38}	PACMAN	P_C	1	2	2	0	0	0	0	3	1	[78]
M_{39}	PAMSSEM I	P_{A1}	1	2	2	1	3	2	3	3	1	[79]
M_{40}	PAMSSEM II	P_{A2}	1	2	2	1	3	2	3	3	2	[79]
M_{41}	PRAGMA	P_G	1	2	2	0	0	0	0	3	1	[80]
M_{42}	PROMETHEE I	P_1	1	2	2	1	2	0	3	3	1	[81]
M_{43}	PROMETHEE II	P_2	1	2	2	1	2	0	3	3	2	[81]
M_{44}	QUALIFLEX‘	Q_F	1	1	1	0	0	0	0	3	1	[82]
M_{45}	REGIME	R_G	1	1	1	0	0	0	0	3	1	[83]
M_{46}	SAW	S_A	1	2	2	0	0	0	0	3	2	[74]
M_{47}	SMART	S_M	1	2	2	0	0	0	0	3	2	[84]
M_{48}	TACTIC	T_C	1	2	2	1	2	0	1	1	0	[15]
M_{49}	TOPSIS	T_P	1	2	2	0	0	0	0	3	2	[85]
M_{50}	UTA	U_T	1	2	2	0	0	0	0	3	2	[86]
M_{51}	VIKOR	V_K	1	2	2	0	0	0	0	3	2	[87]
M_{52}	AHP + fuzzy TOPSIS	$A_H + T_F$	1	3	2	1	1	2	0	3	2	[56]
M_{53}	Fuzzy AHP + TOPSIS	$A_F + T_P$	1	3	2	1	1	1	0	3	2	[56]
M_{54}	AHP + VIKOR	$A_H + V_K$	1	3	2	0	0	0	0	3	2	[88]
M_{55}	DEMATEL	D_M	1	3	3	0	0	0	0	3	2	[89]
M_{56}	REMBRANDT	R_M	1	3	3	0	0	0	0	3	2	[90]

MIN_MAX[1] - Methods of extracting the minimum and maximum values of the attribute.

c3.1.1 – if the uncertainty concerns the data, does it refer to the weights of criteria or to the variants' criterial performance; possible values are: 1 – criteria, 2 – variants, 3 – both;

c3.1.2 – if the uncertainty concerns the DM's preferences, what thresholds will be used in the decision problem; possible values are: 1 – indifference, 2 – preference, 3 – both.

3.2.2. MCDA methods' properties

As it was noted above, the descriptors c correspond to the characteristics m. In such a manner, the considered descriptors were encoded for all considered 56 MCDA methods. Table 2 provides a full description of the MCDA methods depending on all the indicated characteristics (0 means lack of ability). It is worth noting that the inclusion of characteristics relating to all levels of the hierarchy allows to divide MCDA methods into relatively few groups.

3.2.3. Practical mapping between decision problem descriptors and MCDA methods' properties

The relationships between the set of the MCDA methods' characteristics and the set of a decision problem's descriptors can be presented by analyzing an exemplary decision problem and the procedure of the MCDA method selection for solving it. In [91], a decision-making problem of constructing a ranking of premises for urban distribution centers was considered. It considered three alternative locations in terms of 11 criteria. During the selection of the MCDA method for the given decision-making problem, a full set of descriptors was used, i.e. $\tilde{c} = c$. The decision-making

Table 3
The rules of selecting a suitable MCDA method.

MCDA method properties		m_{i1}	m_{i2}	m_{i3}	m_{i4}	$m_{i1.1}$	$m_{i3.1}$	$m_{i4.1}$	$m_{i3.1.1}$	$m_{i3.1.2}$	Subset of MCDA methods	
											Names	Abbreviations
Rules	R_1	0	1	0	1	0	0	0	0	0	Maximax, Maximin, MIN_MAX[1]	$\{M_X, M_N, E_M\}$
	R_2	0	1	1	4	0	1	0	2	0	FuzzyMIN_MAX[1]	$\{E_F\}$
	R_3	0	1	1	3	0	2	1	0	3	ELECTRE IV	$\{E_4\}$
	R_4	0	2	0	1	0	0	0	0	0	Goal Programming	$\{G_P\}$
	R_5	0	2	1	3	0	1	1	2	0	NAIADE I	$\{N_1\}$
	R_6	0	2	1	3	0	1	2	2	0	COMET, NAIADE II	$\{C_T, N_2\}$
	R_7	1	1	0	1	1	0	0	0	0	ARGUS, Lexicographic method	$\{A_G, L_M\}$
	R_{11}	1	1	0	1	2	0	0	0	0	ELECTRE I	$\{E_1\}$
	R_8	1	1	0	3	1	0	1	0	0	QUALIFLEX, REGIME	$\{Q_F, R_G\}$
	R_{12}	1	1	0	3	2	0	1	0	0	ELECTRE II	$\{E_2\}$
	R_{13}	1	2	0	3	2	0	1	0	0	IDRA, MAPPAC, PACMAN, PRAGMA	$\{I_D, M_P, P_C, P_G\}$
	R_{14}	1	2	0	3	2	0	2	0	0	EVAMIX, MAUT, MAVT, SAW, SMART, TOPSIS, UTA, VIKOR	$\{E_V, M_U, M_V, S_A, S_M, T_P, U_T, V_K\}$
	R_{26}	1	2	0	3	3	0	2	0	0	AHP + TOPSIS, AHP + VIKOR	$\{A_H + T_P, A_H + V_K\}$
	R_{15}	1	2	1	1	2	1	0	2	0	Maximin fuzzy method	$\{M_F\}$
	R_{17}	1	2	1	1	2	2	0	0	1	TACTIC	$\{T_C\}$
	R_{18}	1	2	1	1	2	2	0	0	3	ELECTRE IS	$\{E_S\}$
	R_{19}	1	2	1	2	2	2	0	0	3	ELECTRE TRI	$\{E_T\}$
	R_9	1	2	1	3	1	2	1	0	1	ORESTE	$\{O_R\}$
	R_{10}	1	2	1	3	1	2	1	0	3	MELCHIOR	$\{M_C\}$
	R_{16}	1	2	1	3	2	1	2	3	0	Fuzzy SAW, Fuzzy TOPSIS, Fuzzy VIKOR	$\{S_F, T_F, V_F\}$
	R_{20}	1	2	1	3	2	2	1	0	3	ELECTRE III, PROMETHEE I	$\{E_3, P_1\}$
	R_{21}	1	2	1	3	2	2	2	0	3	PROMETHEE II	$\{P_2\}$
	R_{22}	1	2	1	3	2	3	1	2	3	PAMSSEM I	$\{P_{A1}\}$
	R_{24}	1	2	1	3	2	3	1	3	3	Fuzzy PROMETHEE I	$\{P_{1F}\}$
	R_{23}	1	2	1	3	2	3	2	2	3	PAMSSEM II	$\{P_{A2}\}$
	R_{25}	1	2	1	3	2	3	2	3	3	Fuzzy PROMETHEE II	$\{P_{2F}\}$
	R_{27}	1	2	1	3	3	1	2	1	0	Fuzzy AHP + TOPSIS	$\{A_F + T_P\}$
	R_{28}	1	2	1	3	3	1	2	2	0	AHP + fuzzy TOPSIS	$\{A_H + T_F\}$
	R_{29}	1	2	1	3	3	1	2	3	0	Fuzzy AHP + fuzzy TOPSIS, Fuzzy ANP + fuzzy TOPSIS	$\{A_F + T_F, A_{NF} + T_F\}$
	R_{30}	1	3	0	3	3	0	2	0	0	AHP, ANP, MACBETH, DEMATEL, REMBRANDT	$\{A_H, A_N, M_B, D_M, R_M\}$
	R_{31}	1	3	1	3	3	1	2	3	0	Fuzzy AHP, Fuzzy ANP	$\{A_F, A_{NF}\}$

MIN_MAX[1] - methods of extracting the minimum and maximum values of the attribute.

problem includes the weights of criteria in quantitative form, so the descriptors of the decision problem have taken the values $c_1 = 1$, $c_{1.1} = 2$. In addition, the efficiency of the variants was expressed on a quantitative scale ($c_2 = 2$). The decision-making problem was characterized by uncertainty ($c_3 = 1$), where the uncertainty referred to the input data ($c_{3.1} = 1$), and in particular to the weightings of the criteria and performance of the decision variants ($c_{3.1.1} = 3$, $c_{3.1.2} = 0$). The considered decision problematic was the problematic of ranking, and the obtained solution was a complete ranking, i.e. ranking without incomparability ($c_4 = 3$, $c_{4.1} = 2$). It is easy to notice that the individual descriptors c correspond to the m characteristics of the same values, i.e. $m_{i1} = 1$, $m_{i1.1} = 2$, $m_{i2} = 2$, $m_{i3} = 1$, $m_{i3.1} = 1$, $m_{i3.1.1} = 3$, $m_{i3.1.2} = 0$, $m_{i3.4} = 3$, $m_{i4.1} = 2$ etc. Analysis of Table 2 allows to notice three MCDA methods having such characteristics vectors: Fuzzy SAW(M_{20}), Fuzzy TOPSIS (M_{21}) and Fuzzy VIKOR (M_{22}).

3.2.4. MCDA method properties' explanation

When we have a given decision problem, its requirements with relations to properties of individual MCDA methods can be determined.

The property m_1 refers to the weights of the criteria. MCDA methods may use qualitative, quantitative or relative weights, as well as may not use criteria weights. For example, in [92], the criteria weights are not used, which results in the properties related to the criteria m_1 and $m_{1.1}$ not being met. On the other hand, in [93], quantitative weights of criteria were applied, which means that the property m_1 is met, and the property $m_{1.1}$ obtained the value of 2. Finally, in [94], the weights of criteria were compared pairwise in the form of a comparison matrix, thus providing a weights vector. Therefore, the property m_1 was met, and the property $m_{1.1}$ obtained the value of 3.

The second property describes the scale at which the performance of the variants in each of the criteria are compared or determined. As in the case of the criteria weights, this scale can be qualitative, quantitative or relative. In [89,95], only the significance of individual criteria related to Green Supply Chain Management (GSCM) was examined, without considering any decision variants. This means that the property m_2 of the decision problem is not met. In contrast, property m_2 is met for example in [96], where the variants were compared on a qualitative scale and, therefore, property m_2 is given the value of 1. In [69], a quantitative scale was used to compare the variants, so that m_2 obtained the value of 2. Eventually, in [97], the comparative scale was used for comparisons of variants (pairwise comparison matrix), therefore, property m_2 obtains the value of 3.

Property m_3 refers to the uncertainty of the decision problem. The uncertainty may refer to the input data describing the criteria weights or the variants' performance in each criterion. In such case, the data is expressed with the use of fuzzy numbers. On the other hand, the uncertainty may also apply to the preferences of the decision makers. This kind of uncertainty is expressed with the use of the thresholds of indifference and preference. The indifference threshold determines the difference in the criterion performance of individual variants, at which they can be considered to be equally good. On the other hand, the threshold of preference defines the difference in the performance of the variants, in which one of the variants is considered to be definitely better than the other. Uncertainty is included e.g. in [98]. It is an uncertainty related to data at the level of the criteria weights and the performance of the criteria. Therefore, the properties m_3, $m_{3.1}$ and $m_{3.1.1}$ are fulfilled, with $m_{3.1}$ being 1 and $m_{3.1.1}$ being 3. In contrast, in [99], uncertainty about the decision maker's preferences occur, so the thresholds of indifference and preference were applied. Therefore, the properties m_3, $m_{3.1}$ and $m_{3.1.2}$ are met in this case, with $m_{3.1}$ being 2 and $m_{3.1.2}$ being 3.

Last, but not least, the m_4 property refers to the decision problematics. It should be clarified that the methods dealing with the ranking problem, also allow to solve the choice problem, and, therefore, one of the possible values of the property m_4 includes both the ranking and the choice problems. If the MCDA method considers a ranking problem, it may provide the results in the form of a full (total order) or partial ranking (partial order). A method supporting the total order usually allows to obtain global performance values of the variants in numerical form and to determine for each pair of variants which one is better. In contrast, methods supporting the partial order do not provide a full comparability of the variants and most often express the global efficiency of variants on an ordinal scale, which, additionally, does not allow indicate which variant is better for any pair of variants. For example, in [96], the issue of choice is considered, so the property m_4 takes the value of 1. The problem of ranking is considered e.g. in [100] and [101], where the property m_4 obtains the value of 3. In the former, a total order of variants is produced, thus the property $m_{4.1}$ obtains the value of 2, whereas in the latter a partial order of variants was obtained, so the property $m_{4.1}$ takes the value of 1.

3.3. Formal representation of the MCDA methods' properties

When looking for a formal approach for selecting an MCDA method, applying a classifier, as it was done in a number of studies [21,22,32], seems to be an interesting concept. In current work we propose the set of descriptors identifying MCDA methods' properties for a particular decision situation, presented below:

The descriptor c_1 checks if weights of any kind will be used in the decision problem:

$$c_1(DP) = \begin{cases} 1 & (\exists\{g_1 \succ g_2 \succ \ldots \succ g_m\}) \otimes \left(\underset{g_i, g_j \in G}{\forall} \exists r : \left| g_i - g_j \right| = r \right) \\ & \otimes \left(\underset{g_i, g_j \in G}{\forall} \exists W_{|G|\times|G|} : w_{ij} = g_i / g_j \right) \\ 0 & otherwise \end{cases}$$

where:

- r – the quantitative difference between the weights of criteria (g_i and g_j),
- w_{ij} – a relative weight of a criterion g_i in respect to a criterion g_j,
- $W_{|G|x|G|}$ – matrix of pairwise comparisons, where size is equal to the size of a set of criteria G.

The descriptor $c_{1.1}$ is responsible for distinguishing the type of weights used, which can be expressed on the scale of one of the following types: qualitative, quantitative or relative:

$$c_{1.1}(DP) = \begin{cases} 1 & \exists\{g_1 \succ g_2 \succ \ldots \succ g_m\} \\ 2 & \underset{g_i, g_j \in G}{\forall} \exists r : \left| g_i - g_j \right| = r \\ 3 & \underset{g_i, g_j \in G}{\forall} \exists W_{|G|\times|G|} : w_{ij} = g_i / g_j \end{cases}$$

The descriptor c_2 distinguishes the type of scale on which the decision variants will be compared. These can be: qualitative, quantitative or relative scales. The descriptor c_2 also includes a situation when the variants are not compared. It is due to the fact that in situations where MCDA methods are used, such situations still may occur, even though all of the considered MCDA methods include comparisons of variants. The descriptor c_2 value is determined according to formula:

$$c_2(DP) = \begin{cases} 1 & \underset{\substack{a \in A \\ g_i \in G}}{\forall} \exists\{g_i(a_1) \succ g_i(a_2) \succ \ldots \succ g_i(a_m)\} \\ 2 & \underset{\substack{a_j, a_k \in A \\ g_i \in G}}{\forall} \exists r : \left| g_i(a_j) - g_i(a_k) \right| = r \\ 3 & \underset{\substack{a_j, a_k \in A \\ g_i \in G}}{\forall} \exists E_{|A|\times|A|} : e_{jk} = g_i(a_j) / g_i(a_k) \\ 0 & otherwise \end{cases}$$

where:

$g_i(a)$ – the performance of a variant a concerning a criterion i and a weight g_i,
r – the quantitative difference between performances of variants a with respect to criterion g,
e_{jk} – relative criterion performance (for the criterion g) of variant a_j with respect to a variant a_k,
$E_{|A|x|A|}$ – matrix of pairwise comparisons, where size is equal to the cardinality set of variants A.

The descriptor c_3 checks if there is data or preference uncertainty in the decision problem:

$$c_3(DP) = \begin{cases} 1 & \left(\underset{g_i\in G}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i = \widetilde{N}_{fuzzy}\right) \vee \left(\underset{\substack{a_j\in A\\ g_i\in G}}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i(a_j) = \widetilde{N}_{fuzzy}\right) \vee \ldots \\ & \left(\underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(q) : \begin{cases} g_i(a_j) - g_i(a_k) \le q \Rightarrow g_i(a_j) \sim g_i(a_k) \\ g_i(a_j) - g_i(a_k) > q \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases}\right) \vee \left(\underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(p) : \begin{cases} g_i(a_j) - g_i(a_k) = 0 \Rightarrow g_i(a_j) \sim g_i(a_k) \\ p > g_i(a_j) - g_i(a_k) > 0 \Rightarrow g_i(a_j) \succ_{weak} g_i(a_k) \\ g_i(a_j) - g_i(a_k) \ge p \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases}\right) \\ 0 & otherwise \end{cases}$$

where:

N_{fuzzy} – triangular of a trapezoidal fuzzy number,
$(n_l, n_u, \alpha_F, \beta_F)$ – parameters of a fuzzy number membership function,
q – indifference threshold,
p – preference threshold.

The descriptor $c_{3.1}$ verifies if the uncertainty is related particularly to input data, to preferences, or to both of them:

$$c_{3.1}(DP) = \begin{cases} 1 & \left(\underset{g_i\in G}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i = \widetilde{N}_{fuzzy}\right) \vee \left(\underset{\substack{a_j\in A\\ g_i\in G}}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i(a_j) = \widetilde{N}_{fuzzy}\right) \\ 2 & \left(\underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(q) : \begin{cases} g_i(a_j) - g_i(a_k) \le q \Rightarrow g_i(a_j) \sim g_i(a_k) \\ g_i(a_j) - g_i(a_k) > q \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases}\right) \vee \left(\underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(p) : \begin{cases} g_i(a_j) - g_i(a_k) = 0 \Rightarrow g_i(a_j) \sim g_i(a_k) \\ p > g_i(a_j) - g_i(a_k) > 0 \Rightarrow g_i(a_j) \succ_{weak} g_i(a_k) \\ g_i(a_j) - g_i(a_k) \ge p \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases}\right) \\ 3 & \left(\underset{g_i\in G}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i = \widetilde{N}_{fuzzy}\right) \vee \left(\underset{\substack{a_j\in A\\ g_i\in G}}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i(a_j) = \widetilde{N}_{fuzzy}\right) \vee \ldots \\ & \left(\underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(q) : \begin{cases} g_i(a_j) - g_i(a_k) \le q \Rightarrow g_i(a_j) \sim g_i(a_k) \\ g_i(a_j) - g_i(a_k) > q \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases}\right) \vee \left(\underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(p) : \begin{cases} g_i(a_j) - g_i(a_k) = 0 \Rightarrow g_i(a_j) \sim g_i(a_k) \\ p > g_i(a_j) - g_i(a_k) > 0 \Rightarrow g_i(a_j) \succ_{weak} g_i(a_k) \\ g_i(a_j) - g_i(a_k) \ge p \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases}\right) \end{cases}$$

The descriptor $c_{3.1.1}$ further divides the input data uncertainty into the ones in which the fuzzy sets were used to the criteria's weights, to the variants' performance or to both of them:

$$c_{3.1.1}(DP) = \begin{cases} 1 & \underset{g_i\in G}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i = \widetilde{N}_{fuzzy} \\ 2 & \underset{\substack{a_j\in A\\ g_i\in G}}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i(a_j) = \widetilde{N}_{fuzzy} \\ 3 & \underset{\substack{a_j\in A\\ g_i\in G}}{\forall}\ \exists \widetilde{N}_{fuzzy} = (n_l; n_u; \alpha_F; \beta_F)_{LR} : g_i, g_i(a_j) = \widetilde{N}_{fuzzy} \end{cases}$$

The descriptor $c_{3.1.2}$ further divides the preference uncertainty by distinguishing the situations in which the thresholds of indifference, preference or both of them were used:

$$c_{3.1.2}(DP) = \begin{cases} 1 & \underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(q) : \begin{cases} g_i(a_j) - g_i(a_k) \le q \Rightarrow g_i(a_j) \sim g_i(a_k) \\ g_i(a_j) - g_i(a_k) > q \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases} \\ 2 & \underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(p) : \begin{cases} g_i(a_j) - g_i(a_k) = 0 \Rightarrow g_i(a_j) \sim g_i(a_k) \\ p > g_i(a_j) - g_i(a_k) > 0 \Rightarrow g_i(a_j) \succ_{weak} g_i(a_k) \\ g_i(a_j) - g_i(a_k) \ge p \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases} \\ 3 & \underset{\substack{a_j,a_k\in A\wedge j\neq k\\ g_i\in G}}{\forall}\ \exists(p,q) : \begin{cases} g_i(a_j) - g_i(a_k) \le q \Rightarrow g_i(a_j) \sim g_i(a_k) \\ p > g_i(a_j) - g_i(a_k) > q \Rightarrow g_i(a_j) \succ_{weak} g_i(a_k) \\ g_i(a_j) - g_i(a_k) \ge p \Rightarrow g_i(a_j) \succ g_i(a_k) \end{cases} \end{cases}$$

The descriptor c_4 checks which problematic is considered in the decision problem:

$$c_4(DP) = \begin{cases} 1 & f\left(\underset{u\in A'\subset A}{\max}\left\{S_D(u) : \dim(u) = \min\left\{\dim(v);\ \underset{v\in A\setminus A'}{\forall}\ \underset{q\in A'}{\exists}\ (\neg S_D(q))\right\}\right\}\right) \\ 2 & f(u_B);\ \underset{u\in A}{\exists}\ u_B = u \wedge \underset{v\in A, v\neq u}{\forall}\ \eta(u) > \eta(v) \\ 3 & f(k_R);\ \underset{k\in KR}{\exists}\ k_R = k \wedge \underset{kp\in KR, kp\neq k}{\forall}\ \upsilon(k) \approx \upsilon(kp) \\ 4 & f\left(\underset{u\in A'\subset A}{\max}\left\{S_D(u) : \dim(u) = \min\left\{\dim(v);\ \underset{v\in A\setminus A'}{\forall}\ \underset{q\in A'}{\exists}\ (\neg S_D(q))\right\}\right\}\right) \vee f(u_B);\ \underset{u\in A}{\exists}\ u_B = u \wedge \underset{v\in A, v\neq u}{\forall}\ \eta(u) > \eta(v) \end{cases}$$

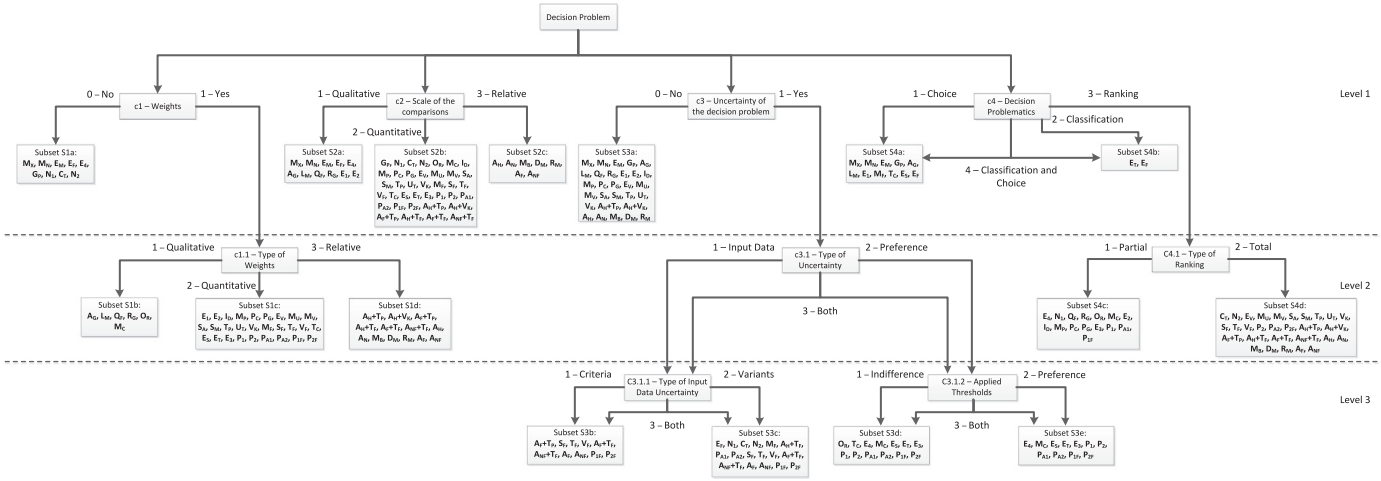


Fig. 2. The decision tree of selecting a suitable MCDA method on the basis of the proposed descriptors.

where:

dim – size,
$S_D()$ – DM's satisfaction with the variant,
η – the norm related to the certain values,
KR – the set of equivalence classes of variants from the set A,
$\approx$ – the relation of partial or complete order.

If the problematic of ranking is considered, the descriptor $c_{4.1}$ checks whether a partial or full order of variants is expected:

$$c_{4.1}(DP) = \begin{cases} 1 & \underset{E=G(A)}{\forall} \exists a_i, a_j : E(a_i)\, R\, E(a_j) \\ 2 & \underset{E=G(A)}{\forall} \exists \{E(a_1) \succ E(a_2) \succ \ldots \succ E(a_m)\} \end{cases}$$

where:

$E(a)$ – a global performance of variant a,
R – incomparability relation.

3.4. Presentation of the MCDA methods' properties using tree representation

The applied classifier can also be presented in a form of decision trees and, in consequence, the whole classification process would have the form of a method selection tree, as it is called by Guitouni and Martel [20]. The decision tree of selecting a suitable MCDA method on the basis of the proposed descriptors is presented on Fig. 2.

The tree presents the problem of an MCDA method selection depending on the information about the decision problem known to the DM. To specify a subset of methods that meet the descriptors describing the decision problem, the algebra of sets should be used.

If the DM has full information about the decision problem, and, therefore, knows what kind of weights should be used, on what scale the variants should be compared, what kind of uncertainty should be included in the decision problem and what the decision problematic is, then the appropriate subset of MCDA methods is determined as intersection of relevant subsets *S1a: S4d* (horizontal approach). For example, if the decision problem is described by descriptors $c_1 = 1$, $c_{1.1} = 2$, $c_2 = 3$, $c_3 = 0$ and $c_4 = 1$, the subset of methods is the intersection of the sets $S1c \cap S2b \cap S3a \cap S4a$. The descriptors $c_{3.1.1}$ and $c_{3.1.2}$ taking the value of 3 and the descriptor c_4 taking the value of 4 are special cases. In particular, regarding the descriptor $c_{3.1.1} = 3$, to account for the MCDA methods considering the data uncertainty for both weights of criteria and the variants, the intersection of sets $S3b \cap S3c$ should be used. In the other cases mentioned above, the intersection of sets should be used analogously.

On the other hand, for the decision problem which the DM cannot fully define, it may be necessary to apply the union of sets (vertical approach). This allows the inclusion of more general descriptors, occurring at levels 1 and 2 of the hierarchy, to which no subsets of methods have been directly assigned. For example, if weights are included in the decision problem, but the scale on which they should be expressed is unknown, the determination of the appropriate methods takes place using the union of the sets $S1b \cup S1c \cup S1d$.

Consequently, the adaptation of the MCDA method to an incompletely defined decision problem is a combination of the vertical approach, referring to the lack of information about the decision-making problem, and the horizontal approach, referring to the certain information. For example, if the DM knowns that in the decision problem: quantitative weights should be used ($c_1 = 1$ and $c_{1.1} = 2$), there exists data uncertainty related both to the weights of criteria and to the variants ($c_3 = 1$, $c_{3.1} = 1$, $c_{3.1.1} = 3$), the ranking problematic should be considered, and a full order should be obtained ($c_4 = 3$ and $c_{4.1} = 2$), but there is uncertainty on what scale the variants should be compared, the subset of the appropriate methods will result from the $S1c \cap S3b \cap S3c \cap S4d \cap (S2a \cup S2b \cup S2c)$ operation.

3.5. Rules database generation

It needs to be noted that the characteristics describing properties of individual multi-criteria methods would also be used as conditional attributes, whereas specific MCDA methods would be decision attributes which constitutes foundations for the MCDA selection rules set. Of the 56 MCDA methods considered, there were only 31 unique sequences of encoded characteristics. This implicates that some of the methods have identical characteristics. A subset of suitable MCDA methods can be recommended. The characteristics of the problem are not reproducible as it is in the case of the MCDA methods. The problem characteristics should be identified each time separately. On the basis of the proposed properties the expert rule base is obtained. The set of rules is presented in Table 3. In addition, rules at various levels of the hierarchy are included here. The first level, limited to the most general properties (m_1, m_2, m_3, m_4), allows defining 13 distinct rules. On the second level of the hierarchy, which includes more specific properties of MCDA methods, a total of 25 rules was distinguished, allowing

the selection of MCDA methods depending on the properties of the decision problem. Last, but not least, the third and most detailed level of the hierarchy allows to define the aforementioned 31 rules.

The presented table shows that the MCDA method can be selected depending on the values of the descriptors at a successive level. However it is clearly visible that the lack of knowledge about a particular level decreases the quality of recommendation. A detailed analysis is presented in Supplementary material – Section 4.When analyzing the number of methods assigned to the decision problem by individual rules, it is easy to notice that the R_{14} rule is the most capacious one. Based on this rule, eight methods are indicated as appropriate to solve a problem of a specific character: EVAMIX, MAUT, MAVT, SAW, SMART, TOPSIS, UTA, VIKOR. The high number of methods in this rule results from the great similarity of the majority of the methods included in it. The MAVT method is basically a simplification of the MAUT method, with the only significant difference being the fact that during the aggregation MAVT uses a value function, and MAUT – a utility function. The value function, in contrast to the utility function, does not take into account the risk (probability) [102]. The SAW method, on the other hand, is the simplest case of the MAVT method, where the additive value function is normalized to the [0,1] interval [103]. Similarly, the SMART method is a simplification of MAVT / MAUT, in which an additive model is used with a linear approximation of the utility / value function [104]. In turn, the UTA method uses partial, criterial usability / values functions that are monotonic and sectionally linear. The partial functions are then aggregated using the additive value function [38]. It can be, therefore, concluded that the MAVT, SAW, SMART and UTA methods are special cases of the MAUT method. Equally significant similarity can be observed between the TOPSIS and VIKOR methods, which are based on the same principles. They differ only in the aggregation function and the normalization method used. In TOPSIS, the aggregation function minimizes the distance to the ideal solution and maximizes the distance from the anti-ideal solution, whereas in VIKOR, the aggregation function only minimizes the distance to the ideal solution. As for normalization, vector normalization is used in TOPSIS and linear normalization is used in VIKOR [87]. Based on the aforementioned observations, it can be therefore concluded that the R_{14} rule essentially includes three different subsets of methods similar to each other.

It needs to be noted that the rule-based approach presented in Table 3 is different than the approach based on the algebra of sets and the tree structure (Fig. 2) presented in Section 3.4. In the tree structure, descriptors $c_{3.1}$, $c_{3.1.1}$, $c_{3.1.2}$ and c_4 assign an MCDA method to one of two disjoint sets or to the intersection of the sets. In turn, in the rules presented in Table 3, each of the possibilities (both subsets, as well as their intersection) is coded separately.

3.6. Uncertainty handling in the decision problem description

Nevertheless, it needs to be noted that the aforementioned analysis can be oversimplifying of the real decision making problem. Often, the DM might not have the full knowledge of the decision problem, or the possessed knowledge might not provide the full knowledge of any of the levels of hierarchy, thus introducing uncertainty to the decision-making. Therefore, the DM can know the values of as much as 4, 7 or 9 classifiers for a single, two or three levels of descriptors respectively, or as little as a single classifier.

For this reason, the authors decided to introduce a more profound modelling of the uncertainty of data and the decision-making problem. In the next step of the empirical research, a set of all possible 450,000 rules for all possible values of each classifier was generated. However, in majority of the rules, some of the classifiers were in conflict – e.g. $c_1 = 0$ (no weights) but $c_{1.1} = 3$ (relative weights) – and, therefore, a subset of the rules was extracted, based on the criteria presented in Table S5 in Supplementary material – Section 5. The extraction process consisted of four steps, depicted on Fig. S5 in Supplementary material – Section 5. In the first step, the full set of rules was filtered four times to extract the rules with the valid values of c_1 and $c_{1.1}$ classifiers. In the second step, the output of the first step was then filtered with the valid values of the c_2 classifier. The output was then filtered with the valid values of the c_3, $c_{3.1}$, $c_{3.1.1}$ and $c_{3.1.2}$ classifiers in step three. Eventually, the output was filtered with the valid values of the classifiers c_4 and $c_{4.1}$. As a result, the original set of 450 thousand rules was reduced to 4,536 rules. Furthermore, after the removal of the rules returning 0 methods, a final set of 656 rules was obtained. A similar procedure was then performed for the hierarchies consisting of two levels and a single level of classifiers.

Fig. 3 illustrates a histogram-like analysis of the possible MCDA selection rules, based on the number of unknown MCDA characteristics, as well as the complexity of the structure. The x-axis and right y-axis represent the number of the characteristics that are unknown, whereas the left y-axis represents the growth in number of methods that possibly meet these unknown characteristics. The bars in the chart represent the precise count of methods matching every single rule, whilst the dashed line represents their average count for the corresponding number of unknowns. The presented rules are limited to the 656 ones that returned at least one method. A detailed analysis of the rule sets containing all rules, including the ones returning empty sets of methods, is provided in Supplementary material – Section 6.

The analysis of Fig. 3 and Table 4 allows to observe that in case of the 1-level hierarchy of the decision rules, if the DM cannot decide on a single characteristic, on average, the number of matching MCDA methods would be almost 2 times higher than in case of a single unknown in the 2-level hierarchy. On the other hand, the difference in case of the 2 levels and 3 levels of characteristics is equal to only 0.8046, for a single unknown. In case of two unknown values of the MCDA characteristics, the average number of possible matching methods is over 2.5 times higher than in case of 2 levels and almost 4 times higher than in case of 3 levels. In order to match the number of methods produced by a single unknown in the 1-level rule set, at least 3 variables should be unknown for the 2-level one and 5 for the 3-level one.

When Fig. 3c is analyzed in detail, it can be observed that when a single characteristic is unknown to the decision maker, the 3-level decision framework still allows to limit the matching number of MCDA methods to a range of 1 to 12, with average value equal to 2.1446. If the number of unknown decision problem descriptors grows to two values, a significant increase of non-empty rules can be observed, to a total of 131. The average number of matched methods increases only slightly to 2.5191. A similar growth of possibly matching methods can be observed also when the count of unknown abilities grows to 3 and 4, with the average equal to 3.0511 and 3.7719 respectively. However, if the number of unknown characteristics increases any further, the speed of growth of the number of methods starts to increase rapidly which fact is illustrated on Fig. 4a.

The average count of matching methods for cases when 5, 6, 7 or 8 of the total of 9 characteristics are unknown is equal to 5.2099, 8.0400, 14.4545 and 29 respectively. This growth can be mapped by a 4-degree polynomial function with the R^2 equal to 0.9997. It is also important to note from Fig. 3c, that along with the increase of the count of methods, the number of rules decreases when more MCDA characteristics are unknown.

Fig. 3a, 3b depict the two remaining scenarios when the knowledge about the decision problem is structured only into two levels (Fig. 3b) or into a sequence of 4 main classifiers (Fig. 3a). A linear

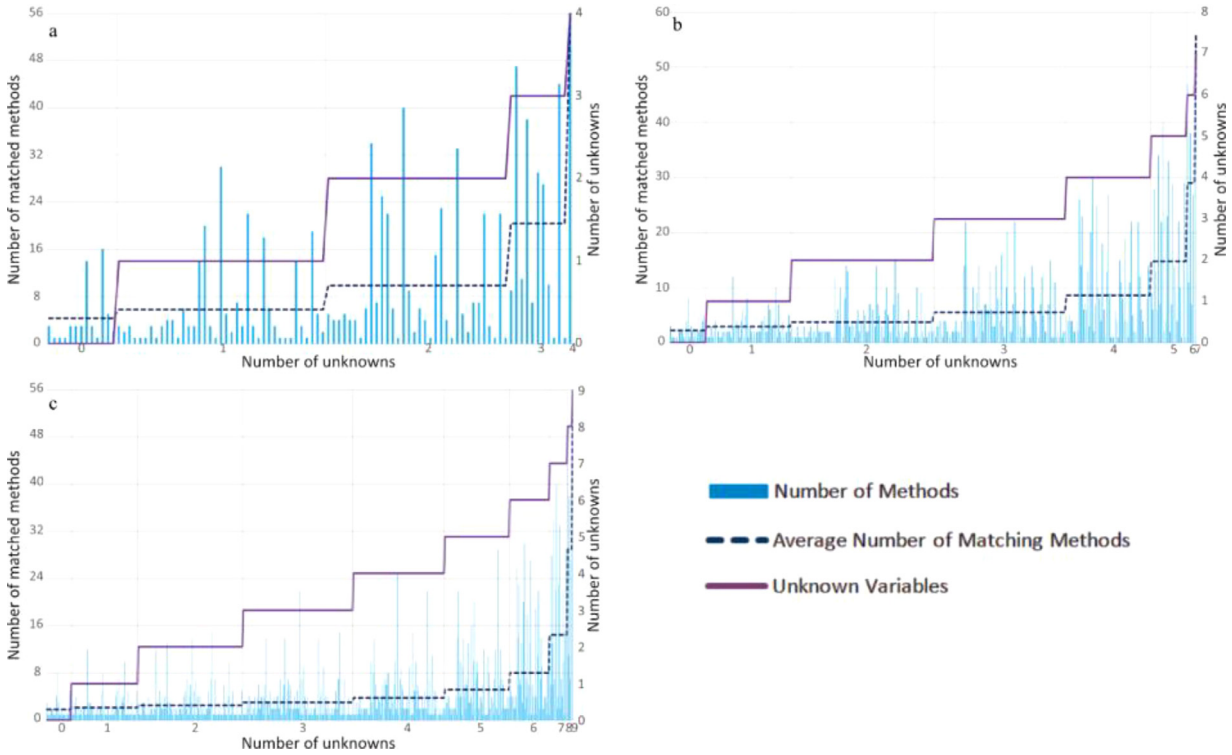


Fig. 3. A histogram-like analysis of the possible MCDA selection rules depending on the number of unknown characteristics, in cases of a single level (a), two levels (b) and three levels (c) of MCDA methods' properties, excluding the rules returning 0 methods.

Table 4
Comparison of the minimum, average and maximum number of methods for classifiers organized into one, two or three levels.

Unknowns	1 Level			2 Levels			3 Levels		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
0	1	4.3077	16	1	2.2400	8	1	1.8065	8
1	1	5.7436	30	1	2.9492	12	1	2.1446	12
2	1	9.8824	40	1	3.7374	15	1	2.5191	15
3	1	20.3636	47	1	5.5000	22	1	3.0511	22
4	56	56.0000	56	1	8.5763	30	1	3.7719	25
5				1	14.8000	40	1	5.2099	29
6				7	29.0000	47	1	8.0400	30
7				56	56.0000	56	1	14.4545	40
8							7	29.0000	47
9							56	56.0000	56

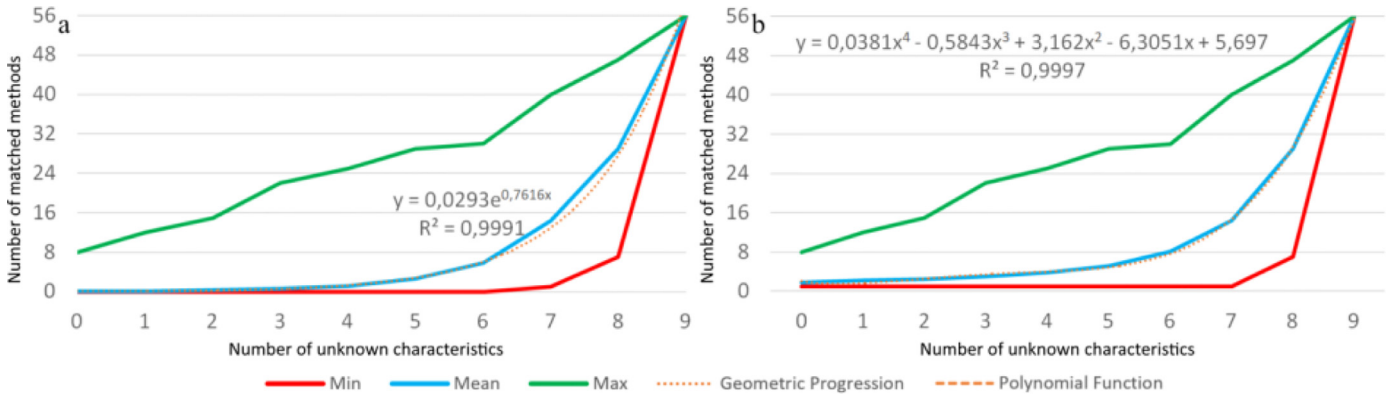


Fig. 4. Minimal, mean and maximal number of matching methods depending on the number of unknown characteristics for a 3-level hierarchy of classifiers, including (a) and excluding (b) the rules returning empty sets of methods.

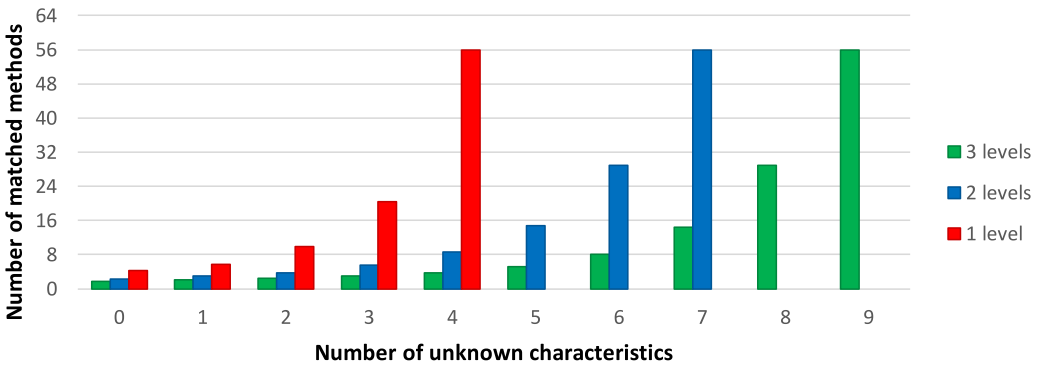


Fig. 5. Average number of methods returned by the rules on each level of classifiers' hierarchy depending on the number of unknown classifiers.

increase of the matching methods' count can be observed along with the increase of the number of unknown method properties. Similarly to the 3-level case, a significant increase of the number of methods can be observed at the expense of the number of rules.

In case of the single-level sequence of classifiers (Fig. 3a), the decision maker needs to take into account that even a single unknown value of the decision problem classifier results in a numerous set of rules and the count of rules ranging from 1 to as much as 30. It should be noted, however, that out of all 30 rules for a single-unknown scenario, only 7 of the rules stand out by providing 10 or more methods (19.5714 in average), whereas for the remaining rules 3 is the mode of the set. If the decision maker cannot produce two values of the methods' properties, the average number of methods returned grows to 9.8824. Moreover, if the DM can produce only a single value of the MCDA methods' classifiers, the average number of methods produced grows to 20.3636 with the minimum value of 1 and maximum value of 47 methods.

The aforementioned facts confirm that the introduction of additional levels to the rule set hierarchy largely increases its precision (Fig. 5), with correlation between the rule sets remaining at the very high level of 0.97 between 1-level and 2-level set or between 1-level and 3-level and 1 between the 2-level and 3-level set. This confirms the fact that the three rule sets can be used interchangeably, when a higher precision is expected.

Moreover, to verify the decision support abilities of the proposed framework, a methods' selection algorithm was developed (Supplementary material – Section 7). A prototype version of an online application supporting the MCDA method selection process handling uncertainty in decision problem description was implemented and is available at dedicated website at http://www.mcda.it.

4. Practical confirmation of the framework

The verification of the proposed framework was conducted with the use of decision making problems from the area of sustainable transport and logistics. The area of applications of multi criteria decision analysis methods is widely discussed in relation to sustainability assessment [30]. The focus of decision support systems dedicated for sustainable logistics is emphasized as well as the need for proper methods selection [34]. Various applications of the MCDA methods area observed in this field due to many conflicting criteria in the area of sustainable transport development [105], sustainable logistics practices [106], green supplier selection [107], green supply chain management [94], selection of transport technologies [108,109] or alternative fuels evaluation [110]. Cinelli et al. [30] emphasizes that various information types including uncertain parameters are required to perform sustainability assessment. To perform validation of proposed framework the set of reference cases of MCDA usage from above areas was prepared and is presented in Table 5. The examples include mainly problems of sustainable supplier selection, supply chain management, location choice, performance evaluation in green SCM, transport infrastructure design, alternative fuels selection, reverse logistics, sustainability assessment of urban systems with the focus on innovations. Each reference case is treated as expert recommendation to solve particular problem with specific MCDA method and is compared with the result delivered by the proposed framework. The obtained results are presented in Table 5.

Table 5 shows a high level of conformity between a selected MCDA method (literature source) and the results obtained by using the proposed framework. A few considered cases need additional clarification. This means a situation, where the proposed framework cannot assign any MCDA method (seven cases) or the characteristic of the considered problem causes that the wrong method is chosen (two cases). The framework does not return any methods for cases number: 5 [95], 12 [92], 16 [89], 29 [121], 32 [110], 33 [123] and 36 [125] (see Table 5):

- In cases 5 [95] and 16 [89], the AHP method was used only to determine relative importance of individual criteria and not to compare decision variants. A value greater than 0 is returned only by properties m_{i1}, and $m_{i1.1}$. All properties referring to decision variants and their comparisons obtain the value 0, including the most basic property determining whether decision variants are compared.
- As for the case 29 [121] the problem is that in this case, authors use equal weights to all the criteria of what was formally interpreted as a lack of weight. Properties m_{i1} and $m_{i1.1}$ returned the value 0 instead of 1 and 2 respectively. Accordingly, the rule base fails to identify a suitable method, since it does not cover the situation where the weights are not used within the methods where it is possible to assign them.
- For cases 32 [110] and 33 [123], in which the AHP method was used to solve a problem, lack of choice of the method results from determining weights of criteria in an uncharacteristic manner. In the article [110], a sensitivity analysis was carried out and weights of criteria were expressed on a percentage scale. On the other hand, in case 33 [123], weights were attributed directly by the decision-maker and expressed on a point scale. In these cases, pairwise comparison matrices (along with a nine-degree Satty's scale) were not used in order to determine weights of criteria, the weights were not relative. Consequently, the property $m_{i1.1}$ returned the value 2 instead of 3.
- In Norese and Carbone [125] (case 36), criterial performances were qualitative not quantitative. That is why, property m_{i2} assumed the value 1 instead of 2. As a result, for a given decision problem, a set of methods $\{E_T\}$ was not assigned.
- In example 12 [92] the wrong adjustment stems from the fact that the authors failed to use indifference and preferences

Table 5
Practical verification of decision rules with respect to the use of referential sources.

No.	Particular MCDA problem descriptors									Description of the problem	The used MCDA method	The activated rule	Recommended subset of MCDA methods	Reference
	c_1	$c_{1.1}$	c_2	c_3	$c_{3.1}$	$c_{3.1.1}$	$c_{3.1.2}$	c_4	$c_{4.1}$					
1	1	2	2	1	2	0	1	1	0	A model of selection the best innovation policies based on a number of criteria reflecting sustainability issues.	$\mathbf{E_S}$	R_{17}	$\{T_C\}$	[101]
2	1	3	2	1	1	1	0	3	2	An integrated approach of fuzzy analytical hierarchy process (fuzzy AHP) and TOPSIS in evaluating the performance of global third party logistics service providers.	$\mathbf{A_F + T_P}$	R_{27}	$\{\mathbf{A_F + T_P}\}$	[63]
3	1	3	3	0	0	0	0	3	2	Green supplier selection for an automobile manufacturing firm using AHP method.	$\mathbf{A_H}$	R_{30}	$\{\mathbf{A_H}, A_N, M_B, D_M, R_M\}$	[107]
4	1	3	3	0	0	0	0	3	2	A performance evaluation model for the operations of the supply chain of an organization of the refrigeration equipment sector.	$\mathbf{M_B}$	R_{30}	$\{A_H, A_N, \mathbf{M_B}, D_M, R_M\}$	[111]
5	1	3	0	0	0	0	0	0	0	The aim of the study is to investigate and to rank the pressures for GSCM based on experts' opinion using an Analytical Hierarchy Process (AHP) in the mining and mineral industry context.	$\mathbf{A_H}$	–	$\emptyset$	[95]
6	1	3	3	0	0	0	0	3	2	Choice of strategy for dealing with defective equipment in reverse logistics.	$\mathbf{A_N}$	R_{30}	$\{A_H, \mathbf{A_N}, M_B, D_M, R_M\}$	[112]
7	1	3	3	1	1	3	0	3	2	Assessment of alternative suppliers for a business.	$\mathbf{A_{NF}}$	R_{31}	$\{A_F, \mathbf{A_{NF}}\}$	[113]
8	1	2	2	1	2	0	3	3	2	Evaluation of energy business cases implemented in the North Sea Region and strategy recommendations using PROMETHEE II method.	$\mathbf{P_2}$	R_{21}	$\{\mathbf{P_2}\}$	[114]
9	1	3	3	0	0	0	0	3	2	Choice of scenario for changes of used fuel for transportation.	$\mathbf{A_H}$	R_{30}	$\{\mathbf{A_H}, A_N, M_B, D_M, R_M\}$	[115]
10	1	3	3	0	0	0	0	3	2	Choice of urban bypass project.	$\mathbf{A_H}$	R_{30}	$\{\mathbf{A_H}, A_N, M_B, D_M, R_M\}$	[94]
11	1	3	2	0	0	0	0	3	2	Choice of fuel for public transport vehicles.	$\mathbf{A_H + T_P}$	R_{26}	$\{\mathbf{A_H + T_P}, A_H + V_K\}$	[56]
12	0	0	1	0	0	0	0	3	1	Ranking of logistics platforms.	$\mathbf{E_4}$	–	$\emptyset$	[92]
13	1	3	2	1	1	3	0	3	2	Ranking of knowledge management solutions adopted in supply chain management.	$\mathbf{A_F + T_F}$	R_{29}	$\{\mathbf{A_F + T_F}, A_{NF} + T_F\}$	[55]
14	1	2	2	1	1	3	0	3	2	The choice of location for urban distribution centers.	$\mathbf{T_F}$	R_{16}	$\{S_F, \mathbf{T_F}, V_F\}$	[91]
15	1	2	2	1	1	3	0	3	2	A multi-criteria framework for comparative assessment of energy technologies in road transport taking into account technologies in terms of their environmental and economic impacts.	$\mathbf{T_F}$	R_{16}	$\{S_F, \mathbf{T_F}, V_F\}$	[108]
16	1	3	0	0	0	0	0	0	0	The decision-making model handling relationships between GSCM practices and performances based on DEMATEL method.	$\mathbf{D_M}$	–	$\emptyset$	[89]
17	1	3	2	0	0	0	0	3	2	MCDA based system for the best municipal solid waste management scenario selection.	$\mathbf{A_H + V_K}$	R_{26}	$\{A_H + T_P, \mathbf{A_H + V_K}\}$	[88]
18	1	3	3	1	1	3	0	3	2	Performance measurement of reverse logistics for the battery manufacturer.	$\mathbf{A_F}$	R_{31}	$\{\mathbf{A_F}, A_{NF}\}$	[97]
19	1	2	2	1	1	3	0	3	2	Supplier choice of equipment from the customer to the manufacturer (reverse supply chain).	$\mathbf{T_F}$	R_{16}	$\{S_F, \mathbf{T_F}, V_F\}$	[116]
20	1	3	3	0	0	0	0	3	2	Applying the analytic hierarchy process to the offshore outsourcing location decision.	$\mathbf{A_H}$	R_{30}	$\{\mathbf{A_H}, A_N, M_B, D_M, R_M\}$	[117]

(*continued on next page*)

Table 5 (*continued*)

No.	Particular MCDA problem descriptors									Description of the problem	The used MCDA method	The activated rule	Recommended subset of MCDA methods	Reference
	c_1	$c_{1.1}$	c_2	c_3	$c_{3.1}$	$c_{3.1.1}$	$c_{3.1.2}$	c_4	$c_{4.1}$					
21	1	2	2	1	1	3	0	3	2	Assessment of balanced supplier performance (Green SCM).	$\mathbf{T_F}$	R_{16}	$\{S_F, \mathbf{T_F}, V_F\}$	[9]
22	1	2	2	1	2	0	3	3	1	Evaluation of alternative transport solutions for the urban transport system.	$\mathbf{E_3}$	R_{20}	$\{\mathbf{E_3}, P_1\}$	[93]
23	0	0	2	0	0	0	0	1	0	Supply chain optimization.	$\mathbf{G_P}$	R_4	$\{\mathbf{G_P}\}$	[69]
24	1	2	1	0	0	0	0	1	0	Evaluation of the performance of national transport systems in terms of impact on the economy, environment and society.	$\mathbf{E_1}$	R_{11}	$\{\mathbf{E_1}\}$	[96]
25	1	3	3	0	0	0	0	3	2	A decision to build a second airport in the metropolis.	$\mathbf{A_H}$	R_{30}	$\{\mathbf{A_H}, A_N, M_B, D_M, R_M\}$	[118]
26	1	2	2	1	1	3	0	3	2	Decision-making model in reverse logistics addressing green issues.	$\mathbf{V_F}$	R_{16}	$\{S_F, T_F, \mathbf{V_F}\}$	[119]
27	1	2	2	1	1	3	0	3	2	Sustainability assessment of urban transportation systems under uncertainty.	$\mathbf{T_F}$	R_{16}	$\{S_F, \mathbf{T_F}, V_F\}$	[98]
28	1	3	2	1	1	3	0	3	2	Green supplier evaluation.	$\mathbf{A_{NF} + T_F}$	R_{29}	$\{A_F + T_F, \mathbf{A_{NF} + T_F}\}$	[120]
29	0	0	2	1	2	0	3	2	0	A two-phase decision-making model integrating design and management of a Supply Chain from an outcome-driven perspective.	$\mathbf{E_T}$	–	∅	[121]
30	1	2	2	1	2	0	3	3	2	Selection of the best sustainable concept using PROMETHEE II method.	$\mathbf{P_2}$	R_{21}	$\{\mathbf{P_2}\}$	[99]
31	1	3	2	0	0	0	0	3	2	A benchmarking framework evaluating the cold chain performance of a company.	$\mathbf{A_H + T_P}$	R_{26}	$\{\mathbf{A_H + T_P}, A_H + V_K\}$	[122]
32	1	2	3	0	0	0	0	3	2	An evaluation of alternative fuels for the road transport sector taking into account cost and policy criteria.	$\mathbf{A_H}$	–	∅	[110]
33	1	2	3	0	0	0	0	3	2	MCDA based methodology for the Flanders in Action Process.	$\mathbf{A_H}$	–	∅	[123]
34	1	3	3	0	0	0	0	3	2	An integrated balanced scorecard (BSC) and analytical hierarchy process (AHP) approach for supply chain management (SCM) performance evaluation.	$\mathbf{A_H}$	R_{30}	$\{\mathbf{A_H}, A_N, M_B, D_M, R_M\}$	[100]
35	1	3	3	0	0	0	0	3	2	An integrated model to evaluate Green Supply Chain's environmental performance.	$\mathbf{A_N}$	R_{30}	$\{A_H, \mathbf{A_N}, M_B, D_M, R_M\}$	[124]
36	1	2	1	1	2	0	3	2	0	An application of ELECTRE Tri to evaluate the airports' innovation.	$\mathbf{E_T}$	–	∅	[125]
37	1	2	2	0	0	0	0	3	2	MCDA based tool supporting selection of the best biowaste management alternatives for stakeholders.	$\mathbf{T_P}$	R_{14}	$\{E_V, M_U, M_V, S_A, S_M, \mathbf{T_P}, U_T, V_K\}$	[126]
38	1	3	2	0	0	0	0	3	2	Sustainability-focused decision support system for supplier selection.	$\mathbf{A_H + V_K}$	R_{26}	$\{A_H + T_P, \mathbf{A_H + V_K}\}$	[127]
39	1	3	2	1	1	2	0	3	2	A fuzzy hierarchical TOPSIS based approach to evaluate different green initiatives and assess improvement areas when implementing green initiatives.	$\mathbf{A_H + T_F}$	R_{28}	$\{\mathbf{A_H + T_F}\}$	[128]
40	1	2	2	1	2	0	3	3	2	Logistics centre location choice.	$\mathbf{E_3}$	R_{21}	$\{P_2\}$	[129]

thresholds, even though it is possible for the ELECTRE IV. Therefore, the properties m_{i3}, $m_{i3.1}$, *and* $m_{i3.1.2}$ were equal to 0.

Referring to the examples in which proposed framework selects a different method than the authors of the reference publication, we must consider the following examples: 1 [101] and 40 [129] (see Table 5):

- In the example 40 [129], the authors stated that they had received a total order of variants, while the ELECTRE III method allows only for a partial order. The property $m_{i4.1}$ mistakenly assumes a value of 2, so that, instead of the ELECTRE III method, PROMETHEE II is adapted (rule R_{21} rather than R_{20}).
- The last situation, where used framework mistakenly matched the MCDA method applies to the example in row number 1 [101] from Table 5. The invalid assignment is caused by the fact that formally the authors used only the indifference threshold (q and p thresholds were equal). Therefore, the property $m_{i3.1.2}$ obtained value 1 instead of 3 and, consequently, rule R_{17} was activated instead of rule R_{18}.

The study demonstrated usefulness of the proposed set of rules for selection of MCDA methods for real applications. The resulting accuracy of the selection is satisfactory. It should be noted that the occurrence of the missing or incorrect classifications was not a shortcoming of the proposed framework, but rather inadequacy of the problem analysis, often resulting from the inappropriate use of MCDA methods. Such applies even to popular methods such as AHP, where the publicly available algorithms were not always correctly implemented.

5. Summary

The selection of an MCDA method suitable for solving a specific decision problem is a vital element of the decision-making process. It is closely related to the issue of striving for the objectification of the decision support process itself [37]. A review of the literature confirms the dilemmas of researchers, and reveals that different methods are used for similar decision problems, leading to difficulties when comparing the results. The problem is addressed in various up-to-date studies, and the authors' conclusions clearly indicate the need for further studies in the search for generalized solutions independent of the current areas of usage of MCDA methods. Like it was showed, earlier approaches focused on MCDA selection problems only partially covered the problem because of limited set of considered methods and assumed precisely defined characteristics of the decision problem. Real situations are usually based on uncertain inputs not only at the level of detailed values of parameters but at more general specifications of decision problem as well.

The presented article contains a successful attempt to build a generalized MCDA method selection framework. The large-scale literature review allowed to extract a set of 56 up-to-date MCDA methods. Their profound analysis made it possible to identify sets of properties and to build on their basis a complete taxonomy of MCDA methods. It constituted the foundations for a formal presentation of the framework of the MCDA methods' selection in the form of a decision tree and a set of descriptors, as well as for the extraction of the set of decision rules. The presented framework for the selection of an MCDA method is based on the identified set of properties of the multi-criteria decision problem. The properties, which relate to decision problematics, comparison of variants, characterisation of weights, performance of alternatives, fuzzy data representation and aspects of imprecise in decision-makers' preferences, are used as a basis for the MCDA method selection.

The proposed framework constitutes formal guidelines for the selection of a particular MCDA method which is independent of the problem domain. While the earlier approaches for method selection take into account a limited number of methods, the current study uses the complete set of solutions available to date. It was shown in the research, that the hierarchical representation of the set of descriptors allowed the selection of MCDA methods also in situations of limited knowledge about the decision problem. The inclusion of uncertainty of the input data to the MCDA method selection rules in the presented approach allowed to address the issue of lack of knowledge in the description of the decision-making process. The modelling of the complete uncertainty space as a part of the proposed approach, enabled the analysis of the impact of the number of missing input data on the final form of the set of the recommended MCDA methods. These studies, along with the proposed framework, were also the basis for transferring the proposed solution to the public scope in the form of a complete, responsive, publicly available expert system supporting the selection of MCDA methods. The solution is available at http://www.mcda.it. When analyzing the effectiveness of the proposed approach, it is worth pointing out that in the empirical research the accuracy of recommendation of particular MCDA methods for a given decision-making situation was satisfactory.

Concluding, the main contributions of the work include:

- a generalised MCDA method selection framework for decision problems with theoretical background and wide applicability,
- the formal presentation of decision rules for MCDA method selection with the potential for direct application,
- guidelines for practitioners and a set of rules applicable in different areas of multicriteria decision making,
- hierarchical process of gathering knowledge about the decision problem from the decision maker,
- a complete analysis of the uncertainty influence on the final set of recommended MCDA methods,
- an algorithm for MCDA methods selection, as well as its implementation in a form of a web-based expert system.

In general, the presented framework provides a basis for construction of a knowledge database containing the rules for selection of a specific MCDA method from a set of all defined options, on the basis of detailed characteristics of the problem. The framework has some limitations, however. The presented set of rules in its current form is not always able to recommend a specific MCDA method. It can only propose a selection of potential methods. Additionally, due to the fact that the framework is based on the formal characteristics of a decision-making situation, the decision-making situation context aspect is omitted [20]. Consequently, the additional factors influencing the MCDA method selection for a given decision-making situation, such as the analyst's familiarity of the methods or the domain of the decision-making problem, were not studied.

The potential future works include the extension of the current collection of MCDA methods to include group decision making, and expansion of the database of reference cases. An additional challenging task would be the knowledge conceptualization and the construction of an ontology of decision problems and MCDA methods which would be used to select methods on the basis of classification results. This could lead to the development of a complete expert system supporting multi-criteria decision making. All the same, the authors would like to further improve their approach by adding support for more modern methods and adjusting the descriptors to match them. Therefore, the authors encourage and are looking forward to any forms of community input.

Acknowledgment

This work was partially supported by the National Science Centre, Poland, grants no. 2017/25/B/HS4/02172 and 2017/27/B/HS4/01216.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.omega.2018.07.004.

References

[1] Ceballos B, Lamata MT, Pelta DA. A comparative analysis of multi-criteria decision-making methods. Prog Artif Intell 2016;5:315–22.
[2] Siskos E, Askounis D, Psarras J. Multicriteria decision support for global e–government evaluation. Omega 2014;46:51–63.
[3] Garcia S, Cintra Y, Torres R, de CSR, Lima FG. Corporate sustainability management: a proposed multi-criteria model to support balanced decision-making. J Clean Prod 2016;136:181–96.
[4] Stewart TJ. Goal directed benchmarking for organizational efficiency. Omega 2010;38:534–9.
[5] Doumpos M, Zopounidis C. Assessing financial risks using a multicriteria sorting procedure: the case of country risk assessment. Omega 2001;29:97–109.
[6] Zopounidis C, Galariotis E, Doumpos M, Sarri S, Andriosopoulos K. Multiple criteria decision aiding for finance: an updated bibliographic survey. Eur J Oper Res 2015;247:339–48.
[7] Liu J, Liao X, Huang W, Liao X. Market segmentation: A multiple criteria approach combining preference analysis and segmentation decision. Omega 2018.
[8] Podviezko A. Use of multiple criteria decision aid methods in case of large amounts of data. Int J Bus Emerg Mark 2015;7:155.
[9] Govindan K, Khodaverdi R, Jafarian A. A fuzzy multi criteria approach for measuring sustainability performance of a supplier based on triple bottom line approach. J Clean Prod 2013;47:345–54.
[10] Ishizaka A, Nemery P. Multi-criteria decision analysis: methods and software. Chichester, West Sussex, United Kingdom: Wiley; 2013.
[11] Stewart TJ, French S, Rios J. Integrating multicriteria decision analysis and scenario planning—review and extension. Omega 2013;41:679–88.
[12] Ehrgott M, Figueira JR, Greco S, editors. Trends in multiple criteria decision analysis, vol. 142. Boston, MA, US: Springer; 2010.
[13] Greco S. A new PCCA method: IDRA. Eur J Oper Res 1997;98:587–601.
[14] Mardani A, Jusoh A, Zavadskas EK. Fuzzy multiple criteria decision-making techniques and applications – two decades review from 1994 to 2014. Expert Syst Appl 2015;42:4126–48.
[15] Vansnick J-C. On the problem of weights in multiple criteria decision making (the noncompensatory approach). Eur J Oper Res 1986;24:288–94.
[16] Saaty TL, Ergu D. When is a decision-making method trustworthy? Criteria for evaluating multi-criteria decision-making methods. Int J Inf Technol Decis Mak 2015;14:1171–87.
[17] Roy B, Słowiński R. Questions guiding the choice of a multicriteria decision aiding method. Eur J Decis Process 2013;1:69–97.
[18] Roy B, Bouyssou D. Aide multicritère à la décision: méthodes et cas. Paris: Economica; 1993.
[19] Zanakis SH, Solomon A, Wishart N, Dublish S. Multi-attribute decision making: a simulation comparison of select methods. Eur J Oper Res 1998;107:507–29.
[20] Guitouni A, Martel J-M. Tentative guidelines to help choosing an appropriate MCDA method. Eur J Oper Res 1998;109:501–21.
[21] Ulengin F, Ilker Topcu Y, Onsel Sahin S. An artificial neural network approach to multicriteria model selection. In: Köksalan M, Zionts S, editors. Multiple criteria decision making in the new millennium, 507. Berlin, Heidelberg: Springer; 2001. p. 101–10.
[22] Cicek K, Celik M, Ilker Topcu Y. An integrated decision aid extension to material selection problem. Mater Des 2010;31:4398–402.
[23] Wang X, Triantaphyllou E. Ranking irregularities when evaluating alternatives by using some ELECTRE methods. Omega 2008;36:45–63.
[24] Peng Y, Wang G, Wang H. User preferences based software defect detection algorithms selection using MCDM. Inf Sci 2012;191:3–13.
[25] Chang Y-H, Yeh C-H, Chang Y-W. A new method selection approach for fuzzy group multicriteria decision making. Appl Soft Comput 2013;13:2179–87.
[26] Kolios A, Mytilinou V, Lozano-Minguez E, Salonitis K. A comparative study of multiple-criteria decision-making methods under stochastic inputs. Energies 2016;9:566.
[27] Figueira JR, Mousseau V, Roy B. ELECTRE methods In multiple criteria decision analysis. In: Greco S, Ehrgott M, Figueira JR, editors. State of the art surveys. New York: Springer-Verlag; 2016. p. 155–85.
[28] Guitouni A, Martel J-M, Bélanger M. A multiple criteria aggregation procedure for the evaluation of courses of action in the context of the Canadian airspace protection; 2001.
[29] Bana e Costa CA, Vincke P. Multiple criteria decision aid: an overview. In: Bana e Costa CA, editor. Readings in multiple criteria decision aid. Berlin, Heidelberg: Springer; 1990. p. 3–14.
[30] Cinelli M, Coles SR, Kirwan K. Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment. Ecol Indic 2014;46:138–48.
[31] Buchholz T, Rametsteiner E, Volk TA, Luzadis VA. Multi criteria analysis for bioenergy systems assessments. Energy Policy 2009;37:484–95.
[32] Celik M, Topcu YI. Analytical modelling of shipping business processes based on MCDM methods. Marit Policy Manag 2009;36:469–79.
[33] Hanne T. Meta decision problems in multiple criteria decision making. In: Gal T, Stewart TJ, Hanne T, editors. Multicriteria decision making, 21. Boston, MA, US: Springer; 1999. p. 147–71. vol.
[34] Qaiser FH, Ahmed K, Sykora M, Choudhary A, Simpson M. Decision support systems for sustainable logistics: a review and bibliometric analysis. Ind Manag Data Syst 2017;117:1376–88.
[35] Roy B. Paradigms and challenges. In: Figueira J, Greco S, Ehrgott M, editors. Multiple criteria decision analysis: state of the art surveys, 78. New York: Springer-Verlag; 2005. p. 3–24.
[36] Kodikara PN. Multi-objective optimal operation of urban water supply systems Ph.D thesis. Victoria University; 2008.
[37] Roy B. Multicriteria methodology for decision aiding. Boston, MA: Springer; 1996.
[38] Jacquet-Lagreze E, Siskos Y. Preference disaggregation: 20 years of MCDA experience. Eur J Oper Res 2001;130:233–45.
[39] Martel J-M, Matarazzo B. Other outranking approaches. Multiple criteria decision analysis: state of the art surveys, 78. New York: Springer-Verlag; 2005. p. 197–259.
[40] Saaty TL. The analytic hierarchy process: planning, priority setting, resource allocation. New York, London: McGraw-Hill International Book Co; 1980.
[41] Öztürké M, Tsoukiàs A, Vincke P. Preference modelling. Multiple criteria decision analysis: state of the art surveys, 78. New York: Springer-Verlag; 2005. p. 27–59.
[42] Celik M, Deha Er I. Fuzzy axiomatic design extension for managing model selection paradigm in decision science. Expert Syst Appl 2009;36:6477–84.
[43] Kornyshova E, Salinesi C. MCDM techniques selection approaches: state of the art. IEEE; 2007. p. 22–9.
[44] Li Y, Thomas MA. A Multiple Criteria Decision Analysis (MCDA) software selection framework. IEEE; 2014. p. 1084–94.
[45] Özcan T, Çelebi N, Esnaf Ş. Comparative analysis of multi-criteria decision making methodologies and implementation of a warehouse location selection problem. Expert Syst Appl 2011;38:9773–9.
[46] Salminen P, Hokkanen J, Lahdelma R. Comparing multicriteria methods in the context of environmental problems. Eur J Oper Res 1998;104:485–96.
[47] Yeh C-H. A problem-based selection of multi-attribute decision-making methods. Int Trans Oper Res 2002;9:169–81.
[48] Al-Shemmeri T, Al-Kloub B, Pearman A. Model choice in multicriteria decision aid. Eur J Oper Res 1997;97:550–60.
[49] Ozernoy VM. Choosing the “Best” multiple criterlv decision-making method. INFOR: Inf Syst Oper Res 1992;30:159–71.
[50] Salinesi C, Kornyshova E. Choosing a prioritization method-case of is security improvement; 2006. Luxembourg p. 51–5.
[51] Guitouni A, Martel J-M, Vincke P. A framework to choose a discrete multicriterion aggregation procedure. Universit Libre de Bruxelles; 2000. SMG.
[52] Kurka T, Blackwood D. Selection of MCA methods to support decision making for renewable energy developments. Renew Sustain Energy Rev 2013;27:225–33.
[53] Chen C-T, Lin C-T, Huang S-F. A fuzzy approach for supplier evaluation and selection in supply chain management. Int J Prod Econ 2006;102:289–301.
[54] Bellman RE, Zadeh LA. Decision-making in a fuzzy environment. Manag Sci 1970;17 B-141–B-164.
[55] Patil SK, Kant R. A fuzzy AHP-TOPSIS framework for ranking the solutions of Knowledge Management adoption in Supply Chain to overcome its barriers. Expert Syst Appl 2014;41:679–93.
[56] Tzeng G-H, Lin C-W, Opricovic S. Multi-criteria analysis of alternative-fuel buses for public transportation. Energy Policy 2005;33:1373–83.
[57] Van Horenbeek A, Pintelon L. Development of a maintenance performance measurement framework – using the analytic network process (ANP) for maintenance performance indicator selection. Omega 2014;42:33–46.
[58] De Keyser WSM, Peeters PHM. Argus — a new multiple criteria method based on the general idea of outranking. In: Paruccini M, editor. Applying multiple criteria aid for decision to environmental management, 3. Dordrecht: Springer Netherlands; 1994. p. 263–78.
[59] Faizi S, Rashid T, Sałabun W, Zafar S, Wątróbski J. Decision making with uncertainty using hesitant fuzzy sets. Int J Fuzzy Syst 2018;20:93–103.
[60] Corrente S, Figueira JR, Greco S, Słowiński R. A robust ranking method extending ELECTRE III to hierarchy of interacting criteria, imprecise weights and stochastic analysis. Omega 2017;73:1–17.
[61] Voogd H. Multicriteria evaluation with mixed qualitative and quantitative data. Environ Plan B: Plan Des 1982;9:221–36.
[62] Ahn BS. The analytic hierarchy process with interval preference statements. Omega 2017;67:177–85.
[63] Kumar P, Singh RK. A fuzzy AHP and TOPSIS methodology to evaluate 3PL in a supply chain. J Model Manag 2012;7:287–303.
[64] Promentilla MAB, Furuichi T, Ishii K, Tanikawa N. A fuzzy analytic network process for multi-criteria evaluation of contaminated site remedial countermeasures. J Environ Manag 2008;88:479–95.
[65] Dubois D, Prade H, Testemale C. Weighted fuzzy pattern matching. Fuzzy Set Syst 1988;28:313–31.
[66] Ziemba P. NEAT F-PROMETHEE – a new fuzzy multiple criteria decision making method based on the adjustment of mapping trapezoidal fuzzy numbers. Expert Syst Appl 2018;110:363–80.
[67] Dubois D, Prade H. Qualitative possibility theory and its applications to constraint satisfaction and decision under uncertainty. Int J Intell Syst 1999;14(1):45–61.

[68] Opricovic S. Fuzzy VIKOR with an application to water resources planning. Expert Syst Appl 2011;38:12983–90.
[69] Nixon JD, Dey PK, Davies PA, Sagi S, Berry RF. Supply chain optimisation of pyrolysis plant deployment using goal programming. Energy 2014;68:262–71.
[70] Fishburn PC. Exceptional paper—lexicographic orders, utilities and decision rules: a survey. Manag Sci 1974;20:1442–71.
[71] Bana E, Costa CA, Vansnick J-C. MACBETH - an interactive path towards the construction of cardinal value functions. Int Trans Oper Res 1994;1:489–500.
[72] Matarazzo B. Multicriterion analysis of preferences by means of pairwise actions and criterion comparisons (MAPPACC). Appl Math Comput 1986;18:119–41.
[73] Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value tradeoffs. New York: Wiley; 1976.
[74] Hwang C-L, Yoon K. Multiple attribute decision making: methods and applications. Cham: Springer-Verlag; 1981.
[75] Leclercq JP. Propositions d'extension de la notion de dominance en présence de relations d'ordre sur les pseudo-critères : MELCHIOR. Revue Belge de Recherche Opérationnelle, de Statistique et d'Informatique 1984;24:32–46.
[76] Munda G. Multicriteria evaluation in a fuzzy environment: theory and applications in ecological economics. Heidelberg: Physica-Verlag; 1995.
[77] Roubens M. Preference relations on actions and criteria in multicriteria decision making. Eur J Oper Res 1982;10:51–5.
[78] Giarlotta A. Passive and Active Compensability Multicriteria ANnalysis (PACMAN). J Multi-Criteria Decis Anal 1998;7:204–16.
[79] Guitouni A, Martel J-M, Bélanger M, Hunter C. Managing a decision-making situation in the context of the canadian airspace protection. Québec: Faculté des sciences de l'administration de l'Université Laval, Direction de la recherche; 1999.
[80] Matarazzo B. Preference ranking global frequencies in multicriterion analysis (Pragma). Eur J Oper Res 1988;36:36–49.
[81] Corrente S, Greco S, Słowiński R. Multiple criteria hierarchy process with ELECTRE and PROMETHEE. Omega 2013;41:820–46.
[82] Paelinck JHP. Qualitative multiple criteria analysis, environmental protection and multiregional development. Pap Reg Sci 1976;36:59–76.
[83] Hinloopen E, Nijkamp P, Rietveld P. Qualitative discrete multiple criteria choice models in regional planning. Reg Sci Urban Econ 1983;13:77–102.
[84] Edwards W, Newman JR. Multiattribute evaluation. Beverly Hills: Sage Publications; 1982.
[85] Bilbao-Terol A, Arenas-Parra M, Canal-Fernandez V, Antomil-Ibias J. Using TOPSIS for assessing the sustainability of government bond funds. Omega 2014;49:1–17.
[86] Jacquet-Lagreze E, Siskos J. Assessing a set of additive utility functions for multicriteria decision-making, the UTA method. Eur J Oper Res 1982;10:151–64.
[87] Opricovic S, Tzeng G-H. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. Eur J Oper Res 2004;156:445–55.
[88] Vučijak B, Kurtagić SM, Silajdžić I. Multicriteria decision making in selecting best solid waste management scenario: a municipal case study from Bosnia and Herzegovina. J Clean Prod 2016;130:166–74.
[89] Govindan K, Khodaverdi R, Vafadarnikjoo A. Intuitionistic fuzzy based DEMATEL method for developing green practices and performances in a green supply chain. Expert Syst Appl 2015;42:7207–20.
[90] Barfod MB. An MCDA approach for the selection of bike projects based on structuring and appraising activities. Eur J Oper Res 2012;218:810–18.
[91] Awasthi A, Chauhan SS, Goyal SK. A multi-criteria decision making approach for location planning for urban distribution centers under uncertainty. Math Comput Model 2011;53:98–109.
[92] Antún JP, Alarcón R. Ranking projects of logistics platforms: a methodology based on the electre multicriteria approach. Procedia – Soc Behav Sci 2014;160:5–14.
[93] Fierek S, Zak J. Planning of an integrated urban transportation system based on macro – simulation and MCDM/A methods. Procedia - Soc Behav Sci 2012;54:567–79.
[94] Ferrari P. A method for choosing from among alternative transportation projects. Eur J Oper Res 2003;150:194–203.
[95] Mathiyazhagan K, Diabat A, Al-Refaie A, Xu L. Application of analytical hierarchy process to evaluate pressures to implement green supply chain management. J Clean Prod 2015;107:229–36.
[96] Bojković N, Anić I, Pejčić-Tarle S. One solution for cross-country transport-sustainability evaluation using a modified ELECTRE method. Ecol Econ 2010;69:1176–86.
[97] Bansia M, Varkey JK, Agrawal S. Development of a reverse logistics performance measurement system for a battery manufacturer. Procedia Mater Sci 2014;6:1419–27.
[98] Awasthi A, Chauhan SS, Omrani H. Application of fuzzy TOPSIS in evaluating sustainable transportation systems. Expert Syst Appl 2011;38:12270–80.
[99] Vinodh S, Girubha RJ. PROMETHEE based sustainable concept selection. Appl Math Model 2012;36:5301–8.
[100] Sharma MK, Bhagwat R. An integrated BSC-AHP approach for supply chain management evaluation. Meas Business Excell 2007;11:57–68.
[101] Popiolek N, Thais F. Multi-criteria analysis of innovation policies in favour of solar mobility in France by 2030. Energy Policy 2016;97:202–19.
[102] Belton V. Multi-criteria problem structuring and analysis in a value theory framework. In: Gal T, Stewart TJ, Hanne T, editors. Multicriteria decision making, 21. Boston, MA, US: Springer; 1999. p. 335–66.
[103] Geldermann J, Schöbel A. On the similarities of some multi-criteria decision analysis methods: on the similarities of some MCDA methods. J Multi-Criteria Decis Anal 2011;18:219–30.
[104] Edwards W, Barron FH. Smarts and smarter: improved simple methods for multiattribute utility measurement. Organ Behav Hum Decis Process 1994;60:306–25.
[105] Govindan K, Kadziński M, Ehling R, Miebs G. Selection of a sustainable third–party reverse logistics provider based on the robustness analysis of an outranking graph kernel conducted with ELECTRE I and SMAA. Omega 2018.
[106] Yazdani M, Zarate P, Coulibaly A, Zavadskas EK. A group decision making support system in logistics and supply chain management. Expert Syst Appl 2017 2017;88:376–92.
[107] Yu Q, Hou F. An approach for green supplier selection in the automobile manufacturing industry. Kybernetes 2016;45:571–88.
[108] Streimikiene D, Baležentis T, Baležentienė L. Comparative assessment of road transport technologies. Renew Sustain Energy Rev 2013;20:611–18.
[109] Fontes CHDO, Freires FGM. Sustainable and renewable energy supply chain: a system dynamics overview. Renew Sustain Energy Rev 2018;82:247–59.
[110] Tsita KG, Pilavachi PA. Evaluation of alternative fuels for the Greek road transport sector using the analytic hierarchy process. Energy Policy 2012;48:677–86.
[111] Della Bruna Jr E, Ensslin L, Rolim Ensslin S. An MCDA-C application to evaluate supply chain performance. Int J Phys Distrib Logist Manag 2014;44:597–616.
[112] Ravi V, Shankar R, Tiwari MK. Analyzing alternatives in reverse logistics for end-of-life computers: ANP and balanced scorecard approach. Comput Ind Eng 2005;48:327–56.
[113] Tseng M-L, Chiang JH, Lan LW. Selection of optimal supplier in supply chain management strategy with analytic network process and choquet integral. Comput Ind Eng 2009;57:330–40.
[114] Xu B, Nayak A, Gray D, Ouenniche J. Assessing energy business cases implemented in the North Sea Region and strategy recommendations. Appl Energy 2016;172:360–71.
[115] Poh K, Ang B. Transportation fuels and policy for Singapore: an AHP planning approach. Comput Ind Eng 1999;37:507–25.
[116] Kannan G, Pokharel S, Sasi Kumar P. A hybrid approach using ISM and fuzzy TOPSIS for the selection of reverse logistics provider. Resour Conserv Recycl 2009;54:28–36.
[117] Boardman Liu L, Berger P, Zeng A, Gerstenfeld A. Applying the analytic hierarchy process to the offshore outsourcing location decision. Supply Chain Manag 2008;13:435–49.
[118] Zietsman D, Vanderschuren M. Analytic Hierarchy Process assessment for potential multi-airport systems – the case of Cape Town. J Air Transp Manag 2014;36:41–9.
[119] Vahabzadeh AH, Asiaei A, Zailani S. Green decision-making model in reverse logistics using FUZZY-VIKOR method. Resour Conserv Recycl 2015;103:125–38.
[120] Büyüközkan G, Çifçi G. A novel hybrid MCDM approach based on fuzzy DEMATEL, fuzzy ANP and fuzzy TOPSIS to evaluate green suppliers. Expert Syst Appl 2012;39:3000–11.
[121] Álvarez-Socarrás A, Báez-Olvera A, López-Irarragorri F. Two-phase decision support methodology for design and planning an outcome-driven supply chain. J Appl Res Technol 2014;12:704–15.
[122] Joshi R, Banwet DK, Shankar R. A Delphi-AHP-TOPSIS based benchmarking framework for performance improvement of a cold chain. Expert Syst Appl 2011;38:10170–82.
[123] Macharis C, De Witte A, Turcksin L. The Multi-Actor Multi-Criteria Analysis (MAMCA) application in the Flemish long-term decision making process on mobility and logistics. Transp Policy 2010;17:303–11.
[124] Felice FD, Petrillo A, Cooper O. An integrated conceptual model to promote green policies. Int J Innov Sustain Dev 2013;7:333.
[125] Norese MF, Carbone V. An application of ELECTRE tri to support innovation: an application of electre tri to support innovation. J Multi-Criteria Decis Anal 2014;21:77–93.
[126] Pubule J, Blumberga A, Romagnoli F, Blumberga D. Finding an optimal solution for biowaste management in the Baltic States. J Clean Prod 2015;88:214–23.
[127] Luthra S, Govindan K, Kannan D, Mangla SK, Garg CP. An integrated framework for sustainable supplier selection and evaluation in supply chains. J Clean Prod 2017;140:1686–98.
[128] Wang X, Chan HK. A hierarchical fuzzy TOPSIS approach to assess improvement areas when implementing green supply chain initiatives. Int J Prod Res 2013;51:3117–30.
[129] Żak J, Węgliński S. The selection of the logistics center location based on MCDM/A methodology. Transp Res Procedia 2014;3:555–64.

A3.

Karczmarczyk, A., Jankowski, J., Wątróbski, J. (2019, September). Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks. In 2019 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 663-673). IEEE.

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 663–673
DOI: 10.15439/2019F199
ISSN 2300-5963 ACSIS, Vol. 18

Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks

Artur Karczmarczyk*, Jarosław Jankowski* and Jarosław Wątróbski†
*Faculty of Computer Science and Information Technology
West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland
Email: {artur.karczmarczyk,jaroslaw.jankowski}@zut.edu.pl
†Faculty of Economics and Management
University of Szczecin
Mickewicza 64, 71-101 Szczecin, Poland
Email: jwatrobski@usz.edu.pl

***Abstract*—Spreading of information within social media and techniques related to viral marketing take more and more attention from companies focused on targeting audiences within electronic systems. Recent years resulted in extensive research centered around spreading models, selection of initial nodes within networks and identification of campaign characteristics affecting the assumed goals. While social networks are usually based on complex structures and high number of users, the ability to perform detailed analysis of mechanics behind the spreading processes is very limited. The presented study shows an approach for selection of campaign parameters with the use of network samples and theoretical models. Instead of processing simulations on large network, smaller samples and theoretical networks are used. Results showed that knowledge derived from relatively smaller structures is helpful for initialization of spreading processes within the target network of larger size. Apart from agent based modeling, multi-criteria methods were used for evaluation of results from the perspective of costs and performance.**

I. Introduction

Online platforms evolved from early stage technical systems to social media with integrated mechanics of social communication and interactions close to the real world [1]. Together with growing audiences, they attracted more attention of marketers. Apart from typical digital marketing channels based on display advertising and search engines new strategies focused on social media emerged. They include mechanism based on detailed targeting, consumer behavior analysis and commercial content dissemination with the use mechanisms of information spreading.

Results delivered from viral campaigns usually outperform traditional campaigns because of the utilized social influence and ability to induce high dynamics even with low budgets [2]. Social recommendations have high impact on customer decisions and, properly integrated with marketing communication [3], help to further increase performance [4].

The recent studies focused on viral marketing take into account data from real platforms as well as theoretical network models [5]. One of the goals is to increase campaign dynamics and coverage with properly selected initial customers during the seeding process [6]. Apart from static networks, dynamic networks with varying structures are taken into account [7]. Other approaches take into account multi-layer structure of networks representing specifics of real social relations based on different networks, for example private and professional contacts [8].

Theoretical and simulation models are used for prediction of network coverage. They can be derived from analytic models used in epidemiology [9] or can be more focused on network structures and characteristics [10]. Other possibility is to use theories and models related to the diffusion of innovations [11].

While most of the research is focused on coverage and number of infected nodes within the network, from the practical point of view, marketing campaigns can have different goals and specifics. They are planned within assumed budget constraints and timing. A different strategy can be used to acquire high number of potential customers in a very short time than for a long term planning and organic growth of customer database. Campaign budget influences the number of initially infected nodes (seeds) and demographic characteristics. The quality of seeds and their number can be a key factor of campaign coverage and overall results. Additional budgets can be used to increase campaign dynamics or lifespan. To take into account various goals multi-criteria campaign evaluation can be used to select campaign parameters and goals according to preferences and priorities [12]. Earlier research has shown that in order to reduce computational complexity, campaigns can be planned with the use of simulations within smaller synthetic networks based on theoretical model. However, since the theoretical models might not always fit the real networks, the current study proposes the use of network samples for the initial simulations and detection of campaign parameters. Both approaches were compared with results obtained from the complete network and showed the ability to obtain approximate results with network samples.

The paper comprises of five main sections. After this introduction, in Section II literature review is presented. It is followed by the methodology presented within Section III and results in Section IV. Paper is concluded in Section V.

II. Literature review

Social platforms gather detailed information about user behavior and social relations with the main goal to better address commercial messages and properly target products and services [13]. The growing complexity and volumes of the collected data is a direct result of the growing number of users and that their activities moved to electronic systems [14], [15]. Social platforms are treated as tools to use social influence mechanisms to spread information between friends with the impact strengthened by social recommendations. Contacts within social networks are used to pass the information and it often induces information cascades as a main driver of viral marketing campaigns. Multidisciplinary nature of phenomena connected with information diffusion integrates efforts from scientists from various fields like sociology, computer science, physics and management with a different theoretical and practical goals [4] [9] [6].

For better understanding of the information spreading processes, theoretical models are used and they are often implemented within agent based environment or used for analytic studies [16]. Methodological background of studies is often based on models initially created for epidemic research like SIR or SIS with taken into account analytic view on processes and their dynamics [9]. Apart from them, more dedicated solutions were created to create models on microscopic level using information about network structures and relations between users. They are based on two key mechanisms represented by linear threshold models [11] and independent cascades [10]. Linear threshold model, with its later extensions, assumes the social influence induced by neighbors with the network and information flow when the number of neighbors exceeds assumed threshold. Cascading models use different mechanics with spreading based on propagation probabilities and communication with surrounding neighbors and passing content to them. These approaches can be treated as pull and push spreading models. Spreading models can be also used for analysis based on aggregated and macroscopic level [17].

Apart from the mechanics of the information spreading, the dynamics of processes are related to network models and their structures. For the simplest approaches, static networks of non-varying structures are used. More closer to reality are approaches focused on dynamic networks with a changing number of social connections or availability of nodes [18]. For better representation of real systems multi-layer networks are used with spreading dependent on connections between layers, their structures or similarities [8].

Many studies related to information spreading take into account the selection of initial customers, in a form of a seeding process, targeted with product samples or other marketing content with the main goal to motivate them to spread the information to friends within the network [6]. Proper selection of seeds is crucial for successful campaigns, but the problem identified as influence maximization problem is NP-hard [10]. Greedy solutions deliver effective results, but with the high computational cost they are difficult to use within real networks [10]. More practical approaches base on heuristics and a selection of nodes with the use of the network metrics like degree or betweenness. Centrality measures can be used for selection of initial influencers with assumed characteristics [19] [20].

Apart from seeding only once at the beginning of the process, knowledge about the process performance can be gathered and used for additional actions to improve the process characteristics. Adaptive approaches can be used [21] to increase the reach and better utilize the available knowledge. Other possibility is to spread the seeds over the time and better utilize the natural spreading processes. It can be applied in a form of sequential seeding [22] or its extension with recomputed nodes' rankings at every simulation step [23]. Further improvement of seeding can be performed with the use of knowledge about community structures within the network [24], voting mechanics [25] or k-shell based approach dedicated for identification of nodes with high spreading potential [26].

Apart from single campaigns spreading, processes can interact or compete [27]. For such scenarios seeding can be planned to increase the chance of process to survive among competitors or reach audiences in a shortest time before other processes acquire them. Similar situation takes place in epidemic research where two or more pathogens are competing with each other or conditional infections are observed with activity of first virus required for next viruses. Competing scenarios are observed when awareness spreading is deceasing dynamics of epidemic [28]. It lead to extension of the single campaign models to multi-spreading processes for viral marketing studies[29].

Another studies take into account content specifics and network structures [30], proper ways to motivate users to forward the content [31], influence of emotions on content propagation processes [32] [33] and other structural or functional factors [34] [35].

The earlier studies focused mainly on influence maximization to increase coverage within the network. Campaign evaluation was was also discussed as a multi-criteria problem [12]. Campaigns performed within agent based simulation environment were evaluated with the use of set of criteria related to budgets, campaign costs and the number of target nodes. Model output was delivering solutions with defined number of seeds or propagation probabilities. Study also showed the ability to perform simulations with theoretical models and apply selected strategies to real network. The current study extends the presented approach and uses network samples created with the use of snowball sampling [36].

III. Methodology

Viral marketing campaigns can be based on various strategies. During the campaign planning, decisions are taken about

optimal number of initial seeds, methods used for their selection, motivation techniques used for users to increase their willingness to spread the content and type of incentives used to increase the propagation probabilities. Similar problems are related to campaign evaluation and selection of campaign metrics dependent on campaign goals. Other performance metrics can be used for campaigns focused on high network coverage than on highly targeted processes addressed only to specific customers.

While social networks store information about users, connections and network structures, it is possible to analyze information before campaign to optimize the strategy and maximize results. With the assumed campaign scenarios and goals it is possible to simulate and test different strategies for selection of campaign parameters. Due to high computational complexity it would be difficult for larger networks.

The approach proposed in this paper assumes the generation of synthetic networks based on theoretical models, generation of network samples based on real network, performing simulations focused on verification of different seeding strategies and campaign parameters and evaluation of results with the use of MCDA methods and, finally, launching the campaign within the real network (see Fig. 1)

Simulations can be performed within synthetic networks based on theoretical models like Barabasi-Albert model (BA) [37], Watts-Strogatz (WS) model [38] and Erdos-Renyi model (ER) [39]. The size of synthetic networks can be adjusted with reference to the size of real network and it can be a fraction of the real network e.g. 10%, 20%, 30% etc. It is also important to select proper network model with high similarity to real network. The presented approach uses Kullback-Leibler measure (KL) to compare network similarities [40]. Number of nodes and edges within synthetic network can be scaled for better performance and accuracy.

Since a real network not always must be similar to idealized theoretical models, another approach can be based on network samples generated as a fraction of the real network. Snowball sampling can be used to obtain smaller structures, which would allow to perform simulations easier, yet with assumed similarity to the full network structures. Samples can be scaled from lower to higher fraction of the complete network. It is assumed that accuracy of simulations in the bigger samples is more close to the real network but computational complexity is lower for the smaller samples.

The simulations for all samples and synthetic networks are performed with the use of various campaign parameters. The number of seeds represented by the seeding fraction (SF) and its effect on total coverage can be verified and is the representation of a campaign budget. Another decisions are related to seed selection strategy (SS). It can be based on different network metrics and it is also related to campaign costs. For example, targeting high degree nodes can be more expensive than low degree nodes.

From the other point of view, the selection of nodes with high closeness can be more expensive than the selection of nodes with high degree because of higher computational complexity required to compute closeness metrics than degree. Another tested parameters are based on propagation probabilities (PP). For lower propagation probabilities, coverage within the network will be lower, but higher probabilities require higher motivation of users to forward the content. It may require incentives and is related to increased budgets.

To compare results from samples and synthetic networks, the proposed study performs analysis for all networks used. The MCDA module takes into account possible campaign success evaluation criteria like coverage, dynamics, campaign costs. In the subsequent step, the performance table obtained from the samples, as well as the criteria and preferences, are used to produce a ranking of possible advertising strategies with the selected MCDA method. After analyzing the ranking and performing robustness / sensitivity analysis, the analyst provides the campaign parameters recommendation for real network campaign.

In the prior research [12], the authors successfully used the PROMETHEE II method [41], [42] to evaluate viral marketing campaign strategies. However, in the proposed research the authors' wanted to emphasize the effect that the marketers' weights assigned to particular criteria have on the final strategies evaluation. Therefore, it was decided that full sensitivity analysis of the obtained solutions should be performed, which eliminated aspect of uncertainty of the decision maker's criteria preference. Moreover, since in the proposed approach the input data comes from simulations, data uncertainty can also be disregarded. However, the evaluation problem at hand still is characterized by weights and data expressed on a quantitative scale. Last, but not least, the obtained solution to the strategy evaluation problem should take the form of a complete ranking to allow the choice of the best strategy. Therefore, based on the analysis of 65 MCDA methods [43], [44] and the guidelines included in [45] and [46], the authors decided to found their approach on the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method [47].

The TOPSIS method is a representative of the American MCDA school [48] which transforms all decision-making problem criteria into a single score value. In case of the TOPSIS method, based on the criterial performance of the evaluated criteria, a positive and negative ideal strategies are created, i.e. one which tops at each criterion and one that bottoms at all criteria. Subsequently, the score of each appraised strategy is computed as a relative distance between the strategy and both the positive and negative ideal solutions. Therefore, the best strategy would be the one which is closest to the positive ideal strategy, yet as far as possible from the negative ideal strategy in terms of criterial performance values.

IV. Results and Discussion

A. Evaluation of viral marketing campaign strategies on a real network

The empirical study was based on a real network, a part of the topology of the Gnutella network as mapped in 2002 in the

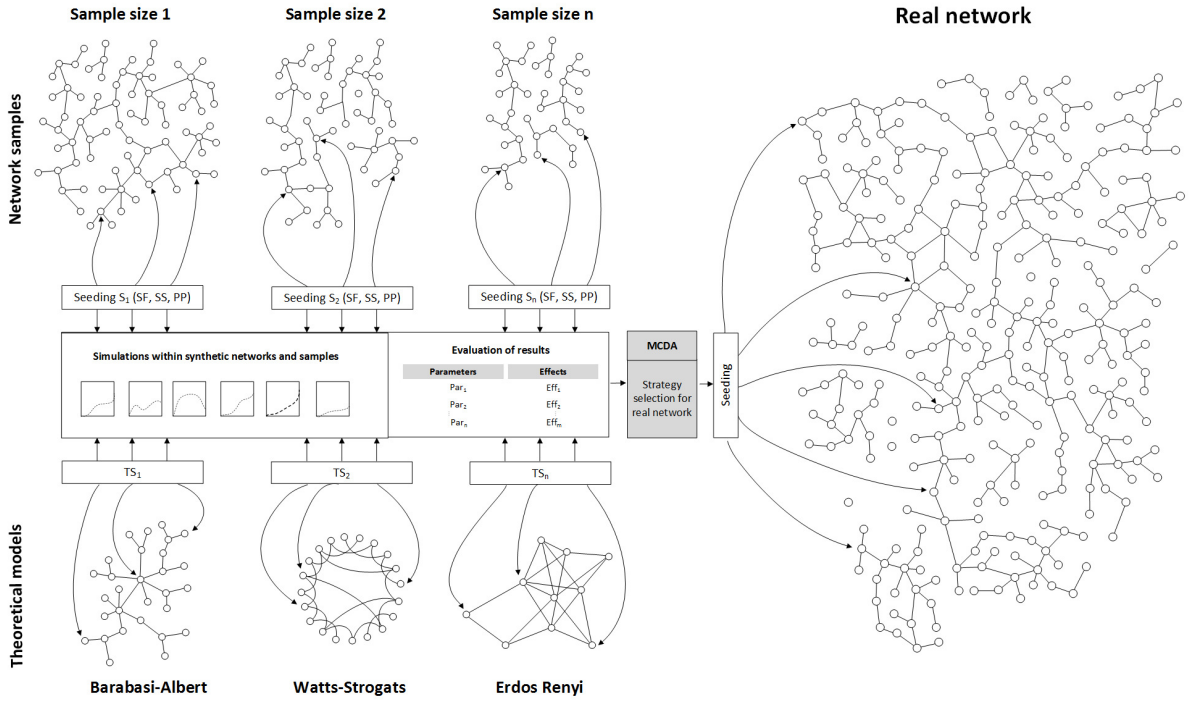


Fig. 1. Conceptual framework for real network strategy selection based on simulation results within network samples and theoretical models

[49] research. The mapped network comprises of 8846 nodes and 31839 edges. The nodes represent hosts in the Gnutella network topology and the edges represents connections between the Gnutella hosts in a single of the network snapshots collected in August 2002. The average values of the main network's metrics are as follows:

1) total degree $D = 7.1985$
2) closeness $C = 1.587441e - 07$
3) Page Rank $PR = 0.0001130454$
4) Eigen Vector $EV = 0.01602488$
5) clustering coefficient $CC = 0.0001130838$
6) betweenness $B = 19104.87$

During the empirical study, the authors used the proposed framework to plan and simulate a viral marketing campaign. Ten simulation scenarios were generated to assure repeatability of the results regardless of the input parameters. Each scenario was composed of the weights drawn for each edge, ranging $< 0; 1 >$. These weights were later compared with the propagation probability of each node to determine whether or not the actual information propagation would occur.

As part of the simulations, a total of 400 sets of parameters were tested, built as a Cartesian product of the following simulation parameter values:

1) Par1 - 0.01, 0.02, ..., 0.09, 0.10;
2) Par2 - 0.01, 0.10, 0.20, ..., 0.90;
3) Par3 - degree (1), closeness (2), eigenvector centrality (3), betweenness (4) – the value is the rank of the method based on its computation speed.

Rank	Alt	CCi	SF	PP	Last Iter	Coverage	Measure
1	A11	0.7494	0.01	0.20	14.4	0.5174	1
2	A10	0.7244	0.01	0.20	14.2	0.5176	2
3	A7	0.7218	0.01	0.10	16.8	0.1334	1
4	A51	0.7148	0.02	0.20	13.2	0.5189	1
5	A50	0.7033	0.02	0.20	13.6	0.5201	2
6	A8	0.6975	0.01	0.10	19.6	0.1127	3
7	A12	0.6960	0.01	0.20	14.7	0.5172	3
8	A6	0.6901	0.01	0.10	15.6	0.1359	2
9	A47	0.6883	0.02	0.10	14.5	0.1625	1
10	A48	0.6878	0.02	0.10	18.9	0.1218	3
11	A52	0.6839	0.02	0.20	14.5	0.5181	3
12	A46	0.6805	0.02	0.10	15.1	0.1638	2
13	A91	0.6799	0.03	0.20	12.2	0.5213	1
14	A88	0.6732	0.03	0.10	18.1	0.1313	3
15	A92	0.6722	0.03	0.20	14.5	0.5194	3
16	A90	0.6635	0.03	0.20	12.4	0.5221	2
17	A128	0.6588	0.04	0.10	17.6	0.1464	3
18	A87	0.6561	0.03	0.10	13	0.1870	1
19	A131	0.6560	0.04	0.20	11.8	0.5230	1
20	A132	0.6537	0.04	0.20	14.3	0.5207	3

Fig. 2. Visualization of the top 20 alternatives from the TOPSIS evaluation of the [49] real network.

Consequently, 4000 simulations were performed for the [49] network. The results of each simulation run were registered, including inter alia the iteration during which the last infection occurred as well as the total coverage achieved, which values were labelled for the further evaluations as Eff4 and Eff5.

After the simulations concluded, the TOPSIS method was

used to evaluate all 400 campaign scenarios. Initially, the weights of all criteria well set equal. The preference direction of the **Par1-Par3** criteria was minimum and of the **Eff4-Eff5** was maximum. Intuitively that would mean the decision maker would prefer low cost of the enterpreneurship, yet long duration and maximum coverage. The top 20 strategies are presented on Fig. 2. The best strategy, A11, obtained ϕ_{net} score of 0.7494. This strategy is based on low values of SF and PP (0.01 and 0.20 respectively) and degree as the method of seeding nodes selection. The runner-up alternative, A10, is based on the same SF and PP values, but uses closeness as the method for selecting the seeding nodes. As a result, slightly broader coverage was achieved in minutely less iterations (0.02s difference). The third-best strategy, A7, maintains the degree measure and the SF of 0.01, however it reduces the PP by half, to 0.10. Such strategy would non-negligibly reduce the costs of the campaign (lower investment in incentives), and, since less nodes at each step would get infected, the procedure would take longer (16.8 iterations on average). However, the obtained coverage is significantly lower, equal to 0.1334 of the network, which is over three-fold worse than the winning A11 strategy.

For the purposes of comparison, the worst strategy, A400, was based on high SF (0.10), high (ignitable) PP (0.9) and eigenvector centrality as the measure. As a result, the contamination process averagely finished within 5.1 iterations, with the mean coverage of 0.9722. Although almost full network gets covered with that strategy, it is important to note that the incentive costs for such strategy would be very high to achieve 90% propagation probability. Also the duration of the campaign would be low, which is against the DM's preferences.

One of the benefits of the TOPSIS method is the fact it allows to build an ideal reference model for the given evaluation problem. In case of the problem at hand, the ideal strategy would be based on degree for selecting the nodes to seed information to and only 1% nodes would be seeded. Incentives would be in place to generate an average propagation probability of 0.01%. With such parameters of the network, the DM would like the outcomes of the marketing campaign to be 19.6 iterations resulting in 97.22% coverage. It is important to note, however, that although ideal, such strategy is only a reference model and does not exist.

The rank presented on Fig. 2 is based on an assumption that the weight of each criterion on the final outcome is equal. However, the DM often gives more significance to some criteria over the others. One of the tremendous benefits of the utilisation of MCDA in the evaluation of viral marketing campaign strategies is the possibility to perform a sensitivity analysis, to learn how even slight changes in preferences of each criterion would affect the final outcome. Therefore, in a subsequent part of the research, a sensitivity analysis was performed to show how the ranking relations between the top 20 alternatives would change if the weights of each criterion would change. The analysis was divided into five parts, one for each criterion. During each phase, the weight of a single criterion was changed from 1 to 100, while the weights of the remaining criteria were set equally to 50.

The results of the sensitivity analysis are presented on Fig. 3. The top row of the figure (A-E) presents how the score of each strategy changed, resulting from each criterion's weight change, whereas the bottom row of the figure (F-J) presents how that change affected the strategies' positions in the ranking. The analysis of Fig. 3A,F shows that no matter how the weight of the criterion Par1 changed, strategy A11 remained the leading one. On the other hand, if the weight of this criterion dropped slightly below 40, strategy A7 would outrun strategy A10. Strategy A51 rank is not affected by the changes of weight of criterion Par1, whilst strategy A50 (ranked fifth) would be outrun by strategy A12 (ranked 7) if its weight was higher than 75. The analysis of the chart on Fig. 3A allows to observe, that while the score of alternatives A128, A131 and A132 is not significantly affected by the changes of Par1 weight, the remaining strategies gain more score as the weight of this criterion increases. A similar tendency can be observed on Fig. 3B, where the scores of all strategies increase along with the increase of significance of criterion Par2. When the weight of that criterion would exceed 90, the runner-up strategy A7 would outrun the strategy A11. An opposite tendency can be observed on Fig. 3E, where all alternatives lose score when the weight of Eff5 grows. Along with this criterion's weight growth, there are only little changes in the order of the three leading alternatives, however, if the weight of that criterion dropped close to 0, the leading strategy A11 would drop six positions to rank 7. This demonstrates the fact that strategy A11 is considerably supported by criterion Eff5. The observation of Fig. 3F-J shows that while for the criteria Par1 and Par2 the majority of rank changes occur when the weight of the criterion changes considerably, in case of criteria Par3 – Eff5, most of the rank changes occur with even minute changes of these criteria's weights.

B. Selection of synthetic networks

As it was presented in section IV-A, the proposed MCDA framework allows to successfully evaluate various viral marketing campaign strategies performed over a real network. However, full networks are rarely available for the entities ordering campaigns. Often, only characteristics of a network are provided. Moreover, running comprehensive simulations on a real networks containing multitude of nodes is also time consuming. Therefore, it is beneficial to perform simulations on smaller synthetic networks before launching the actual campaign on a real network.

Consequently, in the empirical research, apart from evaluating campaign strategies based on full, real network, the authors also used the proposed framework to perform simulations on synthetic networks, similar to the real one, but of a reduced size. The strategies' rankings obtained for synthetic networks were then compared to the ranking obtained for the real network.

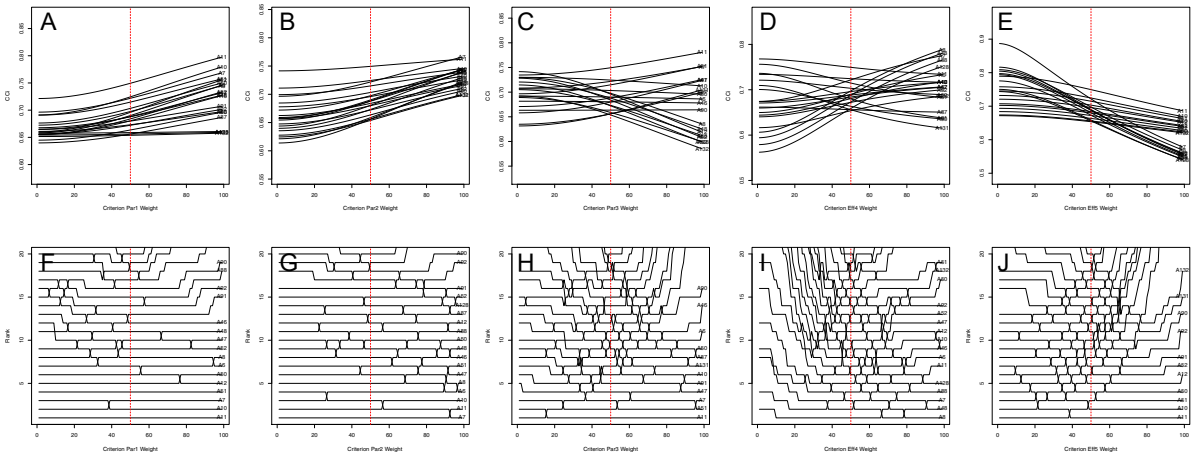


Fig. 3. Ranking sensitivity analysis for the top 20 alternatives from the TOPSIS evaluation of the [49] real network.

For the 10%, 30% and 50% size of the real network, BA, WS and ER networks were generated with the following parameters:

1) BA - number of nodes equal to 10%, 30% and 50% of the real network; number of edges m to add in each step equal to $1, 2, \ldots, 5$ – a total of 15 networks;
2) WS - number of nodes equal to 10%, 30% and 50% of the real network; the neighborhood within which the vertices of the lattice will be connected equal to $1, 2, \ldots, 5$ – a total of 15 networks;
3) ER - number of nodes equal to 10%, 30% and 50% of the real network; number of edges equal to the chosen number of nodes multiplied by $1, 2, \ldots, 5$ – a total of 15 networks.

As a result, a set of 45 networks was generated. In order to avoid arbitrary decisions which network to run the simulations on, the Kullback-Leibler divergence measure was used to compare the degree distribution of all generated networks to the real one. Based on the smallest value of the KLD measure, three networks were selected for further simulations. The selected networks are presented in Table IV-B.

C. Viral marketing campaign strategies planning with synthetic networks

The results of the viral marketing campaign strategies planning with the use of the three aforementioned BA networks is presented in table on Fig. 4. The analysis of the table allows to notice that regardless of the selected network size, in all three cases the same strategy A11 was chosen as the superior one, similarly to the real network case. While in case of the real network this strategy lasted averagely for 14.4 iterations and resulted in 0.5174 coverage, in case of the synthetic networks, the process averagely lasted $10 - 11.2$ iterations (slightly shorter) and resulted in $0.5783 - 0.7049$ coverage (slightly higher). The second best strategy in all three synthetic networks was strategy A51, which above, in case of the real network, was ranked fourth. This strategy is based on small values of SF and PP (0.02 and 0.20 respectively) and lasts averagely in $9.5 - 10.7$ iterations resulting averagely in $0.5783 - 0.7049$ coverage. The measure used here is also degree, as in the winning alternative.

The strategy A10, which for the real network evaluation was ranked second, in case of the synthetic networks reached place 3 for the 50% network and rank 4 for the remaining networks. More interesting is the case of strategy A7. On the real network it is ranked third, for the 30% network it remained at the same ranking position, however, for the 50% network it dropped to the fourth rank, whilst for the 10% network its ranking fell to 15th position. The strategy A7 is characterized by its very low SF and PP values (0.01 and 0.10 respectively) and degree as the measures which makes it one of the cheapest, with maximally extended information propagation process duration, on the cost of small final coverage. The duration of the process is very long for this strategy on the real network and the 30% and 50% networks (16.8, 12.8 and 13 iterations averagely, while the maximum average duration was 19.6, 12.9 and 13.7 iterations respectively). In case of the 10% synthetic network, the average duration is 9.7 iterations and the yielded coverage is lower, equal to 0.1784, which resulted in reduction of the A7's rank.

In case of the strategy A15 which for the 10% network is ranked third, it does not occur on the real network top-twenty list, and on the remaining synthetic networks it is below the first top-ten. This is an interesting difference, which can be further analyzed with the use of the sensitivity analysis (see Fig. 5). In case of the 10% network, the strategy is slightly supported by Par1 criterion. If the weight of criterion Par2 was increased, the strategy A15 would significantly drop in the ranking, down to rank 17. On the other hand, if the weight of the Par3 criterion became insignificant, strategy A15 would be ranked 10th. Regarding the efficiency rankings, Eff5 supports the strategy A15 (rank 11 to rank 1 increase when Eff5 weight increases from 1 to 100) and Eff4 is in conflict with A15 (rank 1 to rank 6 decrease when Eff4 weight increases from 1 to

TABLE I
KULLBACK-LEIBLER DIVERGENCE MEASURE FOR THE SELECTED SYNTHETIC NETWORKS

Expected %	Network	Num. of nodes	Perc. of nodes	Num. of edges	Perc. of edges	KLD
10	BA, $m = 4$	885	0.100045218%	3530	0.110870316%	0.000935498
30	BA, $m = 5$	2654	0.300022609%	13255	0.416313326%	0.000800703
50	BA, $m = 5$	4423	0.5%	22100	0.694117278%	0.000521317

	BA-885-4							BA-2654-5						BA-4423-5						
Rank	Alt	CCi	SF	PP	Last Iter	Coverage	Measure	Alt	CCi	SF	PP	Last Iter	Coverage	Measure	Alt	CCi	SF	PP	Last Iter	Coverage	Measure
1	A11	0.8202	0.01	0.20	10.3	0.5783	1	A11	0.8202	0.01	0.20	10	0.7049	1	A11	0.8190	0.01	0.20	11.2	0.7026	1
2	A51	0.8005	0.02	0.20	9.5	0.5783	1	A51	0.8005	0.02	0.20	9.8	0.7049	1	A51	0.7958	0.02	0.20	10.7	0.7026	1
3	A15	0.8002	0.01	0.30	8.7	0.8129	1	A7	0.8002	0.01	0.10	12.8	0.2567	1	A10	0.7926	0.01	0.20	11.3	0.7026	2
4	A10	0.7905	0.01	0.20	10.3	0.5783	2	A10	0.7905	0.01	0.20	10	0.7049	2	A7	0.7758	0.01	0.10	13	0.2686	1
5	A50	0.7705	0.02	0.20	9.5	0.5783	2	A91	0.7705	0.03	0.20	9.8	0.7050	1	A50	0.7679	0.02	0.20	10.7	0.7026	2
6	A14	0.7705	0.01	0.30	8.7	0.8129	2	A50	0.7705	0.02	0.20	9.8	0.7049	2	A91	0.7647	0.03	0.20	10.2	0.7028	1
7	A91	0.7689	0.03	0.20	8.8	0.5791	1	A47	0.7689	0.02	0.10	11.5	0.2732	1	A6	0.7606	0.01	0.10	13.7	0.2686	2
8	A55	0.7651	0.02	0.30	7.8	0.8129	1	A6	0.7651	0.01	0.10	12.8	0.2567	2	A47	0.7558	0.02	0.10	11.8	0.2831	1
9	A95	0.7454	0.03	0.30	7.7	0.8129	1	A131	0.7454	0.04	0.20	9.7	0.7051	1	A46	0.7430	0.02	0.10	12.4	0.2828	2
10	A19	0.7414	0.01	0.40	7.6	0.9173	1	A90	0.7414	0.03	0.20	9.8	0.7050	2	A131	0.7379	0.04	0.20	10	0.7033	1
11	A131	0.7414	0.04	0.20	8.6	0.5801	1	A46	0.7414	0.02	0.10	11.5	0.2757	2	A12	0.7372	0.01	0.20	11.3	0.7026	3
12	A90	0.7399	0.03	0.20	8.8	0.5786	2	A15	0.7399	0.01	0.30	8.2	0.8867	1	A90	0.7354	0.03	0.20	10.1	0.7027	2
13	A12	0.7369	0.01	0.20	10.6	0.5783	3	A87	0.7369	0.03	0.10	10.3	0.2898	1	A87	0.7225	0.03	0.10	10.6	0.2961	1
14	A54	0.7365	0.02	0.30	7.8	0.8129	2	A12	0.7365	0.01	0.20	10.1	0.7049	3	A15	0.7205	0.01	0.30	8.3	0.8882	1
15	A7	0.7334	0.01	0.10	9.7	0.1784	1	A8	0.7334	0.01	0.10	12.9	0.2583	3	A8	0.7193	0.01	0.10	13.5	0.2681	3
16	A135	0.7233	0.04	0.30	7.7	0.8129	1	A130	0.7233	0.04	0.20	9.6	0.7055	2	A52	0.7156	0.02	0.20	10.7	0.7027	3
17	A59	0.7205	0.02	0.40	7.2	0.9173	1	A55	0.7205	0.02	0.30	7.9	0.8867	1	A130	0.7124	0.04	0.20	10	0.7031	2
18	A94	0.7197	0.03	0.30	7.8	0.8129	2	A171	0.7197	0.05	0.20	9.3	0.7058	1	A55	0.7050	0.02	0.30	8.1	0.8882	1
19	A47	0.7184	0.02	0.10	8.8	0.2058	1	A14	0.7184	0.01	0.30	8.2	0.8867	2	A86	0.7025	0.03	0.10	10.7	0.2960	2
20	A52	0.7177	0.02	0.20	9.7	0.5783	3	A52	0.7177	0.02	0.20	9.8	0.7049	3	A14	0.7008	0.01	0.30	8.4	0.8882	2

Fig. 4. Visualization of the top 20 alternatives from the TOPSIS evaluation of the campaign strategy planning on synthetic networks.

100).

The sensitivity analysis can also provide information about the overall stability of the obtained solution. In case of the 10% network, the ranking is very stable and the A11 strategy either remains on the winning rank or drops to the second position if the weight of Par2 drops below 40%, Par3 drops below 10%, Eff4 drops below 25%. The only significant change occurs for the Eff5 criterion, where A11 would drop to rank 2 if the Eff5's weight increased to over 60% and even further if the weight increased to over 75%. If exclusively Eff5 was considered, the A11 strategy would be ranked 13th.

Similar stability for Par1-Par3 can be observed for the 30% network, however if the weight of Eff4 increased significantly or the weight of Eff5 increased significantly, A11 would be ranked 6th.

Last, but not least, in case of the 50% synthetic network, A11 would remain ranked 1st regardless of Par1 weight, would drop to 2nd position if Par2 had weight exceeding 90 or would drop to 3rd position if Par3 had negligible weight. In case of Eff4, the stability interval of the obtained solution is $0 - 80$, whilst in case of Eff5 the stability interval is $35 - 100$.

D. Viral marketing campaign strategies planning with network samples

As it was stated in the methodology section of this paper, although synthetic networks allow to minimize the computational efforts, their resemblance to the actual real network might be insufficient. Therefore in the subsequent step of the research, the original real network [49] was sampled, resulting in 3 networks containing 10%, 30% and 50% of the original network. The sampling procedure was performed with the *snowball.sampling* R function from the netdep R library [50].

The results of the viral marketing campaign planning based on the real network [49] samples are presented in table on Fig. 6. Contrary to the synthetic networks' results, where the same strategy A11 was best in case of all three networks, in case of the samples of the real network, the rankings are more diversified.

When the 50% network is considered, the best-ranked strategy is the strategy A15, based on very low SF, higher PP (0.30), degree measure mediocre process length (14.1 iterations) and satisfying coverage (0.5075). Strategy A15 is followed by strategy A11, which uses smaller PP (0.20), which resulted in simulations in less dynamic process, leading to extending its duration to 17.9 iterations, but reducing the coverage almost by half, to 0.2685. The third position in the ranking belongs to strategy A55, which is based on 0.02 SF and 0.30 PP and results in efficiency results similar to the leading A15 strategy - 13.3 iterations and 0.5106 coverage respectively. However, the costs of such approach are higher due to the increase of the SF. When the 30% and 10% networks are considered, the A15 strategy is ranked second in the former and sixteenth in the latter, which, as mentioned earlier, is in contrast to the observations made for synthetic BA networks.

The equal-weights TOPSIS analysis was followed by a sensitivity analysis of the top 20 strategies for each of the sampled networks (see Fig. 7). An overall observation of the figures allow to see that the rankings for the 50% and 30% networks are much more stable than in case of the 10% network. To illustrate that fact, one can notice that

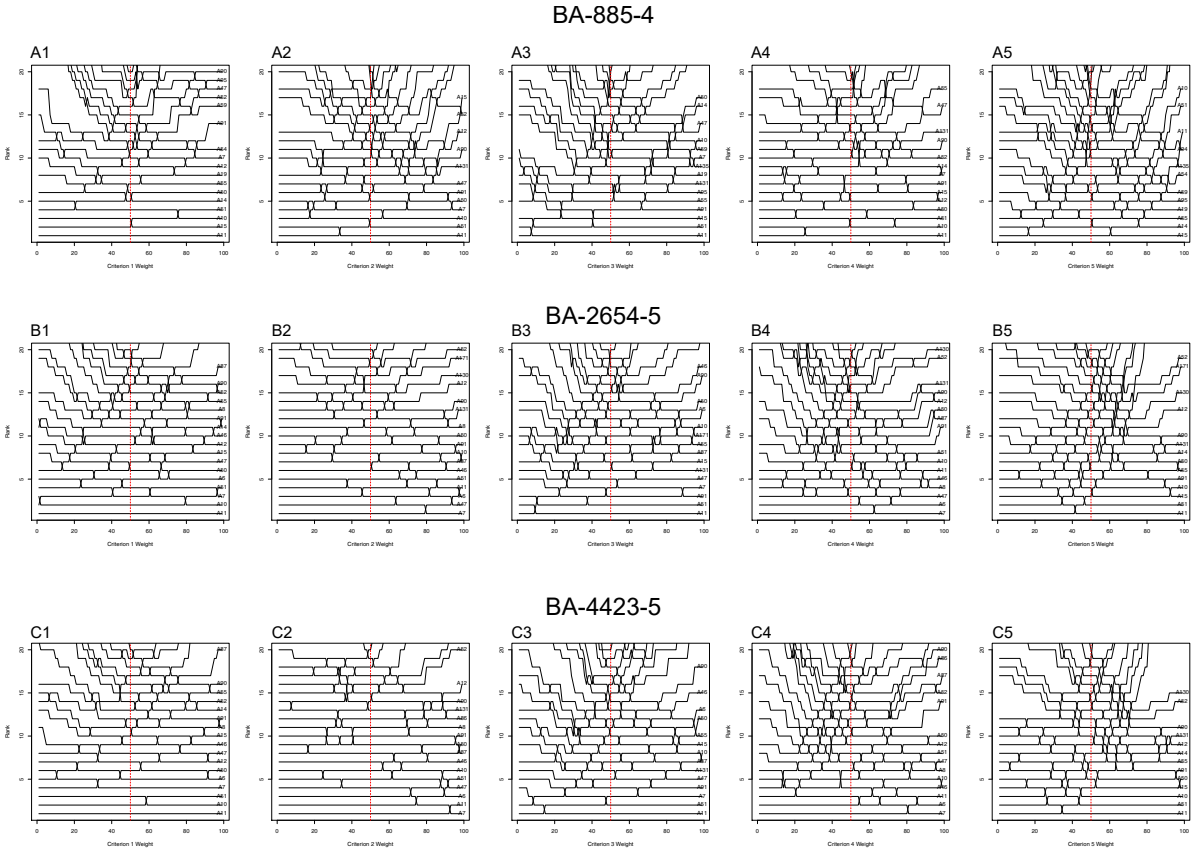


Fig. 5. Ranking sensitivity analysis for the top 20 alternatives from the TOPSIS evaluation of the synthetic networks. A1-A5 – 10% network, B1-B5 – 30% network, C1-C5 – 50% network.

in case of figures C1-C5 and B1-B5 only minute or none changes in the rank of the leading alternative can be observed when the weight of Par1-Eff5 criteria are modified. On the other hand, in case of the 10% network, if Par1 criterion weight was decreased significantly, the leading A23 strategy would drop to position 20 (see Fig. 7A1). Moreover, Fig. 7A2 and A5 demonstrate multiple leader changes in case of even slightest fluctuations of the Par2 and Eff5 criteria. When compared to the stability of the rankings obtained for the actual real network (see Fig. 3), this might suggest that a network obtained as a 10% sample of a real network is too small to maintain the stability of evaluation.

E. Comparison of rankings' evaluation accuracy

The research was concluded by a pairwise comparison of rankings based on equal weights for all analyzed networks. In the comparison, the scores and ranks of all strategies for each network were combined into a single table, ordered by the strategy name. This allowed to obtain correlation matrices for all the networks, presenting how correlated are the ranks (Table IV-E) and scores (Table IV-E) for each pair of networks.

The analysis of the correlation matrices allows to observe that the rankings for BA networks are highly correlated to the ranking for the real network with $0.9390 - 0.9799$ correlation coefficient for scores and $0.9631 - 0.9800$ coefficient for ranks, which means that the relation between them is almost linear. In turn, for the sampled networks, only the ranking for the 50% network achieved high correlation coefficient with the real network, equal to 0.8797 for scores and 0.9222 for ranks. This shows, that the results of the evaluation for the real network and the 50% sampled network are very similar, yet the computational power required to perform the evaluation is significantly smaller. On the other hand, the correlation coefficient values for scores and ranks for the 30% network are much lower, i.e. 0.6043 and 0.6837 respectively, and for the 10% even lower, i.e. 0.4171 and 0.4629 respectively. Such positive yet low values of correlation coefficients indicate there is a positive relation between the rankings obtained for the real network and its 10% and 30% snowball samples. However, the margin of error there might be too high to base the actual campaign on the strategies obtained for such small network samples.

V. Conclusions

Nowadays, when over 45% of the world population are active social media users [51], information spreading in complex social networks begins to bring better results than traditional online advertising campaigns. Online marketers have begun

Snowball Sample 10%								Snowball Sample 30%							Snowball Sample 50%						
Rank	Alt	CCi	SF	PP	Last Iter	Coverage	Measure	Alt	CCi	SF	PP	Last Iter	Coverage	Measure	Alt	CCi	SF	PP	Last Iter	Coverage	Measure
1	A23	0.6298	0.01	0.50	9.1	0.1958	1	A19	0.7257	0.01	0.40	14.5	0.4418	1	A15	0.7521	0.01	0.30	14.1	0.5075	1
2	A27	0.6293	0.01	0.60	9.1	0.2434	1	A15	0.7137	0.01	0.30	16.3	0.2755	1	A11	0.7491	0.01	0.20	17.9	0.2685	1
3	A19	0.6266	0.01	0.40	9	0.1523	1	A59	0.7079	0.02	0.40	13.6	0.4448	1	A55	0.7302	0.02	0.30	13.3	0.5106	1
4	A22	0.6262	0.01	0.50	10.7	0.2018	2	A23	0.7003	0.01	0.50	12.7	0.5560	1	A10	0.7281	0.01	0.20	17.3	0.2748	2
5	A26	0.6213	0.01	0.60	10.2	0.2495	2	A18	0.6990	0.01	0.40	14.1	0.4424	2	A14	0.7258	0.01	0.30	13.9	0.5085	2
6	A18	0.6206	0.01	0.40	10.2	0.1590	2	A55	0.6958	0.02	0.30	14.7	0.2816	1	A51	0.7183	0.02	0.20	14.8	0.2841	1
7	A35	0.6182	0.01	0.80	9.4	0.3411	1	A14	0.6954	0.01	0.30	16.5	0.2781	2	A54	0.7064	0.02	0.30	13.3	0.5109	2
8	A31	0.6159	0.01	0.70	8.4	0.2896	1	A63	0.6832	0.02	0.50	12.1	0.5584	1	A50	0.7034	0.02	0.20	15.2	0.2883	2
9	A67	0.6145	0.02	0.60	8.7	0.2445	1	A99	0.6801	0.03	0.40	12.4	0.4509	1	A95	0.6982	0.03	0.30	12.3	0.5148	1
10	A30	0.6118	0.01	0.70	9.8	0.2959	2	A58	0.6792	0.02	0.40	13.2	0.4461	2	A91	0.6965	0.03	0.20	13.8	0.3017	1
11	A39	0.6101	0.01	0.90	10.1	0.3828	1	A95	0.6765	0.03	0.30	13.6	0.2963	1	A12	0.6912	0.01	0.20	17.3	0.2702	3
12	A34	0.6084	0.01	0.80	10.2	0.3485	2	A22	0.6739	0.01	0.50	12.4	0.5570	2	A16	0.6908	0.01	0.30	14.4	0.5077	3
13	A62	0.6084	0.02	0.50	9.6	0.2087	2	A27	0.6661	0.01	0.60	11.5	0.6388	1	A90	0.6835	0.03	0.20	14.4	0.3022	2
14	A14	0.6083	0.01	0.30	10.2	0.1050	2	A20	0.6658	0.01	0.40	14.8	0.4409	3	A52	0.6813	0.02	0.20	16.9	0.2762	3
15	A63	0.6079	0.02	0.50	8.1	0.1974	1	A103	0.6631	0.03	0.50	11.6	0.5624	1	A56	0.6744	0.02	0.30	13.9	0.5104	3
16	A15	0.6062	0.01	0.30	8	0.1031	1	A54	0.6609	0.02	0.30	13.4	0.2851	2	A19	0.6728	0.01	0.40	10.2	0.6304	1
17	A66	0.6044	0.02	0.60	9.4	0.2547	2	A139	0.6597	0.04	0.40	12.1	0.4573	1	A131	0.6728	0.04	0.20	13.1	0.3126	1
18	A75	0.6041	0.02	0.80	8.9	0.3423	1	A16	0.6588	0.01	0.30	16.4	0.2732	3	A94	0.6721	0.03	0.30	12.2	0.5151	2
19	A107	0.6015	0.03	0.60	8.4	0.2527	1	A62	0.6582	0.02	0.50	12	0.5598	2	A135	0.6715	0.04	0.30	11.8	0.5174	1
20	A102	0.5998	0.03	0.50	9.5	0.2207	2	A98	0.6507	0.03	0.40	12.1	0.4523	2	A92	0.6665	0.03	0.20	16.4	0.2825	3

Fig. 6. Visualization of the top 20 alternatives from the TOPSIS evaluation of the campaign strategy planning on the real network [49] samples.

TABLE II
CORRELATION MATRIX BETWEEN THE RANKS OF EACH OF THE ANALYZED NETWORKS.

Rank	**Real**	**BA-885-4**	**BA-2654-5**	**BA-4423-5**	**SS 10%**	**SS 30%**	**SS 50%**
Real	x	0.9631	0.9794	0.9800	0.4629	0.6837	0.9222
BA-885-4	0.9631	x	0.9840	0.9812	0.4806	0.7760	0.9703
BA-2654-5	0.9794	0.9840	x	0.9980	0.3809	0.6706	0.9289
BA-4423-5	0.9800	0.9812	0.9980	x	0.3647	0.6585	0.9191
SS 10%	0.4629	0.4806	0.3809	0.3647	x	0.8227	0.6159
SS 30%	0.6837	0.7760	0.6706	0.6585	0.8227	x	0.8718
SS 50%	0.9222	0.9703	0.9289	0.9191	0.6159	0.8718	x

TABLE III
CORRELATION MATRIX BETWEEN THE SCORE VALUES OF EACH OF THE ANALYZED NETWORKS.

CCi	**Real**	**BA-885-4**	**BA-2654-5**	**BA-4423-5**	**SS 10%**	**SS 30%**	**SS 50%**
Real	x	0.9390	0.9749	0.9799	0.4171	0.6043	0.8797
BA-885-4	0.9390	x	0.9757	0.9729	0.4954	0.7730	0.9688
BA-2654-5	0.9749	0.9757	x	0.9974	0.3807	0.6373	0.9152
BA-4423-5	0.9799	0.9729	0.9974	x	0.3674	0.6266	0.9049
SS 10%	0.4171	0.4954	0.3807	0.3674	x	0.8204	0.6043
SS 30%	0.6043	0.7730	0.6373	0.6266	0.8204	x	0.8480
SS 50%	0.8797	0.9688	0.9152	0.9049	0.6043	0.8480	x

to invest greater effort into seeding information into social networks and providing incentives to increase the information propagation probability within the networks. These increased efforts have opened the research area for providing evaluation of various social network advertising campaign strategies as well as supporting the process of their planning.

The approach presented in this paper provides a framework for multi-criteria planning of viral marketing campaigns in social networks and their evaluation, in which various preferences and criteria of the marketer are taken into account. The example criteria provided in this paper allow to choose the satisfactory campaign strategy considering the costs related to the seeding of the information and providing incentives to increase its propagation probability in relation to their effect on the process dynamics and obtained coverage.

The authors' contributions in this paper include:

- multi-criteria framework for evaluation of viral marketing campaigns in social networks;
- simulation engine and usage of synthetic network models and real network samples of limited size allowed to provide a viral marketing campaigns planning tool of reduced computational requirements;
- an example set of criteria was provided that allows to choose a satisfactory viral marketing campaign strategy based on multi-criteria consideration of its costs, dynamics and coverage;

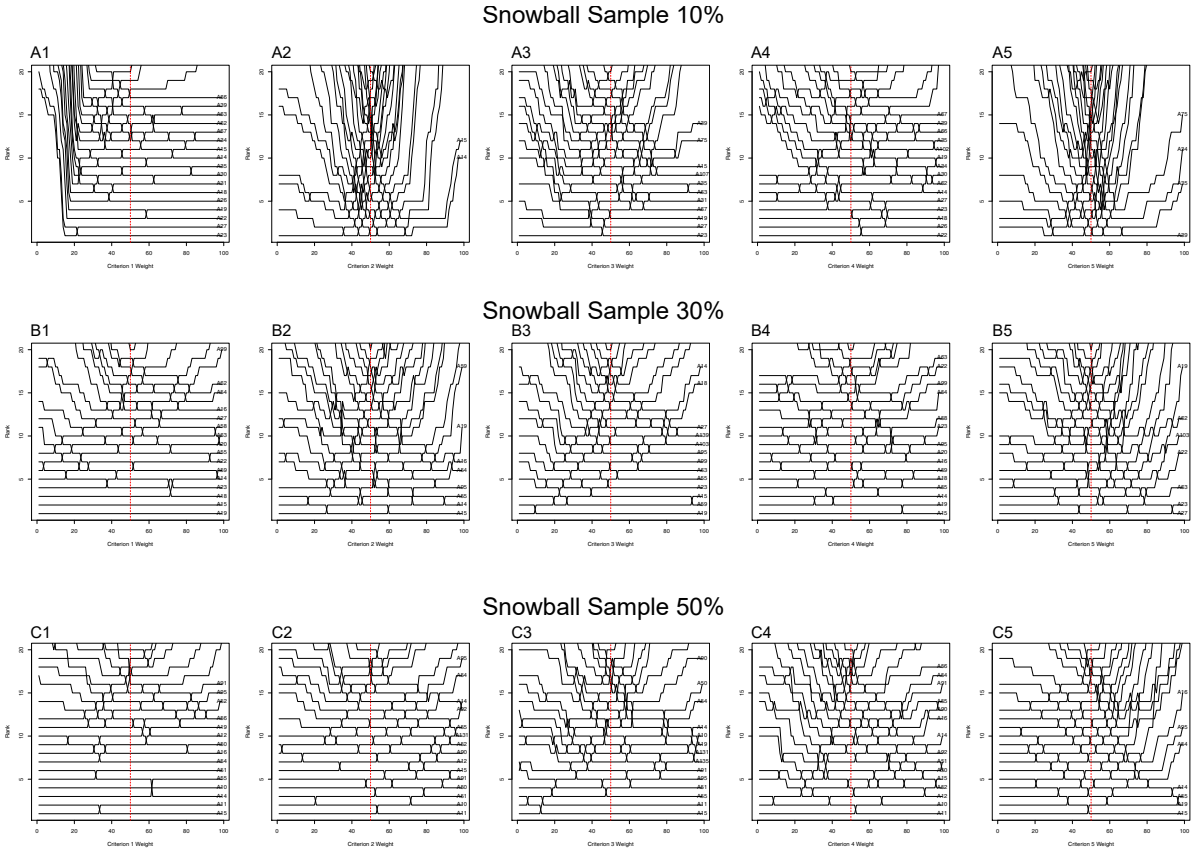


Fig. 7. Ranking sensitivity analysis for the top 20 alternatives from the TOPSIS evaluation of the real network [49] samples. A1-A5 – 10% network, B1-B5 – 30% network, C1-C5 – 50% network.

- the strategies' evaluation accuracy was compared between a full-size real network and a set of reduced-size synthetic and sample networks derived from the original network.

In practical terms, the empirical study has shown that while the synthetic networks, which were selected based on their Kullback-Leibler divergence, provided very similar results to the real network even when as little as 10% of nodes were used, in case of the sampled networks obtained with the snowball sampling approach provided satisfactory results only when the number of nodes was still relatively high. Also, while the rankings obtained from synthetic networks were stable, there was little stability of the rankings from the snowball sample networks.

All in all, the research has identified possible areas of improvement and future works. First of all, a more numerous set of sizes of sample network could be studied to verify how the network size affects its rankings' correlation to the real network's rankings. Secondly, only snowball sampling approach was used in the research. It would be beneficial to explore networks obtained with other sampling approaches. Last, but not least, the list of criteria could be expanded to allow more precise adjustment of the selected strategy to the marketer's needs.

VI. Acknowledgments

This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562.

References

[1] W. Chmielarz and O. Szumski, "Digital distribution of video games-an empirical study of game distribution platforms from the perspective of polish students (future managers)," in *Information Technology for Management: Emerging Research and Applications*. Springer, 2018, pp. 136–154.

[2] D. J. Watts, J. Peretti, and M. Frumin, *Viral marketing for the real world*. Harvard Business School Pub., 2007.

[3] E. Ziemba, *Towards a sustainable information society: People, business and public administration perspectives*. Cambridge Scholars Publishing, 2016.

[4] K. Szopik-Depczynska, A. Kedzierska-Szczepaniak, K. Szczepaniak, K. Cheba, W. Gajda, and G. Ioppolo, "Innovation in sustainable development: an investigation of the EU context using 2030 agenda indicators," *Land Use Policy*, vol. 79, pp. 251–262, Dec. 2018. doi: 10.1016/j.landusepol.2018.08.004. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0264837718306203

[5] W. Chmielarz and O. Szumski, "Analysis of users of computer games," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2016, pp. 1139–1146.

[6] O. Hinz, B. Skiera, C. Barrot, and J. U. Becker, "Seeding strategies for viral marketing: An empirical comparison," *Journal of Marketing*, vol. 75, no. 6, pp. 55–71, 2011.

[7] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia, "Analysing information flows and key mediators through temporal centrality metrics," in *Proceedings of the 3rd Workshop on Social Network Systems*. ACM, 2010, p. 3.

[8] M. Salehi, R. Sharma, M. Marzolla, M. Magnani, P. Siyari, and D. Montesi, "Spreading processes in multilayer networks," *IEEE Transactions on Network Science and Engineering*, vol. 2, no. 2, pp. 65–83, 2015.

[9] K. Kandhway and J. Kuri, "How to run a campaign: Optimal control of sis and sir information epidemics," *Applied Mathematics and Computation*, vol. 231, pp. 79–92, 2014.

[10] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[11] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[12] A. Karczmarczyk, J. Jankowski, and J. Wątróbski, "Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks," *PLOS ONE*, vol. 13, no. 12, p. e0209372, Dec. 2018. doi: 10.1371/journal.pone.0209372. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209372

[13] E. Ziemba, "The contribution of ict adoption to the sustainable information society," *Journal of Computer Information Systems*, vol. 59, no. 2, pp. 116–126, 2019.

[14] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[15] J. Wątróbski, E. Ziemba, A. Karczmarczyk, and J. Jankowski, "An index to measure the sustainable information society: the polish households case," *Sustainability*, vol. 10, no. 9, p. 3223, 2018.

[16] J. Jankowski, J. Hamari, and J. Wątróbski, "A gradual approach for maximising user conversion without compromising experience with high visual intensity website elements," *Internet Research*, vol. 29, no. 1, pp. 194–217, 2019.

[17] R. Pfitzner, A. Garas, and F. Schweitzer, "Emotional divergence influences information spreading in twitter." *ICWSM*, vol. 12, pp. 2–5, 2012.

[18] R. Michalski, T. Kajdanowicz, P. Bródka, and P. Kazienko, "Seed selection for spread of influence in social networks: Temporal vs. static approach," *New Generation Computing*, vol. 32, no. 3-4, pp. 213–235, 2014.

[19] C. Kiss and M. Bichler, "Identification of influencers: measuring influence in customer networks," *Decision Support Systems*, vol. 46, no. 1, pp. 233–253, 2008.

[20] Y. Liu-Thompkins, "Seeding viral content: The role of message and network factors," *Journal of Advertising Research*, vol. 52, no. 4, pp. 465–478, 2012.

[21] L. Seeman and Y. Singer, "Adaptive seeding in social networks," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 459–468.

[22] J. Jankowski, P. Bródka, P. Kazienko, B. K. Szymanski, R. Michalski, and T. Kajdanowicz, "Balancing speed and coverage by sequential seeding in complex networks," *Scientific reports*, vol. 7, no. 1, p. 891, 2017.

[23] J. Jankowski, "Dynamic rankings for seed selection in complex networks: Balancing costs and coverage," *Entropy*, vol. 19, no. 4, p. 170, 2017.

[24] J.-L. He, Y. Fu, and D.-B. Chen, "A novel top-k strategy for influence maximization in complex networks with community structure," *PloS one*, vol. 10, no. 12, p. e0145283, 2015.

[25] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Scientific reports*, vol. 6, p. 27823, 2016.

[26] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, p. 888, 2010.

[27] C. Granell, S. Gómez, and A. Arenas, "Competing spreading processes on multiplex networks: awareness and epidemics," *Physical review E*, vol. 90, no. 1, p. 012808, 2014.

[28] C. Granell, S. Gómez, and A. Arenas, "Dynamical interplay between awareness and epidemic spreading in multiplex networks," *Physical review letters*, vol. 111, no. 12, p. 128701, 2013.

[29] X. Wei, N. C. Valler, B. A. Prakash, I. Neamtiu, M. Faloutsos, and C. Faloutsos, "Competing memes propagation on networks: A network science perspective," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1049–1060, 2013.

[30] M. Bampo, M. T. Ewing, D. R. Mather, D. Stewart, and M. Wallace, "The effects of the social structure of digital networks on viral marketing performance," *Information systems research*, vol. 19, no. 3, pp. 273–290, 2008.

[31] J. Y. Ho and M. Dempsey, "Viral marketing: Motivations to forward online content," *Journal of Business research*, vol. 63, no. 9-10, pp. 1000–1006, 2010.

[32] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media: sentiment of microblogs and sharing behavior," *Journal of management information systems*, vol. 29, no. 4, pp. 217–248, 2013.

[33] A. Dobele, A. Lindgreen, M. Beverland, J. Vanhamme, and R. Van Wijk, "Why pass on viral messages? because they connect emotionally," *Business Horizons*, vol. 50, no. 4, pp. 291–304, 2007.

[34] C. Camarero and R. San José, "Social and attitudinal determinants of viral marketing dynamics," *Computers in Human Behavior*, vol. 27, no. 6, pp. 2292–2300, 2011.

[35] J. Berger and K. L. Milkman, "What makes online content viral?" *Journal of marketing research*, vol. 49, no. 2, pp. 192–205, 2012.

[36] A. Rezvanian, B. Moradabadi, M. Ghavipour, M. M. D. Khomami, and M. R. Meybodi, "Social network sampling," in *Learning Automata Approach for Social Networks*. Springer, 2019, pp. 91–149.

[37] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[38] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[39] P. Erdös and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[40] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[41] J. Wątróbski, K. Małecki, K. Kijewska, S. Iwan, A. Karczmarczyk, and R. Thompson, "Multi-criteria analysis of electric vans for city logistics," *Sustainability*, vol. 9, no. 8, p. 1453, 2017.

[42] P. Ziemba, "Neat f-promethee–a new fuzzy multiple criteria decision making method based on the adjustment of mapping trapezoidal fuzzy numbers," *Expert Systems with Applications*, vol. 110, pp. 363–380, 2018.

[43] J. Wątróbski, J. Jankowski, P. Ziemba, A. Karczmarczyk, and M. Zioło, "Generalised framework for multi-criteria method selection," *Omega*, vol. 86, pp. 107–124, 2019.

[44] J. Wątróbski, J. Jankowski, P. Ziemba, A. Karczmarczyk, and M. Zioło, "Generalised framework for multi-criteria method selection: Rule set database and exemplary decision support system implementation blueprints," *Data in brief*, vol. 22, p. 639, 2019.

[45] J. Wątróbski, J. Jankowski, and Z. Piotrowski, "The selection of multicriteria method based on unstructured decision problem description," in *International Conference on Computational Collective Intelligence*. Springer, 2014, pp. 454–465.

[46] J. Wątróbski and J. Jankowski, "Guideline for mcda method selection in production management area," in *New frontiers in information and production systems modelling and analysis*. Springer, 2016, pp. 119–138.

[47] W. Chmielarz and M. Zborowski, "Analysis of e-banking websites' quality with the application of the topsis method–a practical study," *Procedia computer science*, vol. 126, pp. 1964–1976, 2018.

[48] M. Jankowski, A. Borsukiewicz, K. Szopik-Depczynska, and G. Ioppolo, "Determination of an optimal pinch point temperature difference interval in ORC power plant using multi-objective approach," *Journal of Cleaner Production*, vol. 217, pp. 798–807, Apr. 2019. doi: 10.1016/j.jclepro.2019.01.250. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0959652619302756

[49] M. Ripeanu, I. Foster, and A. Iamnitchi, "Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design," *arXiv:cs/0209028*, Sep. 2002, arXiv: cs/0209028. [Online]. Available: http://arxiv.org/abs/cs/0209028

[50] "Snowball Sampling Function - R Documentation." [Online]. Available: https://www.rdocumentation.org/packages/netdep/versions/0.1.0/topics/snowball.sampling

[51] S. Kemp, "Digital 2019: Global Internet Use Accelerates," Jan. 2019. [Online]. Available: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates

A4.

Karczmarczyk, A., Jankowski, J., Wątróbski, J. (2019). Parametrization of spreading processes within complex networks with the use of knowledge acquired from network samples. Procedia Computer Science, 159, 2279-2293.

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 159 (2019) 2279–2293

Procedia Computer Science

www.elsevier.com/locate/procedia

23rd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Parametrization of Spreading Processes Within Complex Networks with the Use of Knowledge Acquired from Network Samples

Artur Karczmarczyk[a], Jaroslaw Jankowski[a], Jaroslaw Watrobski[b,*]

[a] *Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Zolnierska 49, Szczecin 71-210, Poland*
[b] *Faculty of Economics and Management, University of Szczecin. Mickewicza 64, Szczecin 71-101, Poland*

Abstract

Information spreading processes are main drivers of viral campaigns. They are usually conducted within large scale social networks. Parametrisation of online campaigns is usually related to allocation of budgets, number of seeds and strategies of their selection. It is hard to predict campaign effects. Proposed in this paper approach uses network samples to select appropriate strategy for use within complete network. Results showed how scaling of samples is affecting approximation quality. Multi-criteria analysis of performance allows for strategy selection according to preferences and goals.

Keywords: Information spreading; viral marketing; network sampling; multi-criteria analysis

1. Introduction

The rapid and substantial increase in popularity of social networking platforms [21, 17] has led to a crucial need to understand how millions of online users behave, including their patterns and predispositions [10]. In a number of cases, it can be observed that as a result of information spreading between social media users viral marketing seems to produce better results than traditional advertising campaigns based on ads from commercial sources [45]. Therefore, an increased number of online marketers place efforts in the engagement of potential consumers to benefit from their products and services by propagating information. Recommendations that are socially oriented have a greater impact on targeted consumers [14]. The higher faith in communications within a social network that has ties stronger than for traditional commercial messages is observed. A prior research in this area implemented macroscopic approaches to analyze the quantity of customers acquired using a diffusion of innovations mechanics [29]. More detailed approaches

* Corresponding author.
E-mail address: jwatrobski@usz.edu.pl

10.1016/j.procs.2019.09.403

performed the identification and assessment of those who send and receive messages through a monitoring of the processes involved in the distribution of information [6].

Research that is related to the viral marketing that occurs in complex networks takes into consideration the aspects that lead to campaigns that are successful [3, 12], the selection of initial seeds for the initialization of the campaign [11] as well as epidemic extensions and models usage to model diffusion processes [18]. Multilayer structures [35, 16] and the spread of information in temporal networks have been studied in more recent research [38, 26].

The complexity of mechanics behind viral marketing and information spreading processes has been analyzed from various perspectives. Many prior studies were oriented on theoretical and empirical approaches in order to increase the number of customers reached within the network, i.e. to increase the network coverage. While it is important metrics of campaign success, also other factors can be taken into consideration. Apart from coverage they include campaign costs, duration, seeding intensity and strategies of initial nodes selection. In the authors' previous study [19], a framework for strategic planning of information spreading processes, which helps to select appropriate strategy for selection of initial nodes within the network and adjusting the number of activated nodes in seeding process, was proposed. However, strategic planning on the real network model is time-consuming and requires considerable processing power. Therefore, in this paper the authors' propose an approach in which samples of reduced size yet acceptable accuracy are selected for performing the viral marketing campaign strategy planning.

The main contribution of the presented study is to provide a two-track framework for selection of network sample for viral marketing campaign planning which would reduce the computational requirements for the planning process yet would remain satisfactory in terms of accuracy, and take into consideration the marketer's preferences. In practical terms, a set of evaluation criteria for network sample for viral marketing campaign planning selection is proposed. Moreover, a detailed analysis of relations between network samples of various size and a real network is provided.

The paper comprises of 5 sections. After this introduction, in Section 2 a literature review of the state of the art is presented. Subsequently, in Section 3 the methodological framework of the proposed approach is presented, followed by an empirical study in Section 4. Conclusions and possible future works are presented in Section 5.

2. Literature Review

Social network analysis was initially based on real connections among people with very limited analytical abilities and applications. Together with development of electronic systems social relations become better trackable with possibility to analyses associated with them phenomena. Currently electronic systems cover many aspects of social life and real life behaviors have their equivalents within online social networking platforms [2]. One of key phenomena observed within social networks is information spreading with the use of social influence mechanisms. Apart from social context it is used for commercial messages dissemination and the performance of marketing messages is increased with personal recommendations and trust. Successful viral marketing campaigns using this mechanics can reach extensive audiences with relatively low budges and that why become one of key elements of marketing strategies. Complexity of spreading processes attracted attention from practitioners and researchers from various disciplines like marketing [11], physics [14] or mathematics [18]. From the marketing perspective main goals are related to campaign performance, theoretical studies are focused on computational complexity and generalized models.

In general, large fraction of studies are focused on models used for simulations or formalization of information spreading processes for their better understanding and prediction. Earlier models from the area of epidemic research are adopted to other applications and extension of SIS and SIR models [18]. Newly created models take into account specifics of spreading processes withing well defined network structure and are based on threshold [29] and cascading approaches [22]. Threshold models use social influence mechanisms and increased ability to adopt to new product or ideas together with growing number of activated neighbours. Activation takes place when proportion of activated friends is higher than assumed threshold. Different spreading model is used for independent cascades. Spreading behavior is resulting information cascades observed when information is transmitted to network neighbours, and then their transmit it to fiends and so on.

Studies in this field are conducted not only with the use of different spreading models and approaches but can be performed within different network environments. Simplest approaches are based on static networks representing snapshot of real network with assumed constant number of network nodes and edges. Even though such type of network are rarely observed in real systems they are used for analysis with acceptable computational complexity.

While static networks are commonly used for modeling spreading processes in real systems temporal more typical are networks with changing structures [13]. They better describe real situations with changing number of social contacts over the time. Their usage adds another dimensions to performed analysis which are related to probabilities of nodes or adding and removal as well as scale of changes in analyzed time intervals. Another simplification usually taken is the use of single layer networks. For real social systems behavior is observed usually in several layers, for example electronic and direct communication [35].

One of key identified problem is influence maximization based on such selection of nodes with the network to initiate spreading processes with the highest possible network coverage [11]. Earlier studies showed the ability of effective selection with the use of greedy solutions with high computational complexity [22]. They are difficult to implement within larger networks. More common for usage are heuristics based on high degree or other centrality measures [23] [25].

Seeding process can be performed not only at the beginning to initiate spreading processes but can take an adaptive form with the use of knowledge on ongoing processes [36]. Influence on spreading processes can be performed with additional activation of seeds within network segments more difficult to reach [15]. Additional knowledge about existing communities can be used to spread seeds more effectively citehe2015novel as well as k-shell based identification of nodes with spreading potential [24]. Apart from mentioned areas other studies focused on role of network topology on spreading processes [1], techniques used for increasing motivation of uses to spread the content [12], role of emotions [37] [7] and other factors [5] [3].

From the perspective of effectiveness of used seed selection strategies and other methods usually network coverage is used represented by a fraction of activated nodes. It is important factor for marketing campaigns but other factors like campaign budgets or characteristics of target audiences should be taken into account like it was showed in earlier research [20]. Current study extends earlier approach by using network samples with the use of snowball sampling [27] and different sample sizes. It allows to obtain effectively simulation results without the need of target network analysis. Acquired parameters can be used for campaign parametrization within real environment.

3. Methodological Framework

As it was shown in the literature review in section 2, marketers nowadays place more and more effort on the viral marketing campaigns in the complex networks of social media. Prior to physically executing a real campaign, the information propagation process can be studied with the use of agent-based simulations. During the simulations, the effect of manipulations with the input parameters, such as number of nodes initially infected with the propagated information (seeding fraction, SF) and the information propagation probability within the network can be studied on the potential campaign results (such as obtained information propagation coverage or duration of the process). However, the exact mapping of the complete network is often not available to run simulations on. Moreover, running simulations on a complete real-size network is time-consuming and requires nontrivial computational power.

Therefore, in the first phase of the authors' proposed approach (see Fig. 1), a set of smaller samples of the real network is generated. When the network samples are created, simulation parameters are fed to the simulation engine along with the network samples' structures, and simulations are executed. For this purpose, the independent cascades model (IC, [22]) can be used.

After the initial simulations of all potential strategies on all network samples are complete, the marketer is provided with multiple conflicting results. Therefore, the results obtained from the network samples need to be compared with the results from the real network and their accuracy needs to be verified.

In the authors' approach, the aforementioned accuracy verification is divided into two steps. In the first step, the results of each strategy for a given network sample are juxtaposed with the related results from the real network. The deltas and ratios obtained from such juxtaposition are then plotted and mathematical and geometrical analysis is used to asses the similarity of the results obtained from the reduced-size samples compared to the real network. This step provides broad knowledge about the generated samples and their relations to the real network, based on which, the network sample for further simulations and viral marketing planning should be chosen.

However, in real-life use-cases the network sample that provides the highest accuracy of results compared to the real network might not sufficiently reduce the computational requirements of the planning process. The marketer might decide that lower accuracy would be acceptable if that resulted in different benefits of the eventually selected

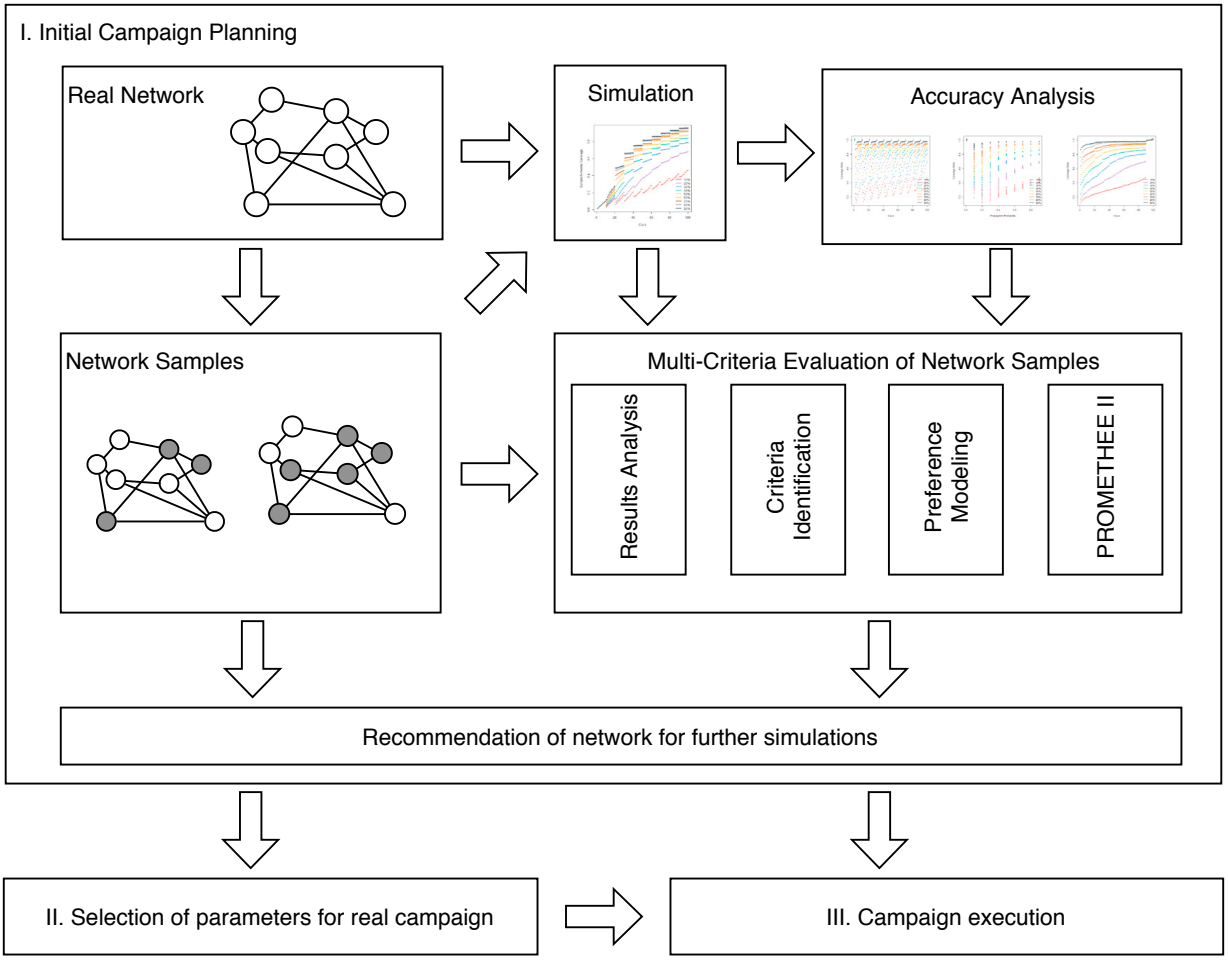


Fig. 1. Methodological framework of the proposed approach.

solution (so-called criteria compensation effect). Therefore, the authors introduced the second step of the approach, in which multi-criteria decision analysis tools are used to identify the criteria for evaluation of the network samples for viral marketing campaign planning, model the marketer's preferences and present recommendation for selection of the network for campaign planning.

There are multiple MCDA methods available, which can generally be divided into two groups, so-called American and European MCDA schools [30]. The methods from the former group focus on aggregating multiple real criteria into a single utility function value pseudo-criterion [9]. They include inter alia AHP, ANP [31], TOPSIS [34], COMET [33, 32]. On the other hand, the methods from the latter group are oriented on the outranking relation between each alternative. They include inter alia ELECTRE [8], PROMETHEE [4] or NAIADE [40] methods.

The selection of the most appropriate MCDA method for a given decicion-making problem is difficult, however research and tools facilitating these choices already exist [39, 41, 42, 43, 44]. In case of the authors' proposed framework, the criteria in the decision-making problem are weighted. Also, the marketer's preferences uncertainty exists, related both to indifference and preference of alternatives. Finally, since a single network sample should be selected as the outcome of the proposed framework, the method should produce a complete ranking of alternatives. Out of 56 considered MCDA methods [44], only PROMETHEE II method accomplishes all the aforementioned requirements.

In PROMETHEE II [4], a complete ranking of alternatives is obtained based on the values of net outranking flows. These flows are based on pairwise comparisons of all alternatives under each criterion, which result in input and output preference flows. During the comparisons, the preference can be expressed as a Boolean value or a more precise intermediate value based one of six preference functions: usual, U-shape, V-shape, level, linear or Gaussian [19]. Moreover, PROMETHEE methods provide a GAIA (Geometrical Analysis for Interactive Aid) tool for visual

Table 1. Parameters of the real network [28].

D	C	PR	EV	CC	B
7.1985	1.59E-07	0.000113045	0.01602488	0.000113084	19104.87

D - degree, C - closeness, PR - page rank, EV - eigenvector, CC - clustering coefficient, B - betweenness

analysis of the relations between criteria and alternatives, which in the authors' proposed framework can be used for understanding which criteria support particular samples of the real network.

In the authors' proposed framework, the PRMETHEE II method is used to build the preference model of the marketer regarding the selection of the real network sample for future viral marketing campaign planning. The authors propose the following set of criteria, divided into two groups: costs and accuracy:

- C1 – costs group – size ratio of the network sample equal to the fraction of real network nodes selected to the network sample;
- C2 – costs group – time required to generate the network sample of a specified size;
- C3 – accuracy group – distance of the coverage ratio computed as the ratio between the average coverage obtained by the network sample of a specified size and the average coverage in the real network from the ideal 1/1 ratio;
- C4 – accuracy group – distance of the duration ratio computed as the ratio between the average number of infection iterations in the network sample and the average number of infection iterations in the real network from the ideal 1/1 ratio.

During the process of PROMETHEE II analysis, rankings of network samples are obtained. Stability intervals of the obtained solutions are verified and preference functions are adjusted in order to provide solution most satisfactory for the marketer. Eventually, a particular network sample is recommended by the framework for future viral marketing campaign planning.

All in all, the authors' methodological contribution in this paper is to provide a two-track framework for selection of network sample for viral marketing campaign planning which would reduce the computational requirements for the planning process yet remain satisfactory in terms of accuracy, and take into consideration the marketer's preferences. In practical terms, a set of four criteria for selecting network sample for viral marketing campaign planning was proposed. Moreover, a detailed analysis of relations between network samples of various size and the actual real network were studied.

4. Empirical Study

The empirical verification of the proposed approach was based on a real network [28]. The network is a part of the topology of the Gnutella network as mapped in 2002. The network is built of 8846 nodes (Gnutella network hosts) and 31839 edges (connections between Gnutaella hosts). The network is a single snapshot collected in August 2002. The average values of the main network's metrics are presented in Table 1.

4.1. Independent Cascade Simulations and Sampling

During the empirical study, the most fundamental scenario of social media viral marketing campaign planning was used, i.e. it was verified how changes in the values of seeding fraction (SF) and propagation probability (PP) affect the information spreading – what coverage is obtained and in how many iterations the information propagation process ends. The independent cascades model was used to perform simulations on the network. A total of 1000 simulations was performed, which number results from the Cartesian product of:

- 10 SF values: 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.10;

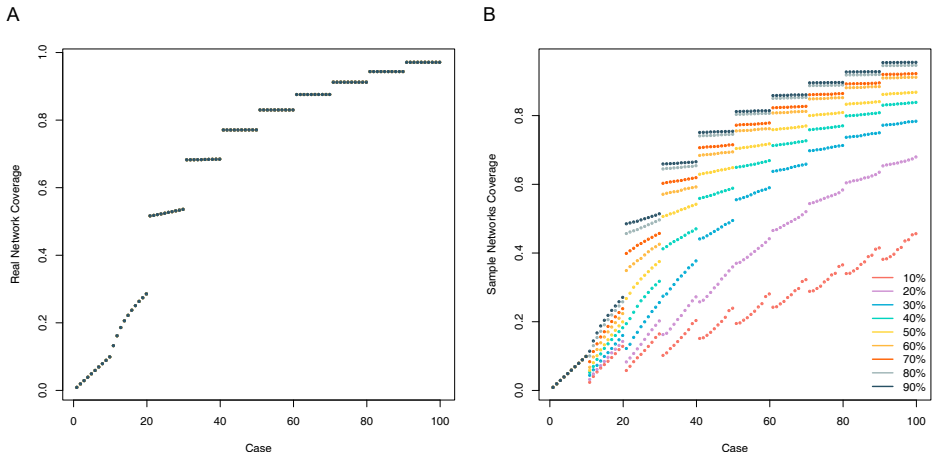


Fig. 2. Coverage obtained by the real network [28] (A) and the sample networks (B). The cases on the charts are ordered ascending by the real network's coverage value obtained.

- 10 PP values: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9;
- 10 weight scenarios in which for each node a random value was drawn. If the value was smaller or equal to the PP value during simulation, an infection. If the value was above the PP value, information was not propagated by the particular node.

Before each simulation, the actual nodes to be seeded with information were selected based on their degree – nodes connected directly to the highest number of other nodes were selected.

The simulation results are presented in Table 2. For each set of SF and PP parameters, the resulting coverage and iterations count (duration of the process) were averaged from the 10 underlying weight scenarios. The coverage obtained in the simulations is visually presented on Fig. 2A. It can be observed, that the coverage value ranges from 0.0099 for the case with lowest SF and PP values to 0.9721 for the case with highest SF and PP values. The information propagation process duration oscillates from 1 to 16.8 iterations.

Based on the results in Table 2 it can be confirmed that the obtained coverage raises along with the increase of the SF and PP values. For example, in a hypothetical scenario where the campaign ordering party wanted to achieve a coverage of at least 25% of the network and it was known that the propagation probability was equal to 0.1, only four campaign strategies would bring satisfactory results - the ones with $PP = 0.1$ and $SF >= 0.07$.

Agent simulations allow to plan viral marketing campaigns on complex networks. However, it is a time- and resource-consuming process. For example, to obtain the results in Table 2, 1000 simulations needed to be performed, which resulted in 6835 iterations, during which a total of 6,013,924 needed to be registered. Also, often the precise mapping of the complete real network is not available to the campaign ordering parties.

Consequently, it would be beneficial to perform the simulations and plan the real network viral marketing campaign with the use of smaller networks. In this paper, snowball network samples of sizes of 10%, 20%, ..., 90% of the real network were used. The average times of generation of each of the network samples for the real network are presented in Table 3.

After the network samples were obtained, for each network 1000 simulations were performed, as for the real network, which resulted in a total of 9000 simulations. The obtained coverage values for each network and each SF and PP parameters case is presented on Fig. 2B. The analysis of Fig. 2 allows to observe that the coverage obtained for the 90% sample is very similar to the real network. The simulation process for this network sample took 6877 iterations, however it required performing only 5,280,770 infections (87% of the infections count for the real network). The results for the 70% network are also visually very similar, yet it required only 3,879,399 infections. Along with the decrease of the size of the sample, the number of infections that need to be tracked drops down to 195,852 for the 10% sample, which significantly reduces the need for time and processing power to perform the simulations. Therefore, it is beneficial for the marketer to perform the planning on a network which provides accurate enough results, yet

Table 2. Simulation results for the real network [28].

SF	PP	Avg. Coverage	Avg. Duration	Avg. Infected	SF	PP	Avg. Coverage	Avg. Duration	Avg. Infected
0.01	0.01	0.009947999	1	88	0.06	0.01	0.060027131	1	531
0.01	0.1	0.133404929	16.8	1180.1	0.06	0.1	0.238390233	11.2	2108.8
0.01	0.2	0.517386389	14.4	4576.8	0.06	0.2	0.527933529	11	4670.1
0.01	0.3	0.683404929	9.5	6045.4	0.06	0.3	0.684467556	7.9	6054.8
0.01	0.4	0.771523853	8.1	6824.9	0.06	0.4	0.77171603	7	6826.6
0.01	0.5	0.830793579	7.4	7349.2	0.06	0.5	0.830850102	6.7	7349.7
0.01	0.6	0.876735247	6.7	7755.6	0.06	0.6	0.876814379	5.9	7756.3
0.01	0.7	0.913395885	6.2	8079.9	0.06	0.7	0.913395885	5.3	8079.9
0.01	0.8	0.944630341	6.2	8356.2	0.06	0.8	0.944630341	5.2	8356.2
0.01	0.9	0.972145603	6.1	8599.6	0.06	0.9	0.972145603	5.1	8599.6
0.02	0.01	0.020009044	1	177	0.07	0.01	0.06997513	1	619
0.02	0.1	0.16253674	14.5	1437.8	0.07	0.1	0.25186525	10.1	2228
0.02	0.2	0.518923807	13.2	4590.4	0.07	0.2	0.529934433	10.7	4687.8
0.02	0.3	0.683563192	8.6	6046.8	0.07	0.3	0.684772779	7.6	6057.5
0.02	0.4	0.771523853	7.6	6824.9	0.07	0.4	0.771749943	6.9	6826.9
0.02	0.5	0.830793579	7.3	7349.2	0.07	0.5	0.830850102	6.6	7349.7
0.02	0.6	0.876735247	6.3	7755.6	0.07	0.6	0.876814379	5.9	7756.3
0.02	0.7	0.913395885	6.1	8079.9	0.07	0.7	0.913395885	5.3	8079.9
0.02	0.8	0.944630341	5.9	8356.2	0.07	0.8	0.944630341	5.2	8356.2
0.02	0.9	0.972145603	5.6	8599.6	0.07	0.9	0.972145603	5.1	8599.6
0.03	0.01	0.029957043	1	265	0.08	0.01	0.080036175	1	708
0.03	0.1	0.187044992	13	1654.6	0.08	0.1	0.264481121	9.6	2339.6
0.03	0.2	0.521309066	12.2	4611.5	0.08	0.2	0.532285779	10.6	4708.6
0.03	0.3	0.683687542	8.1	6047.9	0.08	0.3	0.684987565	7.5	6059.4
0.03	0.4	0.771523853	7.1	6824.9	0.08	0.4	0.771795161	6.6	6827.3
0.03	0.5	0.830793579	7.2	7349.2	0.08	0.5	0.830850102	6.4	7349.7
0.03	0.6	0.876735247	6.1	7755.6	0.08	0.6	0.876814379	5.8	7756.3
0.03	0.7	0.913395885	5.7	8079.9	0.08	0.7	0.913395885	5.2	8079.9
0.03	0.8	0.944630341	5.5	8356.2	0.08	0.8	0.944630341	5.1	8356.2
0.03	0.9	0.972145603	5.1	8599.6	0.08	0.9	0.972145603	5.1	8599.6
0.04	0.01	0.040018087	1	354	0.09	0.01	0.089984174	1	796
0.04	0.1	0.206952295	12.1	1830.7	0.09	0.1	0.275254352	9.5	2434.9
0.04	0.2	0.523016052	11.8	4626.6	0.09	0.2	0.534580601	10.3	4728.9
0.04	0.3	0.683845806	8	6049.3	0.09	0.3	0.685315397	7.3	6062.3
0.04	0.4	0.771523853	7.1	6824.9	0.09	0.4	0.77184038	6.6	6827.7
0.04	0.5	0.830793579	6.9	7349.2	0.09	0.5	0.830850102	6.4	7349.7
0.04	0.6	0.876735247	6.1	7755.6	0.09	0.6	0.876814379	5.7	7756.3
0.04	0.7	0.913395885	5.6	8079.9	0.09	0.7	0.913395885	5.2	8079.9
0.04	0.8	0.944630341	5.5	8356.2	0.09	0.8	0.944630341	5.1	8356.2
0.04	0.9	0.972145603	5.1	8599.6	0.09	0.9	0.972145603	5.1	8599.6
0.05	0.01	0.049966086	1	442	0.1	0.01	0.100045218	1	885
0.05	0.1	0.223129098	11.8	1973.8	0.1	0.1	0.286321501	9.5	2532.8
0.05	0.2	0.52532218	11.1	4647	0.1	0.2	0.53684151	9.8	4748.9
0.05	0.3	0.683992765	7.9	6050.6	0.1	0.3	0.685507574	7.3	6064
0.05	0.4	0.771523853	7	6824.9	0.1	0.4	0.77184038	6.6	6827.7
0.05	0.5	0.830793579	6.7	7349.2	0.1	0.5	0.830850102	6.2	7349.7
0.05	0.6	0.876735247	5.9	7755.6	0.1	0.6	0.876814379	5.7	7756.3
0.05	0.7	0.913395885	5.5	8079.9	0.1	0.7	0.913395885	5.2	8079.9
0.05	0.8	0.944630341	5.5	8356.2	0.1	0.8	0.944630341	5.1	8356.2
0.05	0.9	0.972145603	5.1	8599.6	0.1	0.9	0.972145603	5.1	8599.6

requires little computational power. A detailed comparison of the accuracy of the viral marketing campaign plans for each network sample size are studied in the following sections.

Table 3. Generation times of the real network [28] samples.

% of [28] nodes	10	20	30	40	50	60	70	80	90
Num. of nodes	885	1770	2654	3539	4423	5308	6193	7077	7962
Generation time [s]	6.01	12.10	8.23	19.56	24.47	34.25	37.54	47.28	60.45

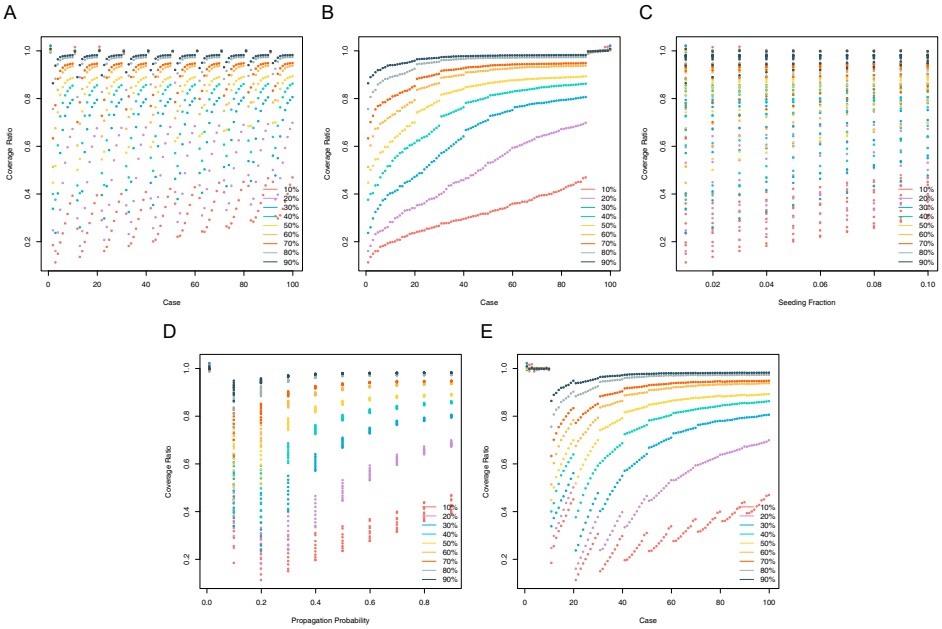


Fig. 3. Coverage ratio of the sample networks to the real network [28]. A: ordered by simulation case, B: ordered by coverage ratio value ascending, C: grouped and ordered by SF, D: grouped and ordered by PP, E: ordered by the real network coverage value ascending.

4.2. Basic comparison of real network and network samples coverage similarity

After the simulations on the sampled networks were complete, the authors were able to perform a basic comparison of similarity between the coverage obtained in the real network and the network samples for particular campaign strategies. In order to perform the analysis, the results were aggregated. First of all, average coverage and average iteration of the last infection (average duration of the propagation process) were averaged from the 10 runs for each simulation settings. This reduced the number of results to 1000, 100 for each network. In the next step, for each network sample, for each strategy (set of SF and PP parameters), the results from the simulation of the network sample and real network were juxtaposed, and the coverage ratio and duration ratio were thus obtained. The former is presented on Fig. 3 and the latter on Fig 4.

The analysis of Fig. 3A shows that in majority of the cases the coverage ratio is lower than 1, i.e. the coverage obtained on sampled network was smaller than on the real network. Fig. 3B confirms the intuitive assumption that the closer the size of the sampled network to the full real network, the closer the coverage ratio to 1. Fig. 3 allows to observe that for SF values from 0.1 to 0.3 the obtained coverage values were the least similar to the real network ones. Additionally, for little values of SF in some simulations the coverage obtained on sampled network was bigger than on the real network. While the values for each network on Fig. 3C were mixed for each SF value, in case of Fig. 3D, the values for each network display in groups, especially for PP values greater than 0.3. These groups are most condensed for the samples with the highest numbers of nodes, i.e. the 60% – 90% samples, which means that the results for these networks were more stable regardless of the weight scenario. On Fig. 3E the cases are ordered based on the ascending value of the coverage in the real network for a given case. This chart again allows observe that the bigger the sample, the closer the coverage ratio to 1 and also the smaller oscillations of the coverage ratio values.

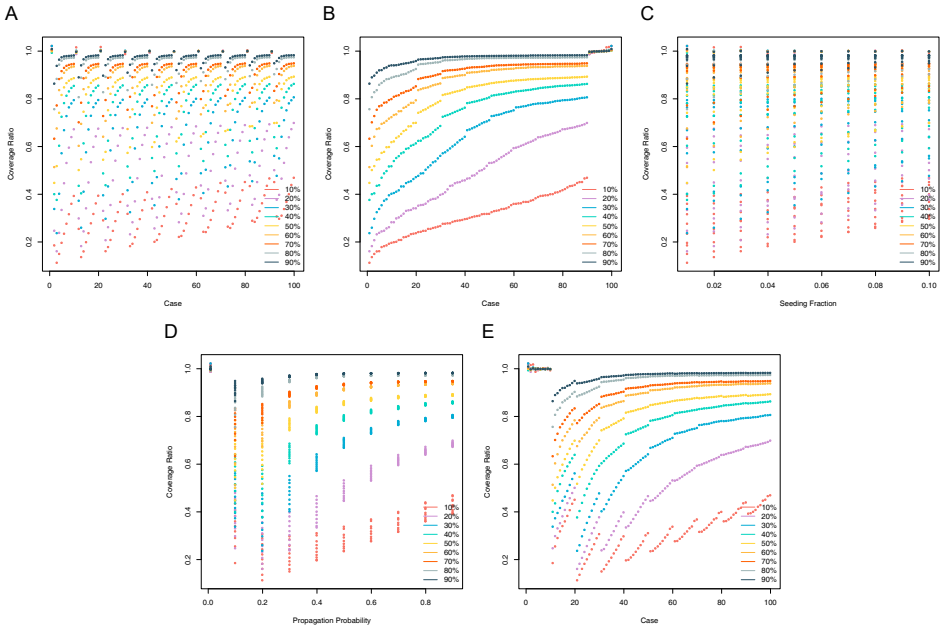


Fig. 4. Duration ratio of the sample networks to the real network [28]. A: ordered by simulation case, B: ordered by duration ratio value ascending, C: grouped and ordered by SF, D: grouped and ordered by PP, E: ordered by the real network coverage value ascending.

Similar analysis can be performed for the duration ratio on Fig. 4. Fig. 4A shows that the duration ratio ranged from 0.2381 for the 10% network, $SF = 0.01, PP = 0.1$; to 2 for the 20% network, $SF \in \{0.03, 0.04\}, PP = 0.6$. When the cases are ordered by the duration ratio value ascending (see Fig. 4B), it can be observed that again the 90% network provides the greatest and the 10% network the lowest accuracy. Figures 4C and 4D show that for individual network sample, while for some PP values the information propagation process tended to last longer than in the real network, for other PP values it was shorter, which fact is not observed for the varying values of SF.

Out of the 900 juxtaposed strategies in network samples of various sizes, only 7 cases had the coverage value equal to the real network – five cases for the 50% network and two for the 80% network. All seven cases were characterized by the lowest possible value of PP, i.e. 0.01. For the remaining network sample sizes, the most accurate coverage ratio result was as follows – 10% : 1.0012, 20% : 0.9995, 30% : 1.0008, 40% : 0.9998, 60% : 0.9999, 70% : 0.9995, 90% : 1.0001.

4.3. MCDA evaluation of social media campaign planning strategies

The analysis presented in subsection 4.2 allows to understand the campaign strategy planning accuracy for network samples of all sizes and for all strategies. In case of the presented real network, selection of the 90% network sample for the viral marketing campaign strategy planning seems to be the most appropriate option. However, in real-life applications the marketer might decide to waive the ideal accuracy of the campaign planning if that would result in other benefits, which would compensate the accuracy loss (criteria compensation). For this purpose, the authors' proposed framework uses the MCDA component to facilitate the selection of the sample size under varying preferences of the marketer.

In the empirical research, all the results from the simulations of the network samples were aggregated into a total of nine results - one for each network. Each results' row was extended by the length of the sample generation (see Table 3). The coverage ratio and duration ratio columns were converted into error values from the 1/1 ratio. As a result, a criterial performance matrix was obtained for the PROMETHEE II analysis, as shown in Table 4.

Initially, a scenario in which the marketer prefers almost exclusively very good accuracy was considered. Therefore, a very high weight was assigned to the C3 criterion and very low weight for the other criteria. The preference direction of all criteria was set to favour minimal values. Usual preference function was used (see Table 5A). The ranking of the network samples with such marketer's preferences are presented in Table 6A. As it was intuitively assumed in section

Table 4. Criterial performance of the 9 sample networks for the PROMETHEE II multi-criteria analysis.

Network	C1	C2 [s]	C3	C4
10%	0.1	6.01	0.619806	0.037354
20%	0.2	12.1	0.462804	0.340597
30%	0.3	8.23	0.330667	0.321128
40%	0.4	19.56	0.240467	0.204471
50%	0.5	24.47	0.179723	0.139564
60%	0.6	34.25	0.118034	0.111196
70%	0.7	37.54	0.088363	0.065311
80%	0.8	47.28	0.046445	0.027444
90%	0.9	60.45	0.028015	0.007866

Table 5. PROMETHEE II method parameters used in the empirical study.

Criterion	**C1**	**C2**	**C3**	**C4**
Preference	min	min	min	min
		A		
Weight	1	1	97	1
Function	usual	usual	usual	usual
		B		
Weight	1	1	1	1
Function	usual	usual	usual	usual
		C		
Weight	1	1	1	1
Function	linear	linear	linear	linear
Q: indifference	0.26	17.53	0.190259	0.117508
P: preference	0.52	35.06	0.380518	0.235016

4.2, the 90% network sample was ranked first. The rank of the remaining alternatives is in inverse proportion to their size, i.e. the smallest 10% network was ranked 9th.

Subsequently, a scenario in which the marketer assigns equal importance to all criteria was studied. Therefore, each criterion C1-C4 was assigned an equal weight of 1 (see Table 5B). The obtained ranking is presented in Table 6B. It can be observed that after the marketer' preferences changed, the 10% network sample, previously ranked last, now is the leader. Selection of second-best network is difficult in this scenario, however, because three networks, i.e. 30%, 80% and 90% samples obtained the same score, and so did 40 – 70% samples. This fact can be visually observed on Fig. 5A. In turn, on Fig. 5A1 - Fig. 5A4 the distance of each network from each criterial axis is marked.

In order to avoid such draws, the most basic usual preference function can be replaced with a more complex one, which instead of a simple Boolean value, returns information how much one network is better from the other under the criterion in question. In the empirical study, the linear (v-shape) function with indifference and preference

Table 6. PROMETHEE II method results from the empirical study.

Network	10%	20%	30%	40%	50%	60%	70%	80%	90%
					A				
Rank	9	8	7	6	5	4	3	2	1
Phi	-0,9450	-0,7250	-0,4800	-0,2425	-0,0025	0,2375	0,4775	0,7200	0,9600
					B				
Rank	1	9	2	5	5	5	5	2	2
Phi	0,3750	-0,1250	0,0000	-0,0625	-0,0625	-0,0625	-0,0625	0,0000	0,0000
					C				
Rank	1	8	6	3	2	4	5	7	9
Phi	0,1246	-0,1227	-0,0161	0,0771	0,1210	0,0420	0,0053	-0,0774	-0,1536

thresholds was used. The indifference threshold was set to the value of standard deviation of the performances of the networks under each criterion, and the preference threshold to its twofold value (see Table 5C). The resulting ranking is presented in Table 6C.

It can be noted that the usage of a more complex preference function allowed to eliminate draws from the ranking. However, not only the ranks of the networks previously in draw changed. The 90% network sample, leading on the first ranking and ranked second on the second ranking, here is ranked 9th. Also, it can be noted that usage of the linear preference function allowed to obtain a more detailed score of each alternative (see Table 6C compared to Table 6B). The remaining differences in rankings based on the usual and linear functions were presented on Fig. 6, on which the closer to the diagonal line a network sample is marked, the smaller its change in the ranking was.

In addition, the use of the PROMETHEE II method allows to study the relation between all criteria in the decision-making problem. The analysis of Fig. 5B shows that criteria C1 and C2 and similar to each other in terms of preferences (because the size of the requested sample is related to the time needed to generate one), but they both are in strong conflict with the C3 criterion (coverage accuracy). Moreover, PROMETHEE II allows to group criteria into clusters. Criteria C1 and C2 can be grouped into ”Cost” cluster, and criteria C3 and C4 into ”Accuracy” cluster. This allows to observe, that the biggest network sample is strongly supported by the ”Accuracy” criteria (Fig. 7B), whereas the most lightweight 10% network sample is strongly supported by the ”Cost” criteria (Fig. 7A).

Last, but not least, it should be noted that results obtained with the MCDA method in this section are not universal. Should the preferences of the marketer regarding the importance of criteria change, the rankings would also be adjusted. For example, if the marketer decided that the generation time of the network samples (criterion C2) is even slightly less significant than the rest of the criteria, the 50% network would outrun the 10% one and would become the most preferable choice. Therefore, each ranking obtained with PROMETHEE II has its stability intervals. In case of the ranking from Table 6C, the ranking is stable within the following intervals of the weights of each criteria: C1 – [24.53% – 100%], C2 – [24.01% – 100%], C3 – [0.00% – 25.27%], C4 – [23.28% – 89.27%].

5. Conclusions

Nowadays, due to rapidly growing popularity of social media platforms [21], and better spreading of information through the complex networks than through traditional marketing tools, more online marketers have been investing effort into seeding marketing information into social networks. Various viral marketing campaign strategies are possible and it is beneficial to perform simulations and choose the best strategy prior to executing the real campaign.

The approach presented in this paper uses network samples of reduced size compared to the real network, yet still of satisfactory accuracy, to facilitate the process of viral marketing campaign selection. The main contributions of the presented study include:

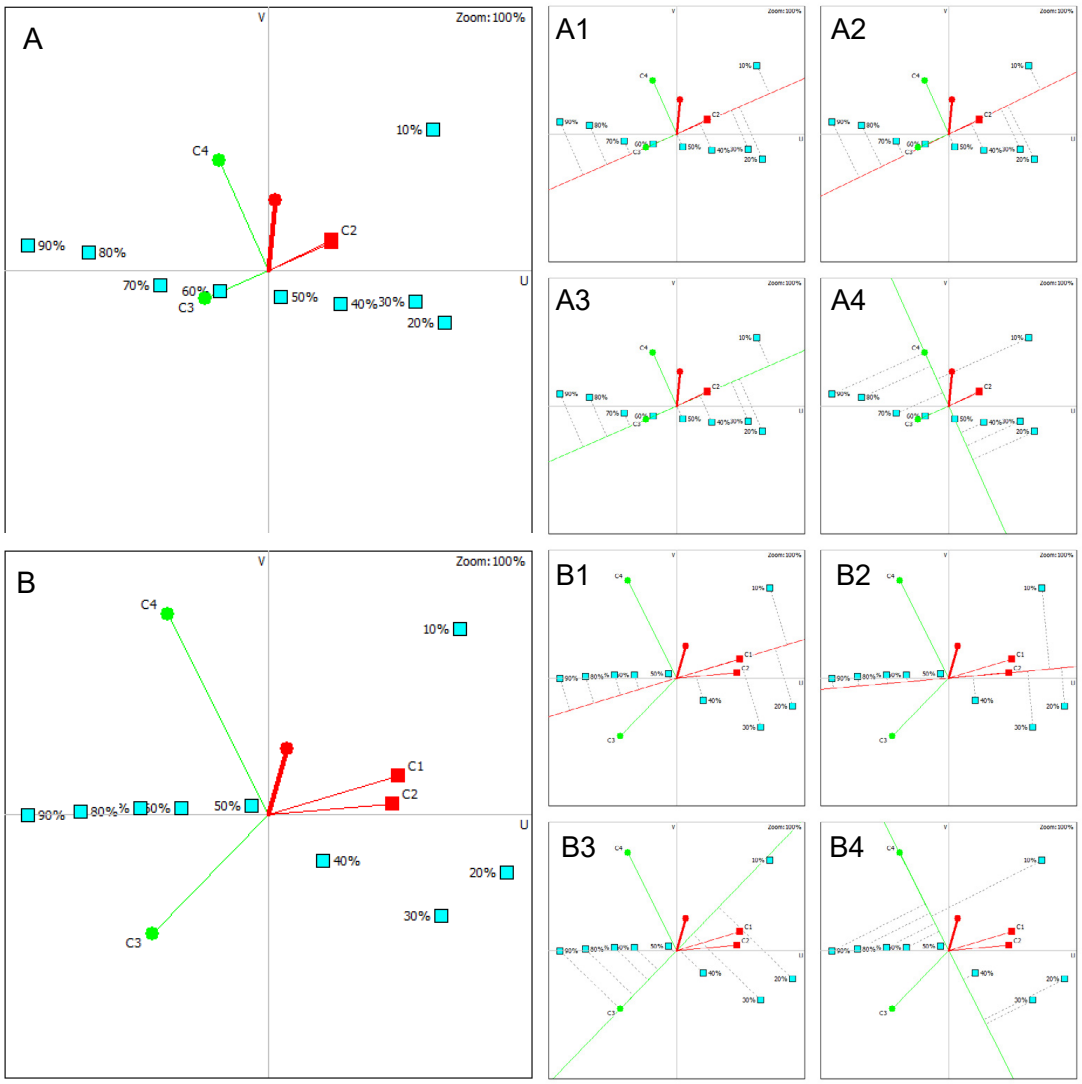


Fig. 5. GAIA visual analysis of the evaluation of sample networks. A - usual preference function. B - linear preference function.

- a two-track framework for selection of network sample for viral marketing campaign planning, which on the one hand reduces the computational requirements for the planning process, yet remains satisfactory in terms of accuracy, and on the other hand considers the preferences of the online marketer;
- a detailed study of how the size of the network sample affects the simulations' coverage ratio and information spreading duration ratio compared to the original real network.

The research has identified possible areas of improvement and future works. First of all, the research was based exclusively on the snowball sampling technique. Other sampling techniques could be explored with the framework proposed in this paper. Additionally, only four criteria were used for the evaluation. More criteria could be introduced to the evaluation model.

6. Acknowledgments

This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562.

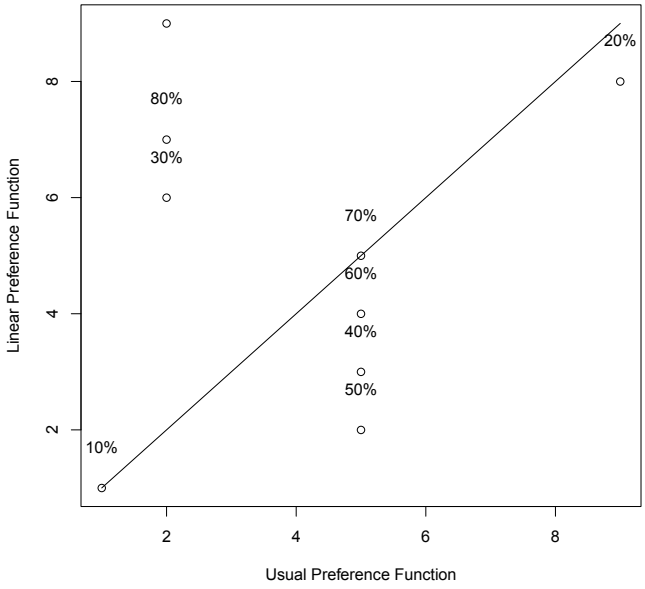


Fig. 6. Comparison of the PROMETHEE II rankings for the usual preference function and the linear preference function with indifference and preference thresholds.

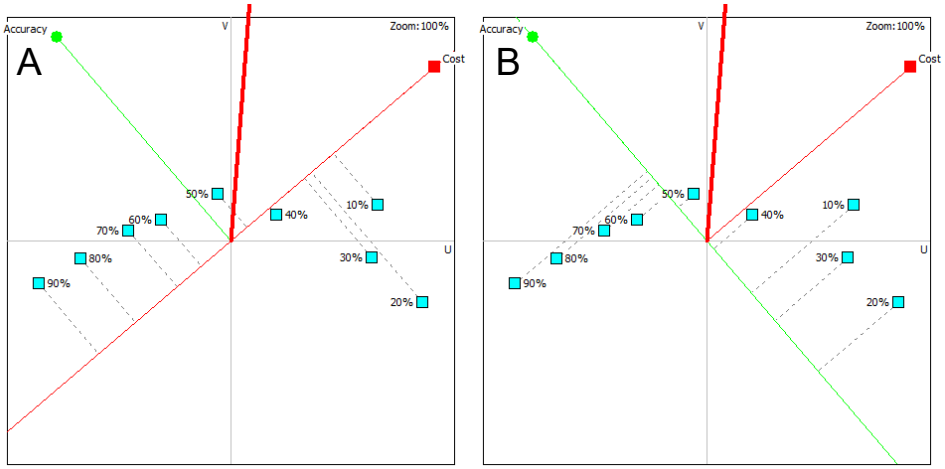


Fig. 7. GAIA visual analysis of the evaluation of sample networks with cost and accuracy criteria grouped.

References

[1] Bampo, M., Ewing, M.T., Mather, D.R., Stewart, D., Wallace, M., 2008. The effects of the social structure of digital networks on viral marketing performance. Information systems research 19, 273–290.
[2] Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: Recent achievements and new challenges. Information Fusion 28, 45–59.
[3] Berger, J., Milkman, K.L., 2012. What makes online content viral? Journal of marketing research 49, 192–205.
[4] Brans, J.P., Mareschal, B., 2005. Promethee methods, in: Multiple criteria decision analysis: state of the art surveys. Springer, pp. 163–186.
[5] Camarero, C., San José, R., 2011. Social and attitudinal determinants of viral marketing dynamics. Computers in Human Behavior 27, 2292–2300.
[6] Chen, W., Wang, C., Wang, Y., 2010. Scalable influence maximization for prevalent viral marketing in large-scale social networks, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 1029–1038.
[7] Dobele, A., Lindgreen, A., Beverland, M., Vanhamme, J., Van Wijk, R., 2007. Why pass on viral messages? because they connect emotionally. Business Horizons 50, 291–304.
[8] Figueira, J., Mousseau, V., Roy, B., 2005. Electre methods, in: Multiple criteria decision analysis: State of the art surveys. Springer, pp.

133–153.
[9] Guitouni, A., Martel, J.M., 1998. Tentative guidelines to help choosing an appropriate MCDA method. European Journal of Operational Research 109, 501–521. URL: http://www.sciencedirect.com/science/article/pii/S0377221798000733, doi:10.1016/S0377-2217(98)00073-3.
[10] Hanna, R., Rohm, A., Crittenden, V.L., 2011. Were all connected: The power of the social media ecosystem. Business horizons 54, 265–273.
[11] Hinz, O., Skiera, B., Barrot, C., Becker, J.U., 2011. Seeding strategies for viral marketing: An empirical comparison. Journal of Marketing 75, 55–71.
[12] Ho, J.Y., Dempsey, M., 2010. Viral marketing: Motivations to forward online content. Journal of Business research 63, 1000–1006.
[13] Holme, P., Saramäki, J., 2012. Temporal networks. Physics reports 519, 97–125.
[14] Iribarren, J.L., Moro, E., 2009. Impact of human activity patterns on the dynamics of information diffusion. Physical review letters 103, 038702.
[15] Jankowski, J., Bródka, P., Kazienko, P., Szymanski, B.K., Michalski, R., Kajdanowicz, T., 2017. Balancing speed and coverage by sequential seeding in complex networks. Scientific reports 7, 891.
[16] Jankowski, J., Hamari, J., Watróbski, J., 2019. A gradual approach for maximising user conversion without compromising experience with high visual intensity website elements. Internet Research 29, 194–217.
[17] Jankowski, J., Kolomvatsos, K., Kazienko, P., Watróbski, J., 2016. Fuzzy modeling of user behaviors and virtual goods purchases in social networking platforms. J. UCS 22, 416–437.
[18] Kandhway, K., Kuri, J., 2014. How to run a campaign: Optimal control of sis and sir information epidemics. Applied Mathematics and Computation 231, 79–92.
[19] Karczmarczyk, A., Jankowski, J., Watróbski, J., 2018a. Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one 13, e0209372.
[20] Karczmarczyk, A., Jankowski, J., Wtrbski, J., 2018b. Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PLOS ONE 13, e0209372. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0209372, doi:10.1371/journal.pone.0209372.
[21] Kemp, S., 2019. Digital 2019: Global Internet Use Accelerates. URL: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates.
[22] Kempe, D., Kleinberg, J., Tardos, É., 2003. Maximizing the spread of influence through a social network, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 137–146.
[23] Kiss, C., Bichler, M., 2008. Identification of influencers: measuring influence in customer networks. Decision Support Systems 46, 233–253.
[24] Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., Makse, H.A., 2010. Identification of influential spreaders in complex networks. Nature physics 6, 888.
[25] Liu-Thompkins, Y., 2012. Seeding viral content: The role of message and network factors. Journal of Advertising Research 52, 465–478.
[26] Michalski, R., Kajdanowicz, T., Bródka, P., Kazienko, P., 2014. Seed selection for spread of influence in social networks: Temporal vs. static approach. New Generation Computing 32, 213–235.
[27] Rezvanian, A., Moradabadi, B., Ghavipour, M., Khomami, M.M.D., Meybodi, M.R., 2019. Social network sampling, in: Learning Automata Approach for Social Networks. Springer, pp. 91–149.
[28] Ripeanu, M., Foster, I., Iamnitchi, A., 2002. Mapping the Gnutella Network: Properties of Large-Scale Peer-to-Peer Systems and Implications for System Design. arXiv:cs/0209028 URL: http://arxiv.org/abs/cs/0209028. arXiv: cs/0209028.
[29] Rogers, E.M., 2010. Diffusion of innovations. Simon and Schuster.
[30] Roy, B., Vanderpooten, D., 1996. The European school of MCDA: Emergence, basic features and current works. Journal of Multi-Criteria Decision Analysis 5, 22–38. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291099-1360%28199603%295%3A1%3C22%3A%3AAID-MCDA93%3E3.0.CO%3B2-F, doi:10.1002/(SICI)1099-1360(199603)5:1<22::AID-MCDA93>3.0.CO;2-F.
[31] Saaty, T.L., 2004. Decision makingthe analytic hierarchy and network processes (ahp/anp). Journal of systems science and systems engineering 13, 1–35.
[32] Sałabun, W., 2014. Reduction in the number of comparisons required to create matrix of expert judgment in the comet method. Management and Production Engineering Review 5, 62–69.
[33] Sałabun, W., 2015. The characteristic objects method: A new distance-based approach to multicriteria decision-making problems. Journal of Multi-Criteria Decision Analysis 22, 37–50.
[34] Sałabun, W., Piegat, A., 2017. Comparative analysis of mcdm methods for the assessment of mortality in patients with acute coronary syndrome. Artificial Intelligence Review 48, 557–571.
[35] Salehi, M., Sharma, R., Marzolla, M., Magnani, M., Siyari, P., Montesi, D., 2015. Spreading processes in multilayer networks. IEEE Transactions on Network Science and Engineering 2, 65–83.
[36] Seeman, L., Singer, Y., 2013. Adaptive seeding in social networks, in: Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on, IEEE. pp. 459–468.
[37] Stieglitz, S., Dang-Xuan, L., 2013. Emotions and information diffusion in social media: sentiment of microblogs and sharing behavior. Journal of management information systems 29, 217–248.
[38] Tang, J., Musolesi, M., Mascolo, C., Latora, V., Nicosia, V., 2010. Analysing information flows and key mediators through temporal centrality metrics, in: Proceedings of the 3rd Workshop on Social Network Systems, ACM. p. 3.
[39] Watróbski, J., 2016. Outline of multicriteria decision-making in green logistics. Transportation Research Procedia 16, 537–552.
[40] Watróbski, J., Jankowski, J., 2015. Knowledge management in mcda domain, in: 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE. pp. 1445–1450.
[41] Watróbski, J., Jankowski, J., 2016. Guideline for mcda method selection in production management area, in: New frontiers in information and

production systems modelling and analysis. Springer, pp. 119–138.
[42] Watróbski, J., Jankowski, J., Piotrowski, Z., 2014. The selection of multicriteria method based on unstructured decision problem description, in: International Conference on Computational Collective Intelligence, Springer. pp. 454–465.
[43] Watróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., Zioło, M., 2019a. Generalised framework for multi-criteria method selection. Omega 86, 107–124.
[44] Watróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., Zioło, M., 2019b. Generalised framework for multi-criteria method selection: Rule set database and exemplary decision support system implementation blueprints. Data in brief 22, 639.
[45] Watts, D.J., Peretti, J., Frumin, M., 2007. Viral marketing for the real world. Harvard Business School Pub.

A5.

Karczmarczyk, A., Wątróbski, J., Jankowski, J. (2019). Multi-Criteria Approach to Planning of Information Spreading Processes Focused on Their Initialization With the Use of Sequential Seeding. In Information Technology for Management: Current Research and Future Directions (pp. 116-134). Springer, Cham.

Multi-criteria Approach to Planning of Information Spreading Processes Focused on Their Initialization with the Use of Sequential Seeding

Artur Karczmarczyk[1], Jarosław Wątróbski[2(✉)], and Jarosław Jankowski[1]

[1] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland
{artur.karczmarczyk,jaroslaw.jankowski}@zut.edu.pl
[2] University of Szczecin, Mickewicza 64, 71-101 Szczecin, Poland
jaroslaw.watrobski@usz.edu.pl

Abstract. Information spreading within social networks and techniques related to viral marketing has begun to attract more interest of online marketers. While much of the prior research focuses on increasing the coverage of the viral marketing campaign, in real-life applications also other campaign goals and limitations need to be considered, such as limited time or budget, or assumed dynamics of the process. This paper presents a multi-criteria approach to planning of information spreading processes, with focus on the campaign initialization with the use of sequential seeding. A framework and example set of criteria was proposed for evaluation of viral marketing campaign strategies. The initial results showed that an increase of the count of seeding iterations and the interval between them increases the achieved coverage at the cost of increased process duration, yet without the need to increase seeding fraction or to provide incentives for increased propagation probability.

Keywords: Social networks · Complex networks · Viral marketing campaign planning · Viral marketing campaign evaluation · MCDA · TOPSIS · Sequential seeding

1 Introduction

Social media platforms have evolved from early stage technical systems to widely used online platforms with integrated mechanics of their users' interactions close to the real world [1–3]. Marketers have turned their focus on social networks due to the fact that the trust users give to their fellow users result in better information propagation than traditional marketing communication methods. This, in turn, results in possibility to obtain high information coverage in the network with relatively slow advertising budgets, which apart from theoretical studies, was presented for real viral campaigns [4–6].

E. Ziemba (Ed.): AITM 2019/ISM 2019, LNBIP 380, pp. 116–134, 2020.
https://doi.org/10.1007/978-3-030-43353-6_7

The up-to-date research focuses on increasing coverage [7], information spreading processes dynamics [8,9] and coverage prediction [10]. The information spreading processes are studied on complex networks of static and dynamic nature [11–13]. Moreover, there is research on modifying the structure of the network to increase the coverage [14–16].

Researchers can base their studies on real network models, which can be obtained from numerous network repositories. However, the availability of real network models is limited, sometimes outdated, or the information about the network nodes is limited. Therefore, oftentimes theoretic models are used to produce synthetic networks for research. These networks are parametrised, which allows to focus the research efforts on particular characteristics of the network. The theoretical models used most often are Barabasi-Albert [17], Watts-Strogatz [18] and Erdos-Renyi [19]. Moreover, a prior research exists based on samples and partially observed networks [20,21].

Viral marketing campaigns, from practical point of view, are often based on various strategies to achieve the assumed campaign goals. Although the majority of research in this field focuses on maximising the number of nodes infected in the network (i.e. increasing coverage), the actual goals and means of the marketer might vary [22]. Whilst full network coverage with the information would be an ideal outcome of the campaign, the limited campaign budget and time might render such outcome unfeasible. Different marketing campaign strategy would be selected for obtaining immense count of potential clients in short period of time, and different when the marketer aims at slow yet steady development of the customer base [23]. For these reasons, tools for planning and evaluating viral marketing campaigns are needed [24]. Approaches based on multi-criteria decision analysis (MCDA) methods proved to be useful for assessing which strategy, based on what parameters' values, would best accomplish the goals of particular campaigns [25,26].

The authors' prior research studied the possibility to reduce the computational complexity of viral marketing campaign planning with the use of theoretical models and network samples. Moreover, it was demonstrated in [27] that re-initialising the campaign multiple times for yet-uninfected initial nodes (seeds) allows to increase the final coverage compared to the traditional approach at the cost of increasing duration of the process, compared to the traditional one-off approaches. Whilst such innovative approach would not be satisfactory for campaign goals focused on achieving high coverage within shortest possible time span, it can bring promising effects for campaign goals focused on maximizing coverage or extending the duration of the company's appearance in the social media. Therefore, in this paper, the authors have extended the parameters' set from [28] with the criteria of count of seeding iterations and the interval between them.

This paper consists of five main sections. After this introduction, a literature review is presented in Sect. 2, followed by the methodological framework in Sect. 3. Section 4 presents the empirical study of the proposed framework and discusses the obtained results. Eventually, the conclusions and possible future works are presented in Sect. 5.

2 Literature Review

Social media platforms collect information on user behaviour and social relations in order to better address marketing campaigns. Social networks can be used as a tool using social influence mechanisms [29,30] to spread information among acquaintances and, in next hops, their acquaintances and so on. Social networks have been used in parliament elections in Poland [31] or presidential elections in USA [32]. Due to its complex nature, the research on information propagation in complex networks is based on interdisciplinary efforts from fields such as sociology, computer studies, physics and management, based on various theoretical and practical foundations and goals [33].

In order to profoundly study the processes of information diffusion in complex networks, often theoretic models and sample real networks are used and implemented in agent-based environments. The methodological background is often based on models such as SIR and SIS, invented in order to study epidemics [34]. The diffusion process is often verified from the microscopic level with the use of linear threshold [30] and independent cascades [35,36] models.

The majority of research on information propagation in complex networks is founded on the selection of the initial clients in the form of seeds. These clients are provided with product samples or other marketing materials with a hope that they will pass the information about the product to their acquaintances in the network [7]. This problem is NP-hard, but some greedy solutions exist, which provide relatively good results, but at high computational cost [35]. More practical solutions are based on heuristic approaches, where the seeds are selected based on centrality measures such as degree, betweenness, page rank, eigenvector or closeness [37].

Moreover, the knowledge about information propagation in given network can be collected and reused in that network for improving the characteristics of the information propagation process. The literature review resulted in finding some adaptive approaches [38], spreading the seeds over time in order to better use the natural information diffusion process [39], voting mechanics [40] or k-shell based approach [41].

On the other hand, there are approaches which instead of focusing on the initial seeds, try to increase the incentives provided to the network users in order to motivate the subsequent users on the information propagation process to pass the information further in the network (increasing the propagation probability) [42].

The authors' prior study showed that viral marketing campaigns can be planned with the use of significantly smaller theoretical models. Although the used synthetic networks were much smaller and less computationally complex, the correlation coefficient of the results on the synthetic network and the real network exceeded the value of 0.9. The literature review shows that the viral campaign strategy evaluation criteria utilised in [28] can be further extended with sequential seeding [43]. More than a single seeding iteration can occur. Moreover, the seeding iterations can either follow each other without breaks, or be dispersed over the campaign with a specified interval. This, in turn, states an interesting research question, if extending the decision model for selection of

viral marketing campaign strategies with these new parameters would allow for better adjustment of the campaign to the marketers' needs.

3 Research Methodology

Viral marketing campaigns can be based on various strategy. During the campaign planning stage, decisions need to be made regarding the fraction of the nodes that need to be initially infected (seeds), how these nodes should be selected, what incentives should be provided to the network members to increase the information propagation probability, to name just a few. The values of the aforementioned parameters are adjusted based on the overall goals of the planned campaign. Certainly, different campaign efficiency metrics would be expected when maximal coverage is desired, than when the campaign ordering party is most focused on making the campaign last as long as possible.

Mappings of real networks, as well as theoretical models having structure similar to the real networks, allow researchers to use tools such as the independent cascades model in order to evaluate various campaign strategies prior to the actual campaign execution and to choose the one which best accommodates the campaign goals. Nonetheless, evaluations based on the real network models are complex computationally. Accordingly, it is often worth to conduct the study and simulations on significantly smaller (e.g. 10%, 30%, 50% of the real network's size) theoretical models. Despite their reduced size, the theoretical models maintain the characteristics of the real network. In the proposed approach (see Fig. 1), multiple theoretical networks are generated based on various configurations of the models' parameters, followed by the selection of the model most similar to the real network with the use of the Kullback-Leibler divergence (KLD) metric.

In the proposed approach, simulations are then computed for various campaign strategies, based on multiple parameters. In this research, the authors suggest a set of 5 input parameters for the simulations:

Par1 – seeding fraction
: The fraction of the total nodes that are selected to be originally provided with the information to further spread.

Par2 – propagation probability
: The assumed probability of passing information from one infected node to other non-infected nodes. The level of the propagation probability can be adjusted by providing incentives to the users to pass the information.

Par3 – seeding iterations' count
: In the original research [28], a single seeding was performed only to bootstrap the information propagation process. Based on the initial successes obtained in [28], the authors have decided to extend the original model with sequential seeding procedures. This parameter specifies how many times, in total, the information will be input to the network (including both the initial seeding and subsequent ones).

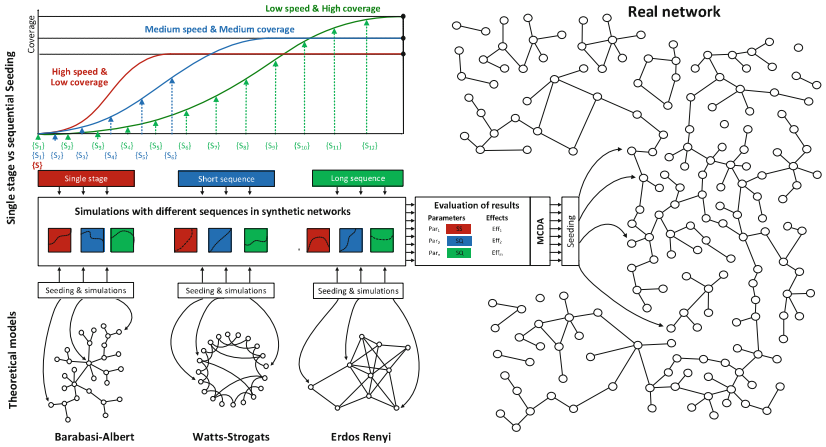


Fig. 1. Conceptual framework of the proposed approach

Par4 – seeding iteration's interval
: The initial seeding is always performed in the first iteration. If there are more than a single seeding iteration (see Par3), Par4 specifies what is the interval between each reseeding procedures in the network.

Par5 – measure used to rank nodes
: The nodes selected for the initial seeding of information are not selected randomly. First they are ordered by a chosen metric and then the top ones are selected. Each possible metric is characterized by a specific computational cost.

Afterwards, the simulations are being computed and the results are saved. Consequently, two performance parameters are obtained:

Eff6 – iteration of last infection
: This represents the moment in simulation when the last infection happened, i.e. when the information propagation process died out. The value of this parameter is at least $1 + (Par3 - 1) \times Par4$.

Eff7 – total coverage
: This is the total coverage achieved by the simulations for a strategy based on the Par1–Par5 parameters, i.e. the ratio of infected nodes to the total nodes.

Last, but not least, after the simulations are concluded and efficiency parameters for all campaign strategies are collected, all gathered information is used to build a decision matrix for the strategies' evaluation.

The procedure of evaluation of the viral campaign strategies requires to consider multiple criteria Par1–Eff7, some of which are conflicting with each other. Depending on the campaign goals, the importance (weight) assigned by the deciding party to each of the criteria might be different. Both weights and

the actual values of the parameters Par1–Eff7 can be expressed on a quantitative scale. Moreover, because the evaluation is based on a complete set of simulations, the problem of uncertainty of data is eliminated here. Eventually, the ultimate goal of the evaluation is to rank all the possible strategies and choose the best one. Due to the said reasons, the authors, using the generalised framework for MCDA methods' selection [44,45], have decided to use the TOPSIS method for evaluating the viral marketing campaign strategies in the proposed approach.

4 Empirical Study

The proposed framework can be used to evaluate viral marketing campaign strategies on real networks, as well as to plan campaigns on considerably smaller synthetic networks. During the empirical study, first such real network campaign strategy evaluation was performed, followed by planning a strategy for social network viral marketing campaign based on two separate goals. Eventually, the effect of Par4 and Par5 parameters on the strategies was studied.

4.1 Evaluation of Viral Marketing Campaign Strategies on a Real Network

The evaluation of real network viral marketing campaign strategies was based on the Gnutella network [46], precisely its snapshot from 2002. The network contains 8846 nodes and 31839 edges. Its metrics are presented in Table 1.

In order to ascertain repeatability of the study regardless of the chosen parameters, ten information propagation scenarios were prepared for the network. For each edge between each two nodes a random value was drawn in such scenarios, which subsequently, during the simulations, was used to decide whether or not the infection passed from one node to the other.

A total of 91000 simulations was performed, based on the parameters' values presented in Table 2.

Table 1. The metrics of the real network [46]

Metric	Symbol	Value
Total degree	D	7.1985
Closeness	C	1.587441e−07
Page rank	PR	0.0001130454
Eigen vector	EV	0.01602488
Clustering coefficient	CC	0.0001130838
Betweenness	B	19104.87

Table 2. Simulation parameters

Criterion	Values
Par1	0.01, 0.02, 0.03, 0.04, 0.05
Par2	0.1, 0.2, 0.3, 0.4, 0.5
Par3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Par4	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Par5	degree (1), closeness (2), EV centrality (3), betweenness (4)

Next, the Eff6 and Eff7 parameter values were computed and the decision matrix for the TOPSIS method was built, containing 9100 strategy alternatives (A1 − A9100). For the evaluation, each of the 7 criteria were given an equal importance weight. The assumed goal of the campaign was to cover as much of the network as possible within fastest possible time and with minimum possible costs. Therefore, the impact of Par1–Eff6 criteria on the strategy score was negative and the Eff7 impact was positive. The top 20 strategies are presented along with their scores in Table 3. The analysis of the table shows that most of the winning strategies were based on a very low value of Par1 (0.01–0.02), low values of Par2 ($\leq$0.3). Regarding Par5, only degree and closeness measures can be found on the top 20 strategies' list, however degree is the measure used in the top-scored alternatives. Because one of the goals of the campaign was to minimize the campaign duration, Par3 and Par4 tend to have low values ($\leq$3). The leading strategy A367 obtained the coverage value of 0.5174 within 14.4 iterations, whereas the runner-up alternative A731 resulted in higher coverage and shorter duration, however the propagation probability was higher, which potentially could lead to higher costs.

4.2 Selection of Synthetic Networks

The proposed MCDA approach allows to successfully evaluate multiple viral marketing campaign strategies conducted over a real network, as it was presented above. Nevertheless, simulations over real networks, characterised by multitude of nodes and edges, require significant computational power. Moreover, the full real network mapping might not always be available. However, it is possible to perform simulations on considerably smaller yet accurate synthetic networks before executing the campaign on the real network.

In the subsequent part of the empirical research, the authors used synthetic networks of 50% size of the real network to perform the simulations and plan a potential campaign on the real network. The results were then compared to the ranking obtained for the real network.

Table 3. Top 20 strategies for the real network

Rank	Alternative	Score	Par1	Par2	Par3	Par4	Par5	Eff6	Eff7
1	A367	0.8454657	0.01	0.2	1	1	1	14.4	0.517386389
2	A731	0.844435	0.01	0.3	1	1	1	9.5	0.683404929
3	A371	0.8391214	0.01	0.2	2	1	1	14.7	0.517386389
4	A735	0.8386404	0.01	0.3	2	1	1	9.8	0.683404929
5	A739	0.8326058	0.01	0.3	2	2	1	9.8	0.683404929
6	A375	0.8323418	0.01	0.2	2	2	1	14.9	0.517397694
7	A2187	0.8320079	0.02	0.2	1	1	1	13.2	0.518923807
8	A2551	0.8298695	0.02	0.3	1	1	1	8.6	0.683563192
9	A775	0.8268715	0.01	0.3	3	1	1	10	0.683404929
10	A411	0.8261902	0.01	0.2	3	1	1	15.1	0.517386389
11	A2191	0.8253991	0.02	0.2	2	1	1	13.7	0.518923807
12	A2555	0.8238216	0.02	0.3	2	1	1	9.2	0.683563192
13	A366	0.8237788	0.01	0.2	1	1	2	14.2	0.517635089
14	A730	0.8233488	0.01	0.3	1	1	2	9.4	0.683495365
15	A779	0.8209285	0.01	0.3	3	2	1	10	0.683563192
16	A415	0.8200903	0.01	0.2	3	2	1	15.1	0.517895094
17	A743	0.8200089	0.01	0.3	2	3	1	9.8	0.683529279
18	A379	0.8193105	0.01	0.2	2	3	1	14.9	0.51786118
19	A2195	0.8189998	0.02	0.2	2	2	1	13.8	0.519262944
20	A734	0.8176981	0.01	0.3	2	1	2	9.7	0.683495365

A set of 15 synthetic networks was generated based on 3 theoretic models with 5 sets of parameters each:

1. BA – number of edges m to add in each step equal to $1, 2, ..., 5$;
2. WS – the neighborhood within which the vertices of the lattice will be connected equal to $1, 2, ..., 5$;
3. ER – number of edges equal to the number of nodes multiplied by $1, 2, ..., 5$.

Kullback-Leibler divergence (KLD) measure was used in order to avoid arbitrary selection of the theoretic model network. Based on the KLD measure of the degree distribution of each of the generated networks with the real network, the network BA-4423-5 was selected, i.e. Barabasi-Albert model with 50% nodes of the real network and $m = 5$. The network characteristics and ints KLD measure value are presented in Table 4.

Table 4. Kullback-Leibler divergence measure for the selected synthetic network

Edges to add	Num of nodes	Num of edges	Perc. of edges	KLD
5	4423	22100	0.694117278%	0.000521317

4.3 Planning of the Viral Marketing Campaign Strategies

Two opposite campaign goals were studied during this part of the empirical study: maximization of coverage within smallest possible time, and maximization of coverage and process duration.

Maximization of Coverage and Minimization of Duration. For the evaluation, each of the 7 criteria were given an equal importance weight. The assumed impact on the strategy score of Par1–Eff6 criteria was negative and the Eff7 impact was positive. The top 20 strategies are presented along with their scores in Table 5.

Similarly to the evaluation of strategies for the real-network, also in case of the synthetic network the top strategies were based on a very low value of the Par1 parameter (0.01–0.02), low values of the Par2 parameter (0.2–0.3). The winning strategies were based on the degree and closeness centrality measures. The coverage for the first 9 strategies is almost equal, i.e. 0.7026 and the duration oscillates around 11 s. The strategy A1936, ranked 9, is based on splitting the seeding process into two steps, in the first and fourth iteration of the process. This results in an equal coverage to the winning strategy, and the duration averagely extended by 0.1 s.

It is important to note, that the above-mentioned ranking was generated for equal weights of all criteria. In order how the weights of individual criteria affect the final rankings, a sensitivity analysis was performed. However, for the reasons of readability, the sensitivity analysis was limited to the top 20 strategies. The results of the analysis are presented on Fig. 2 for the evaluation scores and Fig. 3 for the evaluation ranks.

The analysis of Fig. 2 allows to see how the scores of each strategy are affected by the increase of weight of each criteria. For example, it can be observed that the leading alternative A1914 is supported by criteria Par1, Par3, Par4, Par5 and Eff6, i.e. its score raises along with the raise of the importance of each of these criteria. On the other hand, if more importance was given to the Par2 or Eff7 criteria, the score of strategy A1914 would decrease. The CCi sensitivity chart allows also to see how much each criterion supports or conflicts each alternative.

Table 5. Top 20 strategies for the synthetic network for maximum coverage and minimum duration

Rank	Strategy	Score	Par1	Par2	Par3	Par4	Par5	Eff6	Eff7
1	A1914	0.8776064	0.01	0.2	1	1	1	11.2	0.702645263
2	A1918	0.871181	0.01	0.2	2	1	1	11.3	0.702645263
3	A1922	0.8643984	0.01	0.2	2	2	1	11.3	0.702645263
4	A1940	0.8596561	0.02	0.2	1	1	1	10.7	0.702645263
5	A1929	0.8567826	0.01	0.2	3	1	1	11.4	0.702645263
6	A1943	0.8531567	0.02	0.2	2	1	1	11	0.702645263
7	A1913	0.8513274	0.01	0.2	1	1	2	11.3	0.702645263
8	A1973	0.8507025	0.01	0.2	3	2	1	11.3	0.703640063
9	A1926	0.8489333	0.01	0.2	2	3	1	11.4	0.702645263
10	A3825	0.8464421	0.01	0.3	1	1	1	8.3	0.888175447
11	A1946	0.8462135	0.02	0.2	2	2	1	11.2	0.702645263
12	A1917	0.8456965	0.01	0.2	2	1	2	11.3	0.702645263
13	A3829	0.8408425	0.01	0.3	2	1	1	8.6	0.888175447
14	A1949	0.8395764	0.02	0.2	3	1	1	11.3	0.702645263
15	A1921	0.8394273	0.01	0.2	2	2	2	11.3	0.702645263
16	A2071	0.8384225	0.01	0.2	3	3	1	10.8	0.709586254
17	A1933	0.8363776	0.01	0.2	4	1	1	11.8	0.702645263
18	A1939	0.8357937	0.02	0.2	1	1	2	10.7	0.702645263
19	A3833	0.8350406	0.01	0.3	2	2	1	8.6	0.888175447
20	A2020	0.8347077	0.02	0.2	3	2	1	10.8	0.70594619

For example, Fig. 2C shows that the increase of weight of Par3 decreases the score of both strategies A1933 and A2020. If the weights of all criteria are equal, the strategy A1933 is ranked 17 and A2020 is ranked 20. However, the raise of importance of criterion Par3 affects the strategy A1933 more than A2020, and in case of a 10% increase of the importance of Par3, the rank of strategy A1933 drop below the rank of strategy A2020.

On the other hand, Fig. 3 allows to easily track how the ranks of the strategies change along with changes of weights of individual criteria. It can be clearly observed that the rank of the leading strategy A1914 is very stable and only considerable changes of Eff7 criterion weight can result in change of its rank from 1 to 9 (see Fig. 3G). Furthermore, Fig. 3A and E show that in case of considerable changes of weights of criteria Par1 and Par5, the strategies from the bottom of the top-20 list would be replaced by strategies previously outside of the top-20 list.

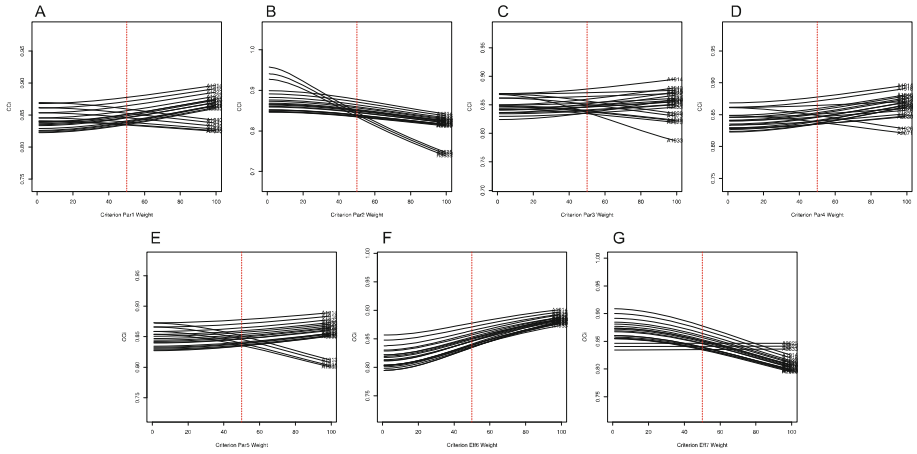


Fig. 2. Sensitivity analysis for the CCi values fot the top 20 strategies for the synthetic network scenario maximizing coverage and minimizing duration

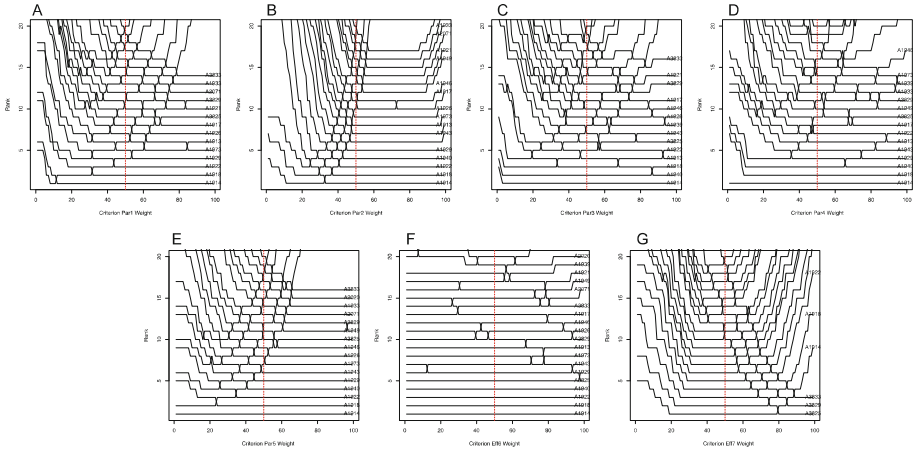


Fig. 3. Sensitivity analysis for the ranks of the top 20 strategies for the synthetic network scenario maximizing coverage and minimizing duration

Maximization of Coverage and Duration. For this scenario, the assumed impact of Par1–Par3 criteria was negative and the Par4–Eff7 impact was positive. Similarly as in the previous experiment, each of the 7 criteria were given an equal importance weight. The top 20 strategies are presented along with their scores in Table 6.

When Table 6 is analysed, it can be observed that the average values of the Par1, Par2, Par5 and Eff7 criteria are very similar for the top 20 alternatives in both scenarios. However, a considerable difference in average values can be observed for Par3, Par4 and Eff6 (see Table 7). While in case of the first scenario,

Table 6. Top 20 strategies for the synthetic network for maximum coverage and maximum duration

Rank	Strategy	Score	Par1	Par2	Par3	Par4	Par5	Eff6	Eff7
1	A2450	0.8872772	0.01	0.2	10	10	1	92.4	0.72233778
2	A2956	0.8714643	0.02	0.2	10	10	1	93.3	0.741419851
3	A2441	0.8604204	0.01	0.2	10	10	2	92.9	0.722156907
4	A2449	0.8566093	0.01	0.2	10	9	1	83.4	0.722292562
5	A2489	0.8533866	0.01	0.2	9	10	1	82.3	0.723965634
6	A4346	0.8468385	0.01	0.3	10	10	1	91.9	0.900474791
7	A2942	0.8455437	0.02	0.2	10	10	2	92.7	0.741058105
8	A2953	0.8443063	0.02	0.2	10	9	1	84.3	0.741397242
9	A2937	0.841499	0.02	0.2	9	10	1	83.3	0.740854624
10	A2437	0.8341213	0.01	0.2	10	9	2	83.9	0.722111689
11	A2487	0.8323953	0.01	0.2	9	10	2	83.1	0.723943025
12	A4822	0.8322019	0.02	0.3	10	10	1	92	0.913361972
13	A3307	0.8297117	0.03	0.2	10	10	1	92.5	0.754171377
14	A4330	0.8263723	0.01	0.3	10	10	2	92	0.900180873
15	A4345	0.8210995	0.01	0.3	10	9	1	82.9	0.900474791
16	A2940	0.8192291	0.02	0.2	10	9	2	83.7	0.740922451
17	A390	0.8187792	0.01	0.1	10	10	1	94.3	0.300836536
18	A4382	0.818645	0.01	0.3	9	10	1	81.9	0.901560027
19	A2931	0.8169775	0.02	0.2	9	10	2	82.8	0.740741578
20	A922	0.8135993	0.02	0.1	10	10	1	95.4	0.332240561

the average duration of the campaign (Eff6) was little lower than 11 iterations, in case of the second scenario the average duration was over 88 iterations. However, at the same time, the average coverage (Eff7) for the top 20 strategies for both scenarios was almost the same. In case of the second scenario, the strategies tended to be based on multiple seeding stages (Par3, averagely 9.75 compared to 2.1 in the first scenario) with long intervals between them (Par4, averagely 9.75, compared to 1.5 in the first scenario).

When the individual winning strategies are analysed in Table 6, it can be observed that the top strategy A2450 was based on 0.01 seeding fraction, 0.2 propagation probability, 10 seeding iterations with 10 iterations interval and degree used as the nodes' selection measure. This resulted in averagely 92.4 iterations and 0.7223 coverage. The second-best strategy A2956 achieved slightly longer duration and better coverage (93.3 and 0.7414 respectively), but it was based on 0.02 seeding fraction, which potentially could lead to higher costs, which would not be compensated by the aforementioned slight increase of coverage and duration of the information spreading process.

Table 7. Comparison of average values of Par1–Eff7 criteria values for the 20 top-ranked alternatives based on synthetic network scenario (A) maximising coverage and minimising duration; and (B) maximising coverage and duration

	Par1	Par2	Par3	Par4	Par5	Eff6	Eff7
(A) Max-Min	0.0130	0.2150	2.1000	1.5000	1.2000	10.7800	0.7310
(B) Max-Max	0.0150	0.2150	9.7500	9.7500	1.3500	88.0500	0.7343
|A − B|	0.0020	0.0000	7.6500	8.2500	0.1500	77.2700	0.0033

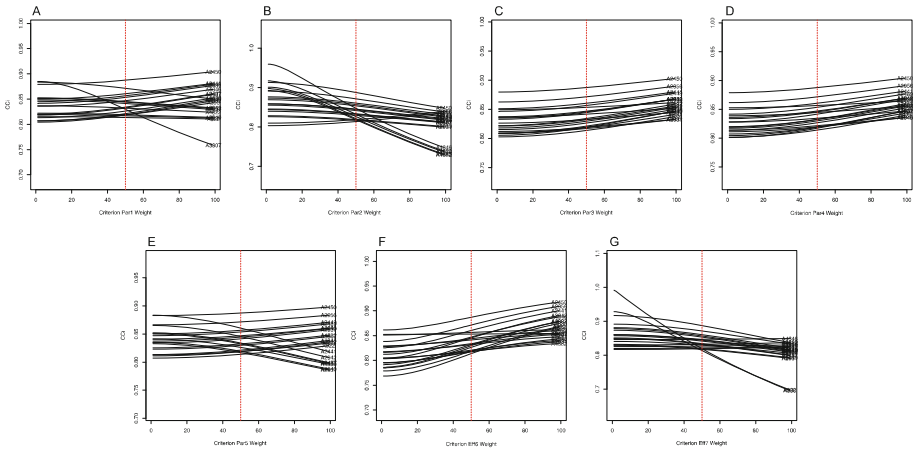


Fig. 4. Sensitivity analysis for the CCi values fot the top 20 strategies for the synthetic network scenario maximizing coverage and duration

Subsequently, a sensitivity analysis was performed. The obtained results are presented on Figs. 4 and 5, however for readability, only the top 20 alternatives are presented on the figures. The analysis of Figs. 4 and 5 allows to observe high stability of the ranks of the leading alternatives. However, the further a strategy is from the top of the ranking in terms of score, the less stable its rank is. This is most clearly visible on Fig. 5B and G.

4.4 Study of Sequential Seeding on Information Propagation Process

The empirical study was concluded by examination of the effect the two new criteria Par3 and Par4, i.e. seeding iterations and interval between them, have on the final coverage and information spreading process duration. Data from simulations performed on the real network [46] was further aggregated and presented on figures Figs. 6, 7 and 8.

The chart on Fig. 6 shows that along with increase of the count of seeding iterations, both the average coverage and process duration increased. The average duration growth can be approximated with a linear function $y = 4.9113x + 2.513$

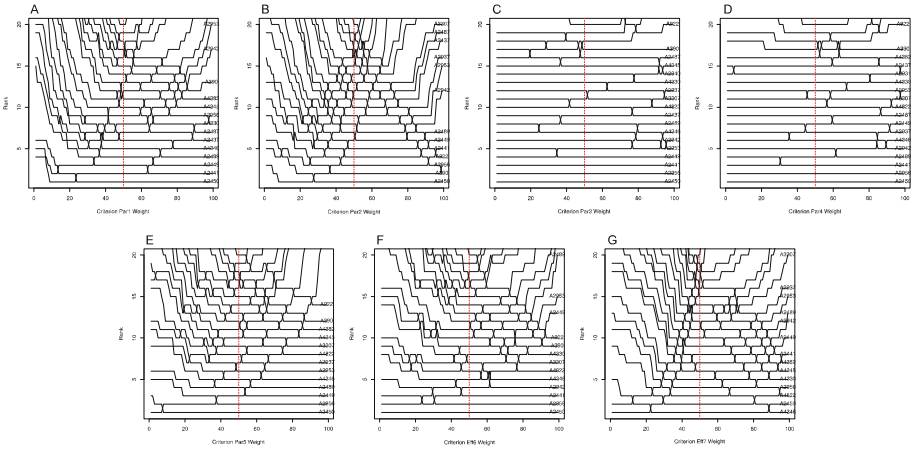


Fig. 5. Sensitivity analysis for the ranks of the top 20 strategies for the synthetic network scenario maximizing coverage and duration

with $R^2 = 0.9934$, whereas the average coverage growth can be approximated with logarithmic function $y = 0.0144ln(x) + 0.5982$ with $R^2 = 0.9831$. Similar increase of average coverage and average information spreading process duration can be observed when the interval between seeding iterations is increased (see Fig. 7). The duration growth can be approximated with a linear function $y = 4.756x + 5.4968$ with $R^2 = 0.9987$, while the average coverage growth can be approximated with logarithmic function $y = 0.0109ln(x) + 0.6062$ with $R^2 = 0.9568$.

Eventually, the simulation results were aggregated and grouped by seeding iterations' count and interval and ordered by both of these factors ascending. The results are presented on Fig. 8. The labels on the X axis of the chart are built of two components C:I, where C denotes the count of seeding iterations and I denotes the interval between them. The visual analysis of this chart allows to observe that the increase of the duration of the process is almost linear to the interval between seeding iterations. This means, that the longer the interval, the proportionally longer the process will last. In case of the average coverage, it can be observed that immediate increase can be achieved if the interval between seeding iterations is increased to 1–6. However, when the interval is increased over 6, the further increase of coverage is only slight. This observation suggests that in case of viral marketing campaign strategies oriented on coverage maximization and process duration minimization it might be more beneficial to increase the seeding iterations' interval only to some extent.

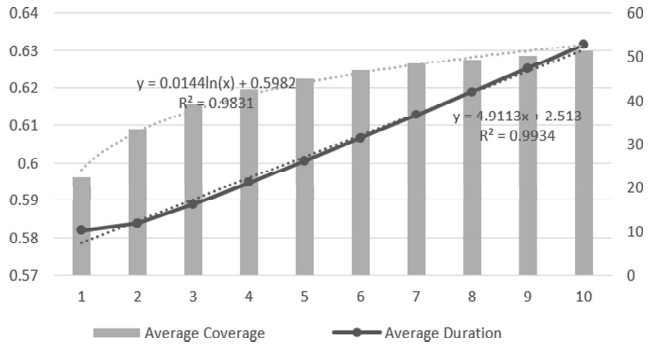


Fig. 6. Effect of seeding iterations' count on coverage and information spreading process duration

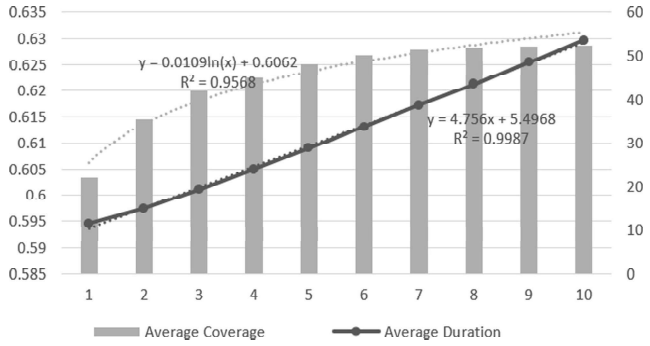


Fig. 7. Effect of seeding iterations' interval on coverage and information spreading process duration

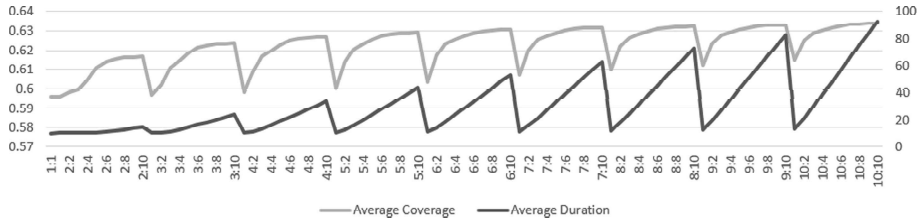


Fig. 8. Effect of seeding iterations' count and interval on coverage and information spreading process duration

5 Conclusions

Social media networks have become very popular and 45% of the population are active social media users. As a result, viral marketing campaigns in social networks began to bring better results than traditional online advertising. Marketers are now investing more effort into seeding information into social networks and

providing incentives to increase the willingness of the users to propagate information further in the network. These increased efforts created a demand for providing manners for campaign planning and evaluation. In recent research the authors proposed a multi-criteria approach for such planning and evaluation.

In this paper, the authors have proposed extension of the multi-criteria approach for viral marketing campaign strategy planning and evaluation, in which strategies utilising sequential seeding are taken into account. This resulted in an evaluation framework containing five parametric criteria and two effectiveness criteria.

The authors' contributions in this paper include:

- multi-criteria framework to planning information spreading processes focused on their initialization with the use of sequential seeding;
- simulation engine for providing data for evaluation of viral marketing campaign strategies performed on real and synthetic networks;
- an example set of criteria to choose a satisfactory viral marketing campaign strategy according to the marketer's goals, taking into account its costs, dynamics, coverage, duration;
- the effect of increasing the count of seeding iterations and of increasing the interval between seeding iterations on the coverage and information propagation process duration was studied.

In practical terms, the empirical study has shown that an increase of the count of seeding iterations in the campaign can increase the achieved network coverage at the cost of the campaign duration increase. Moreover, it was observed that delaying the subsequent seeding iterations by several non-seeding iterations increases the network coverage even more. However, for the studied real network, best coverage increase was observed for 1–6 interval between seeding iterations, and if the interval was increased even further, the effects were less outstanding.

This research has identified some possible areas of improvement and future works. A detailed study of how the sequential seeding affects the process duration and coverage for various sets of other strategy parameters' values could be performed. Moreover, in this research, the seeds in subsequent seeding iterations were chosen based on the original nodes ranking. Possibly, better results could be achieved if the centrality measures for nodes' selection were recalculated in each subsequent seeding iteration. Last, but not least, the introduction of sequential seeding into the viral marketing campaigns calls for studying temporal aspects of the network status and its effect on information diffusion processes.

Acknowledgments. This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562 (AK, JJ) and within the framework of the program of the Minister of Science and Higher Education under the name "Regional Excellence Initiative" in the years 2019–2022, project number 001/RID/2018/19, the amount of financing PLN 10,684,000.00 (JW).

References

1. Greenwood, S., Perrin, A., Duggan, M.: Social media update 2016. Pew Res. Cent. **11**(2) (2016)
2. Couldry, N.: Media, Society, World: Social Theory and Digital Media Practice. Polity Press, Cambridge (2012)
3. Chmielarz, W., Szumski, O.: Digital distribution of video games - an empirical study of game distribution platforms from the perspective of polish students (future managers). In: Ziemba, E. (ed.) AITM/ISM 2018. LNBIP, vol. 346, pp. 136–154. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15154-6_8
4. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. ACM Trans. Web **1**(1), 5–44 (2007). https://doi.org/10.1145/1232722.1232727
5. Camarero, C., José, R.S.: Social and attitudinal determinants of viral marketing dynamics. Comput. Hum. Behav. **27**(6), 2292–2300 (2011). https://doi.org/10.1016/j.chb.2011.07.008
6. Jankowski, J., Bródka, P., Hamari, J.: A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world. Behav. Inf. Technol. **35**(11), 926–945 (2016). https://doi.org/10.1080/0144929X.2016.1212932
7. Hinz, O., Skiera, B., Barrot, C., Becker, J.U.: Seeding strategies for viral marketing: an empirical comparison. J. Mark. **75**(6), 55–71 (2011). https://doi.org/10.1509/jm.10.0088
8. Tang, J., Musolesi, M., Mascolo, C., Latora, V., Nicosia, V.: Analysing information flows and key mediators through temporal centrality metrics. In: Proceedings of the 3rd Workshop on Social Network Systems, p. 3. ACM (2010). https://doi.org/10.1145/1852658.1852661
9. Iribarren, J.L., Moro, E.: Branching dynamics of viral information spreading. Phys. Rev. E **84**, 046116 (2011). https://doi.org/10.1103/PhysRevE.84.046116
10. Jankowski, J., Michalski, R., Kazienko, P.: The multidimensional study of viral campaigns as branching processes. In: Aberer, K., Flache, A., Jager, W., Liu, L., Tang, J., Guéret, C. (eds.) SocInfo 2012. LNCS, vol. 7710, pp. 462–474. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-35386-4_34
11. Liu, C., Zhang, Z.K.: Information spreading on dynamic social networks. Commun. Nonlinear Sci. Numer. Simul. **19**(4), 896–904 (2014). https://doi.org/10.1016/j.cnsns.2013.08.028
12. Kempe, D., Kleinberg, J., Kumar, A.: Connectivity and inference problems for temporal networks. J. Comput. Syst. Sci. **64**(4), 820–842 (2002). https://doi.org/10.1006/jcss.2002.1829
13. Jankowski, J., Michalski, R., Kazienko, P.: Compensatory seeding in networks with varying avaliability of nodes. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), pp. 1242–1249. IEEE (2013). https://doi.org/10.1145/2492517.2500256
14. Ganesh, A., Massoulie, L., Towsley, D.: The effect of network topology on the spread of epidemics. In: Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, pp. 1455–1466, March 2005. https://doi.org/10.1109/INFCOM.2005.1498374
15. Delre, S.A., Jager, W., Bijmolt, T.H.A., Janssen, M.A.: Will it spread or not? The effects of social influences and network topology on innovation diffusion. J. Prod. Innov. Manage. **27**(2), 267–282 (2010). https://doi.org/10.1111/j.1540-5885.2010.00714.x

16. Pazura, P., Jankowski, J., Bortko, K., Bartkow, P.: Increasing the diffusional characteristics of networks through optimal topology changes within sub-graphs (2019). https://doi.org/10.1145/3341161.3344823
17. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999). https://doi.org/10.1126/science.286.5439.509
18. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440 (1998). https://doi.org/10.1038/30918
19. Erdös, P., Rényi, A.: On random graphs I. Publicationes Mathematicae Debrecen **6**, 290 (1959)
20. Onnela, J.P., Christakis, N.A.: Spreading paths in partially observed social networks. Phys. Rev. E **85**, 036106 (2012). https://doi.org/10.1103/PhysRevE.85.036106
21. Génois, M., Vestergaard, C.L., Cattuto, C., Barrat, A.: Compensating for population sampling in simulations of epidemic spread on temporal contact networks. Nat. Commun. **6**, 8860 (2015). https://doi.org/10.1038/ncomms9860
22. Jankowski, J., Hamari, J., Wątróbski, J.: A gradual approach for maximising user conversion without compromising experience with high visual intensity website elements. Internet Res. **29**(1), 194–217 (2019). https://doi.org/10.1108/IntR-09-2016-0271
23. Sałabun, W., Palczewski, K., Wątróbski, J.: Multicriteria approach to sustainable transport evaluation under incomplete knowledge: electric bikes case study. Sustainability **11**(12), 3314 (2019). https://doi.org/10.3390/su11123314
24. Karczmarczyk, A., Wątróbski, J., Jankowski, J., Ziemba, E.: Comparative study of ICT and SIS measurement in polish households using a MCDA-based approach. Procedia Comput. Sci. **159**, 2616–2628 (2019). https://doi.org/10.1016/j.procs.2019.09.254
25. Karczmarczyk, A., Jankowski, J., Wątróbski, J.: Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PLoS ONE **13**(12), e0209372 (2018). https://doi.org/10.1371/journal.pone.0209372
26. Karczmarczyk, A., Jankowski, J., Watrobski, J.: Parametrization of spreading processes within complex networks with the use of knowledge acquired from network samples. Procedia Comput. Sci. **159**, 2279–2293 (2019). https://doi.org/10.1016/j.procs.2019.09.403
27. Jankowski, J., Zioło, M., Karczmarczyk, A., Wątróbski, J.: Towards sustainability in viral marketing with user engaging supporting campaigns. Sustainability **10**(1), 15 (2018). https://doi.org/10.3390/su10010015
28. Karczmarczyk, A., Jankowsk, J., Wątróbski, J.: Multi-criteria approach to viral marketing campaign planning in social networks, based on real networks, network samples and synthetic networks. In: 2019 Federated Conference on Computer Science and Information Systems (FedCSIS), pp. 663–673. IEEE (2019). https://doi.org/10.15439/2019F199
29. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 199–208. Association for Computing Machinery, New York (2009). https://doi.org/10.1145/1557019.1557047
30. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: 2010 IEEE International Conference on Data Mining, pp. 88–97, December 2010. https://doi.org/10.1109/ICDM.2010.118

31. Marcinkiewicz, K., Stegmaier, M.: The parliamentary election in Poland, october 2015. Elect. Stud. **41**, 221–224 (2016). https://doi.org/10.1016/j.electstud.2016.01.004
32. Enli, G.: Twitter as arena for the authentic outsider: exploring the social media campaigns of trump and clinton in the 2016 US presidential election. Eur. J. Commun. **32**(1), 50–61 (2017). https://doi.org/10.1177/0267323116682802
33. Salehi, M., Sharma, R., Marzolla, M., Magnani, M., Siyari, P., Montesi, D.: Spreading processes in multilayer networks. IEEE Trans. Netw. Sci. Eng. **2**(2), 65–83 (2015). https://doi.org/10.1109/TNSE.2015.2425961
34. Kandhway, K., Kuri, J.: How to run a campaign: optimal control of SIS and SIR information epidemics. Appl. Math. Comput. **231**, 79–92 (2014). https://doi.org/10.1016/j.amc.2013.12.164. http://www.sciencedirect.com/science/article/pii/S0096300314000022
35. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146. ACM (2003). https://doi.org/10.1145/956750.956769
36. Wang, C., Chen, W., Wang, Y.: Scalable influence maximization for independent cascade model in large-scale social networks. Data Min. Knowl. Disc. **25**(3), 545–576 (2012). https://doi.org/10.1007/s10618-012-0262-1
37. Kiss, C., Bichler, M.: Identification of influencers — measuring influence in customer networks. Decis. Support Syst. **46**(1), 233–253 (2008). https://doi.org/10.1016/j.dss.2008.06.007
38. Seeman, L., Singer, Y.: Adaptive seeding in social networks. In: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, pp. 459–468. IEEE (2013). https://doi.org/10.1109/FOCS.2013.56
39. Kitsak, M., et al.: Identification of influential spreaders in complex networks. Nat. Phys. **6**(11), 888 (2010). https://doi.org/10.1038/nphys1746
40. Zhang, J.X., Chen, D.B., Dong, Q., Zhao, Z.D.: Identifying a set of influential spreaders in complex networks. Sci. Rep. **6**, 27823 (2016). https://doi.org/10.1038/srep27823
41. Lin, J.H., Guo, Q., Dong, W.Z., Tang, L.Y., Liu, J.G.: Identifying the node spreading influence with largest k-core values. Phys. Lett. A **378**(45), 3279–3284 (2014). https://doi.org/10.1016/j.physleta.2014.09.054
42. Ho, J.Y., Dempsey, M.: Viral marketing: motivations to forward online content. J. Bus. Res. **63**(9), 1000–1006 (2010). https://doi.org/10.1016/j.jbusres.2008.08.010
43. Jankowski, J., Bródka, P., Kazienko, P., Szymanski, B.K., Michalski, R., Kajdanowicz, T.: Balancing speed and coverage by sequential seeding in complex networks. Sci. Rep. **7**(1), 891 (2017). https://doi.org/10.1038/s41598-017-00937-8
44. Wątróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., Zioło, M.: Generalised framework for multi-criteria method selection. Omega **86**, 107–124 (2019). https://doi.org/10.1016/j.omega.2018.07.004
45. Wątróbski, J., Jankowski, J., Ziemba, P., Karczmarczyk, A., Zioło, M.: Generalised framework for multi-criteria method selection: rule set database and exemplary decision support system implementation blueprints. Data Brief **22**, 639 (2019). https://doi.org/10.1016/j.dib.2018.12.015
46. Ripeanu, M., Foster, I., Iamnitchi, A.: Mapping the Gnutella network: properties of large-scale peer-to-peer systems and implications for system design. arXiv:cs/0209028, September 2002

A6.

Karczmarczyk, A., Bortko, K., Bartków, P., Pazura, P., Jankowski, J. (2018, August). Influencing information spreading processes in complex networks with probability spraying. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1038-1046). IEEE.

Influencing Information Spreading Processes in Complex Networks with Probability Spraying

Artur Karczmarczyk[1], Kamil Bortko[1], Piotr Bartków[1], Patryk Pazura[1] and Jarosław Jankowski[1]

***Abstract*— Research related to information diffusion within complex networks tends to focus on the effective ways to maximize its reach and dynamics. Most of the strategies are based on seeding nodes according to their potential role for social influence. The presented study shows how the seeding can be supported by changes in the target users' motivation to spread the content, thus modifying the propagation probabilities. The allocation of propagation probabilities to nodes takes the form of a spraying process following a given probability distribution, projected from the nodes' rankings. The results showed how different spraying strategies affect the results when compared to the commonly used uniform distribution. Apart from the performance analysis, the empirical study shows to which extent the seeding of nodes with high centrality measures can be compensated by seeding the nodes which are ranked lower, but are having higher motivation and propagation probabilities.**

I. Introduction

Online social networking has become an increasingly important and powerful marketing tool that is used to spread information about ideas, opinions and new product information. Much research in the field relates to diffusion processes for modeling information [1], strategies for the initial seed selection [2], [3], social influence mechanisms [4] and factors that affect their dynamics [5], [6]. The focus is placed mainly on the increase of coverage in the network, based on the number of nodes that are being activated [3]. The research that is related to the diffusion of marketing content and viral marketing that occurs in complex networks takes into consideration the factors that lead to campaigns that are successful [7], [8], the initial seed sets that are selected for the initialization of the campaign [2], as well as epidemic extensions and models usage to model diffusion processes [1]. Using measures of centrality, such as degree, to select initial nodes will result in an underrepresentation of some nodes and an over-representation of nodes with high degrees [9]. As a result, intensive viral marketing is usually based on seeding a large volume of content to the nodes that are most influential, with negative effects observed affecting campaign performance [10]. More sustainable solutions to online marketing are observed in a form of mixture seeding [11]. The majority of the research in this field has been concentrated around propagation models and the improvement of seeding strategies in order to encourage increased coverage. The proponents of this type of research typically assume the seed set selection and immediate initialization of the process, without any additional support. In real campaigns, however, the marketers use techniques of increasing the interest in the content, such as incentives and direct communication, to motivate the consumers to spread the information more effectively.

Additional action can take a form of supporting seeding [12] or initialization of supporting campaigns [13]. Additional actions are affecting the information propagation probabilities. The authors' contribution in this paper is to provide an approach based on assumption that apart from selection of initial nodes in the seeding process, supporting actions based on direct changes of propagation probabilities can be performed. The presented approach is based on spraying propagation probabilities within the network instead of using the same averaged value for all nodes like in most of the prior studies. Mechanics of the spraying process has analogy to messages spraying for routing in networks [14], [15].

The performed experimental research examines how the distribution of propagation probabilities is affecting the final coverage. In the experimental research, both synthetic and real networks as were used. The results showed the relation between coverage from various distributions and the way to achieve high coverage without seeding nodes with high centrality measures.

The remainder of this paper provides a literature review in Section II and presents a conceptual framework as well as the assumptions for proposed approach in Section III. This is followed by the empirical results presented in the experimental Sections IV and V and then the conclusions in Section VI.

II. Literature Review

Social platforms have become a key media source for distributing information through viral marketing, the spreading of rumors as well as greatly influencing important political and social changes. Social data analysis and campaigns targeting users have piqued the interest of marketers as the number of social networking platform users has increased [16]. With its interdisciplinary approach, research that has been done in this field attracts sociologists, physicists, computer scientists and marketers with a wide range of approaches and research aims [2], [1], [17].

Companies that implement techniques that use the "word of mouth" demonstrated that well-designed campaigns can deliver better results than those using traditional methods of marketing [18], [19]. Similarly to the case of the research on the spreading of infectious viral diseases, also in case of these social networks the focus is primarily on the application

[1]Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland jjankowski@wi.zut.edu.pl

of epidemic models to predict and understand the behaviors found in them [1]. There are other modeling techniques used, such as proposed by Kempe et al. independent cascades model (IC), which is dedicated particularly to social networks, and the linear threshold (LT) model, which derives from the innovation diffusion field [3].

The processes involved in the distribution of information are affected by a number of factors, including social relations and network structures [5]. For the final coverage, initial nodes selection involved in the flow of information is crucial. The literature includes a wide discussion of the optimal seed set selection, the research problem put forth by Kempe [3], from perspectives of marketing, sociology, physics and combinatorial optimization [2].

In order to deliver a higher coverage in the network, numerous methods are employed to initiate a cascade of information [20]. The solutions that have been proposed are based on a variety of techniques, such as computationally expensive greedy selection [3], including its extensions [21], heuristics, as well as the selection of nodes that have particular characteristic such as eigenvector or a high degree [22]. Based on measures of centrality, seeding has shown high performance when comparing a variety of seeding strategies [2]. Recent research, however, has moved away from static networks in favor of new approaches based on multilayer networks [23], [24] and temporal networks with more focus being placed on time factors and how the network changes affect the process of information distribution [25]. Changes that have occurred in network and process parameters have been taken into account and new approaches that are more adaptive and have continually evolving strategies have been presented [26], [27].

Other approaches emerged in effort to better use the processes of natural diffusion, use sequential seeding [28], avoid nodes from within the same communities with intra connections that are close by using target communities [29], use dynamic rankings with sequential seeding [30] and use mechanisms for voting that have lower weights once activated nodes have been detected [31]. Central nodes in the networks were detected using a k-shell based approach in other studies [32].

Solutions that are exclusively based on seeding are simplified representations of real-world marketing campaigns, in which, apart from seeding, techniques for increasing the motivation of customers are used. The motivation of the target customers can be increased with incentives and the type of the seeded content [10].

More sustainable solutions to online marketing are observed in a form of mixture seeding [11] or initialization of supporting campaigns [13]. Additional seeds can be used to influence campaign dynamic in a form of supporting seeding [12]. While earlier works focused on additional seeds, the presented study verifies the ability of supporting information spreading processes by increasing the motivation represented by propagation probabilities with strategies based on spraying of the propagation probabilities within network according to their rankings. Increasing the potential of nodes with

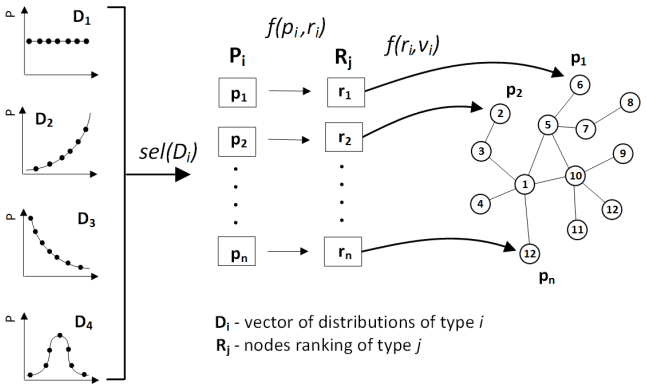


Fig. 1. Mapping PP distribution vector D_i on rank R_j

lower degrees can improve the overall results of a campaign. Despite the numerous connections in the network, content is not always passed along by hubs. For example, in case of higher degree nodes that are engaged in the campaign, weak ties with large number of nodes can result in a low influence on the recipients [10]. A small portion of the many connections found in a network are formed on the basis of strong social relationships [20]. It provides an interesting research gap, to verify the results with increased probability of propagation for nodes with medium or low centrality measures, instead of targeting nodes with higher network positions. Therefore, the authors' contribution in this paper is to provide a approach in which the seeding efforts to obtain a high network coverage are complemented by influencing the propagation probabilities assigned to the network nodes. In terms of practical applications it is modeling the process of motivating customers with potentially lower centrality measures but higher willingness to propagate a content.

III. Conceptual Framework

A. General Assumptions

In the proposed approach, we assume that seeding actions are supported by increasing the motivation of selected nodes within a network. Several strategies focused on increasing the activity of nodes with high centrality measures can be considered. However, such consumers can be demanding and difficult to reach. Therefore, another possibility can be to increase the activity of users with medium centrality measures or even with low measures. Increasing the motivations of users with low measures can be cost effective, with lower incentives assigned than to the hubs. Assume that probability of spreading content is directly related to the motivation of user. Fig.1 shows the process of mapping vector of probability distributions on ranking of nodes and spraying it within the network. From the assumed set of i distributions D we select distribution D_i. A vector $P_i[p_1, p_2, p_n]$ with n elements equal to the number of nodes within a network is created. The vector P_i includes the distribution of probabilities and each element represents the probability which should be assigned to the corresponding rank position. For seeding purposes, within a network with n nodes we create rank of nodes

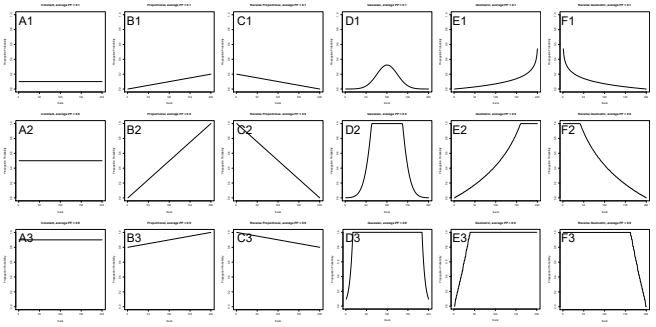


Fig. 2. Example results of the P_i vector generation algorithms for **A** uniform, **B** proportional, **C** reversed proportional, **D** Gaussian, **E** geometric and **F** reversed geometric distributions for average PP equal to **1** – 0.1, **2** – 0.5 and **3** – 0.9.

represented by a vector $R_j[r_1, r_2, r_n]$ with nodes ordered by their centrality measure of j type. Function $f(p_i, r_i)$ is assigning probability p_i to the rank element r_i. Function $f(r_i, v_i)$ is mapping probabilities assigned to nodes in rank on vertices within the network.

As part of the proposed framework, a set of algorithms for obtaining the P_i vectors was created, taking the number of nodes n and the expected average propagation probability P_{avg} as input and returning the P_i vector as output. For example, the P_i vector for the proportional distribution can be obtained and sprayed with the following algorithm:

Input: n, avgPP in
Output: Pi out
Initialisation:
if (avgPP <= 0.5) **then**
  maxV <- avgPP * 2;
  minV <- 0;
else
  maxV <- 1;
  minV <- avgPP - (maxV - avgPP);
end if
Pi <- seq(minV, maxV, (maxV-minV)/(n-1));
return Pi
for $i = 1$ to n **do**
  SprayProbabilityToNode(Pi,Nodei);
end for

In case of the reversed proportional distribution, the P_i vector is generated similarly, yet, in the penultimate step of the algorithm, the generated sequence is reversed. A sample set of P_i vectors obtained from the implemented algorithms for uniform, proportional, reversed proportional, Gaussian, geometric and reversed geometric distributions for the average propagation probability equal to 0.1, 0.5 and 0.9 are presented on Fig. 2.

B. Illustrative Example

Effects of different probability spraying strategies are illustrated on Fig. 3. The presented example is based on real simulation within an actual network consisting of 16 nodes [33]. The seeding fraction parameter set to 25% resulted in four nodes being selected as seeds. The order of the network's nodes corresponds to their degrees. Four probability spraying strategies were selected according to the uniform, Gaussian, proportional and reversed proportional distribution. Tab III-A contains PP values for each node in the network, according to the selected distributions. On Fig. 3, the seed nodes are marked in red. The shade of blue indicates the size of propagation probability relative to a particular distribution. Fig. 3A shows the process based on uniform propagation probability, in which to each node the same PP value of 0.2 is assigned. The process finishes with 68.75% coverage with 11 infected nodes in 4 steps. Fig. 3B illustrates the process based on the Gaussian distribution of probability among nodes. A final coverage of 87.5% was achieved in 6 steps with 14 infected nodes. The proportional coverage illustrated in Fig. 3C delivered 37.5% (6) of infected nodes in 3 steps. The reversed proportional distribution presented in Fig. 3D achieved 87.5% (14) infected nodes in three steps. The example showed how propagation spraying can affect the coverage of information propagation processes. In the next stage of the research, simulations within synthetic and real networks are performed to evaluate the impact of spraying distributions on final coverage.

C. Plan of Experiments

The plan of experiments assumes the usage of agent-based simulations on synthetic networks (SN) and real networks (RN). Main simulation parameters include the propagation probability (PP), fraction of nodes used as seeds (SF), seed selection strategies (SS), distribution of probabilities within network for propagation spraying (PS). The complete experimental space with variants for each parameter is presented in Tab II.

The simulation parameters create an experimental space N x PP x SF x SS x PS resulting into 50400 configurations for synthetic networks and 12000 configurations for real networks. An agent-based model was implemented, with agents connected according to the networks specifications. Comparisons were performed using the same network (N) with the same parameters including propagation probability (PP), seeding fraction (SF) and seed selection strategy (SS). The independent cascades model (IC) was used for simulations. With propagation probability $PP(a, b)$, node a activates node b in the step $t + 1$ under the condition that node a was activated at time t. We assume that the average probability is equal to the assumed probability PP.

IV. Empirical Study Based on Synthetic Networks

A. Characteristics of Used Networks

The experiments where performed with the use of synthetic networks generated with the use of theoretical models: Barabasi-Albert model [34] (BA), Watts-Strogatz model [35] (WS), and Erdos-Renyi model [36] (ER). For each model, networks with 200 nodes were used. Five variants of BA networks were generated with given average out-degree of the vertices with the values of 1, 2, 3, 4 and 5 for networks denoted as BA1, BA2, BA3, BA4 and BA5 respectively. Five

TABLE I

SEEDING FRACTION = 0.25 / AVERAGE PROPAGATION PROBABILITY = 0.2

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Name	6	11	15	16	1	2	12	13	5	7	8	9	3	10	14	4
Degree	10	9	9	9	8	8	8	8	7	7	7	7	6	5	5	3
Uniform	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000	0.2000
Gaussian	0.0002	0.0016	0.0092	0.0383	0.1195	0.2806	0.4957	0.6589	0.6589	0.4957	0.2806	0.1195	0.0383	0.0092	0.0016	0.0002
Proportional	0.0000	0.0266	0.0533	0.0800	0.1066	0.1333	0.1600	0.1866	0.2133	0.2400	0.2666	0.2933	0.3200	0.3466	0.3733	0.4000
Reversed proportional	0.4000	0.3733	0.3466	0.3200	0.2933	0.2666	0.2400	0.2133	0.1866	0.1600	0.1333	0.1066	0.0800	0.0533	0.0266	0.0000

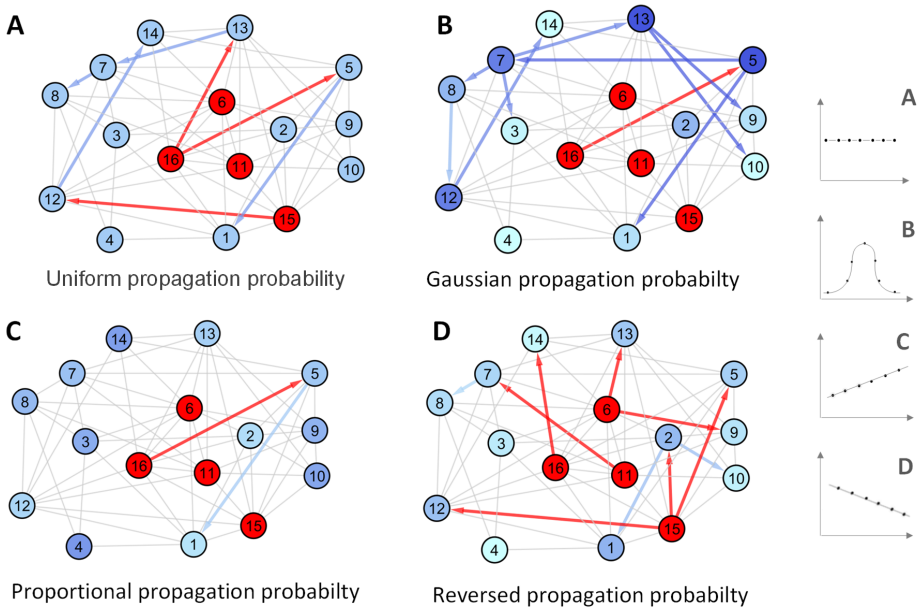


Fig. 3. **A** Example process with uniform propagation probability distribution; **B** Example process with Gaussian propagation probability distribution; **C** Example process with proportional propagation probability distribution; **D** Example process with reversed proportional propagation probability distribution;

ER networks ER1, ER2, ER3, ER4 and ER5 were used with different number of edges equal to 200, 400, 600, 800 and 1000. Fifteen networks following WS model were generated with neighborhood within which the vertices of the lattice equal to 1, 2 and 3. For each value, there were assigned five rewiring probabilities with values 0, 0.25, 0.5, 0.75 and 1, with assigned symbols A, B, C, D and E used for network naming like (WS1A, WS1B etc). Average values of main network parameters for each used network including degree, closeness, PageRank, eigenvector, clustering coefficient and betweenness are presented in Tab III.

B. Overall Results from Simulations

The average coverage from the uniform propagation probability (PP) distribution was observed at the level of 55%. The total coverage was increased by 14 p.p. to 68% when the reverse geometric (RevGeom) distribution was applied with p-value<2.2e-16 for Wilcoxon signed rank test (with Hodges-Lehmann estimator H=-0.136). A smaller increase was observed (64%), when reverse proportional distribution was used (RevProp) (pvalue<2.2e-16, H=-0.093). Decrease in coverage was observed for other spraying strategies. The average coverage was observed at the level of 23% (pvalue<2.2e-16, H=0.407), 25% (pvalue<2.2e-16, H=0.34) and 38% (pvalue<2.2e-16, H=0.158) for Gaussian distribution, Geometric distribution (Geom) and proportional distribution (Prop) respectively. The relation between results for used strategies are presented in Fig. 4A with simulation cases above and below the uniform distribution.

In the next stage, performance factors were computed showing coverage for specific strategy divided by coverage of uniform distribution with values presented in Tab IV. The highest improvement of results was observed for RevGeom with an almost two-fold increase (1.966) when compared to uniform distribution of propagation probability. Performance factor for geometric distribution (Geom) of propagation probabilities was at the level 0.492. Reverse proportional spraying (RevProp) achieved 1.511 of the value observed for uniform distribution. The average factor for proportional spraying (Prop) with assigned higher propagation probabilities to nodes with lower centrality measures achieved 0.670, which represents a substantial decrease when compared to uniform distribution. It is only 44% of the value achieved by the reverse proportional spraying. The lowest factor was obtained for Gaussian distribution with a value of 0.475.

The overall analysis of the simulation cases for all parameters shows that the spreading processes based on geometric distribution of PP delivered better results than uniform distribution in 82.039% of cases. Slightly lower number of cases was observed for reversed proportional approach (75.727%).

TABLE II

SIMULATION PARAMETERS AND THEIR VARIANTS

Symbol	Parameter	Variants	Values
N	Set of synthetic and real networks	26	21 synthetic networks and 5 real networks
PP	Propagation probability	10	0.01 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
SF	Initial seeding fraction	10	1% 2% 3% 4% 5% 6% 7% 8% 9% 10%
SS	Seed selection strategy	4	D - degree, CL - closeness, EV - eigenvector, BT - betweenness
PS	Propagation spraying	6	Uniform, Geometric, Proportional, Gaussian, Reverse Geometric, Reverse Proportional

TABLE III

SPECIFICATION OF USED NETWORKS WITH AVERAGED NETWORK METRICS

Network	Param	Nodes	Edges	Degree	Closeness	PageRank	Eigenvector	Clustering Coefficient	Betweenness
BA1	1	200	199	1.99	0.16	0.004	0.004	0	528.52
BA2	2	200	397	3.97	0.28	0.004	0.099	0.03	250.99
BA3	3	200	594	5.93	0.34	0.005	0.149	0.05	189.14
BA4	4	200	790	7.90	0.38	0.004	0.154	0.08	163.66
BA5	5	200	985	7.90	0.40	0.005	0.213	0.09	149.84
ER1	200	173	200	2.31	0.03	0.005	0.145	0	366.66
ER2	400	196	400	4.08	0.25	0.005	0.196	0.02	284.24
ER3	600	199	600	6.03	0.32	0.005	0.416	0.02	210.58
ER4	800	200	800	8.00	0.36	0.004	0.443	0.04	176.10
ER5	1000	200	1000	10.00	0.39	0.005	0.564	0.05	154.04
WS1C	1 & 0.50	177	200	2.25	0.04	0.005	0.121	0.008	516.35
WS1B	1 & 0.25	190	200	2.10	0.03	0.005	0.100	0	704.43
WS1D	1 & 0.75	173	200	2.31	0.04	0.005	0.186	0.007	359.97
WS1A	1 & 0.00	200	200	2.0	0.01	0.005	0.999	0	4900.5
WS1E	1 & 1.00	172	200	2.32	0.05	0.005	0.145	0.007	455.36
WS2C	2 & 0.50	199	400	4.02	0.25	0.005	0.357	0.02	294.51
WS2B	2 & 0.25	200	400	3.99	0.23	0.004	0.323	0.07	321.75
WS2D	2 & 0.75	199	400	4.02	0.25	0.005	0.293	0.009	289.93
WS2A	2 & 0.00	899	400	4.0	0.03	0.004	0.999	0.5	2425.5
WS2E	2 & 1.00	200	400	4.08	0.25	0.005	0.207	0.01	283.08
WS3C	3 & 0.50	200	600	6.00	0.31	0.004	0.413	0.02	213.11
WS3B	3 & 0.25	200	600	5.99	0.30	0.005	0.464	0.10	228.55
WS3D	3 & 0.75	199	600	6.03	0.32	0.005	0.329	0.03	209.62
WS3A	3 & 0.00	200	600	6.00	0.05	0.005	0.999	0.6	1600.50
WS3D	3 & 1.00	200	600	6.00	0.31	0.004	0.402	0.03	214.55
R1	Real network [37]	1899	20296	21.37	0.11	0.0005	0.079	0.05	1938.04
R2	Real network [38]	7610	15751	4.13	0.0004	0.0001	0.003	0.32	13478.93
R3	Real network [39]	1224	19090	31.19	0.21	0.0008	0.079	0.22	1059.02
R4	Real network [40]	1461	2742	3.75	0.0007	0.0006	0.013	0.69	251.35
R5	Real network [41]	1133	5451	9.62	0.28	0.0008	0.077	0.16	1475.01

Geometric, proportional and Gaussian spraying delivered lower results in more than 99% of cases. The number of simulation cases with increased values is illustrated in Fig. 4B for all used strategies, with values above 1.0 for cases with coverage better than for the uniform propagation. As it is visible in Fig. 4C, the seeding fraction did not change relation between the distribution of the propagation probabilities.

C. Results for Used Network Types

Further analysis focused on the results observed for various network types. Fig. 4D2 shows the results obtained for BA networks generated with different parameters. In general, the averaged values of factor computed for proportional distributions from BA networks, equal to 0.600, is almost two times lower than the average factor for reverse proportional distribution with the value of 1.386. An analysis of the average factor for geometric distribution 0.445 and factor for the reverse geometric distribution with the value of 1.744 shows a four-fold difference. The lowest average value for all BA networks is achieved with Gaussian distribution factor with the value of 0.421.

The results show that the highest performance with factor 2.649 for reverse geometric was observed for BA1, while the lowest value 1.347, lower by 49%, was observed for BA5. Within BA networks, a geometric factor growing by 41% is observed with the value from 0.336 for BA1 up to the highest value 0.570 for BA5. The averaged factors for Gaussian distributions in four cases are stabilized at the level of 0.44, apart from the BA1 network, with the

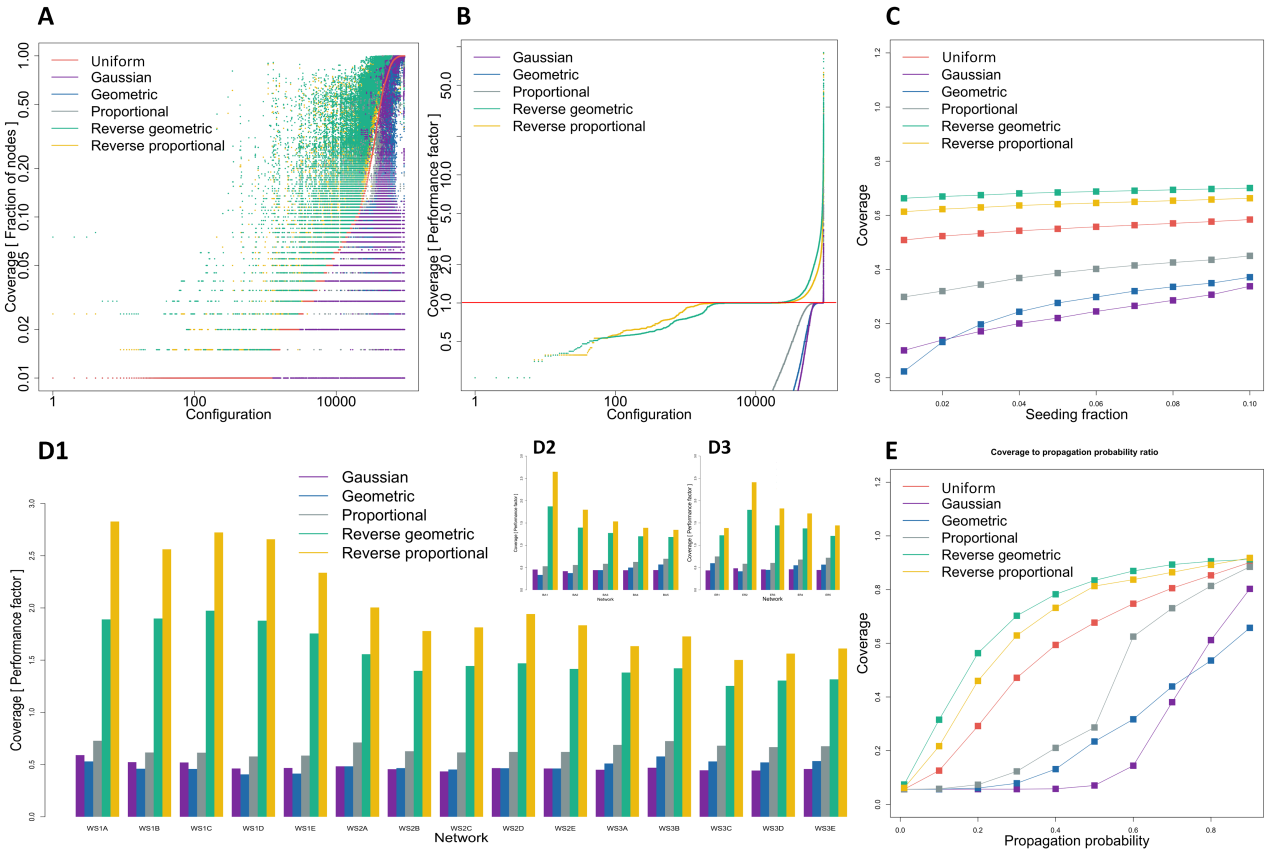


Fig. 4. **A** Results from synthetic networks with distribution of simulation cases above and below results from uniform probability distribution; **B** Number of cases for each distribution with results better than from uniform probability distribution; **C** Relation between seeding traction and coverage for all used probability spraying strategies; **D1** Coverage for all spraying strategies within WS networks; **D2** Coverage for all spraying strategies within BA networks; **D3** Coverage for all spraying strategies within ER networks; **E** Results for spraying strategies for each propagation probability;

average value at the level of 0.352 with 20% difference, when compared to other networks. For the factor computed for reverse proportional growth is observed for BA1 when compared with other networks, with 25-35% increase when compared to all other BA networks. Different relation is observed for the proportional distribution with value for BA5 higher by 11% to 24%.

The results for ER networks are presented in Fig. 4D3. The averaged factor for the proportional distribution, equal to 0.668, is more than 50% lower than the average factor for the reverse proportional distribution, with the value of 1.410. The analysis of average values of performance factor for geometric distribution, with the value of 0.516, and average factor for reverse geometric, 1.756, shows nearly a three-fold difference. The lowest average value for all networks, similarly like for the BA networks, was obtained for the average factor based on the Gaussian distribution with the value equal to 0.458.

The values difference for the factor computed for the proportional distribution between network with the highest value 0.75 for network ER5 and the lowest for ER1 equal to 0.586 is equal to 20%. Different relation was observed for the reversed proportional distribution. The difference between ER1 network with value 1.794 has differences from 20% to 33%, when compared with other networks. The averaged values for the Gaussian distribution are localized in the range from 0.44 to 0.48. A small difference (10%) is observed when ER5 is compared with ER1. For the ER networks, a growing geometric distribution factor is observed by 31% for ER1, with average value of 0.417 when compared to ER5, with the highest value equal to 0.598. The results from the ER networks show the highest value 2.412 for the reverse geometric factor for ER1, while it equaled 1.386 for ER5.

The results from the WS networks are presented in Fig. 4D1. The average value of factor of proportional distribution equal to 0.711 is nearly two times lower than the averaged factor for reverse proportional, with the value 1.410. An average factor for the reverse geometric distribution with the value 2.123 is four time higher than 0.509 observed for the geometric one. The lowest average factor was observed for the Gaussian distribution, with the value 0.509 higher than for BA and ER network.

The difference of factor for proportional distribution from 0.586 for WS1E up to 1.107 for WS1A with 47% difference is observed. A different relation is observed for the reverse proportional distribution, with the highest value for WS1C 1.973 when compared to 1.254 for WS3C and a difference of nearly 37%.

The averaged values for the Gaussian distribution take values from 0.44 to 0.57. A 50% difference to other networks is observed, when compared to WS1A with value 0.858. In the WS networks, a stabilized factor for geometric distribution is observed. All networks apart from WS1A, with the factor value of 0.799, have factors localized in the range from 0.413 to 0.576. The average factor for reverse geometric distribution is dropping by 45% for WS3C to 1.501 when compared to WS1A with the highest value of 3.365.

D. Results for Used Simulation Parameters

Another part of the analysis shows how the propagation probability is affecting the final results. Analyzing the average values for individual probabilities, we can observe significant relationships. The values of factors for proportional distributions can be divided into three groups relative to intervals. The first group is only one probability of 0.1, for which the value is 0.576. The second of them is in the range from 0.330 to 0.399. It includes the probabilities from 0.2 to 0.5. The third one is in the range from 0.813 to 1. It consists of the probabilities 0.01 and from 0.6 to 0.9. Thus, a significant difference in the value span, equal to 0.672, can be observed. This difference perfectly fits into the average value of all probabilities with the value of 0.674. For the average factor of reverse proportional values a nearly two-fold difference can be seen between the lowest value with the probability of 0.9 and the highest value of 0.2. The average of all measures for all probabilities is 1.499, which is twice as high as for the proportional factor. Analyzing the factor for Gaussian distribution, we can see up to six times the difference in the probability of 0.01, where it is 1, relative to the probability of 0.5, with a value of 0.16. The average value of factor for Gaussian distribution for all probabilities is equal to 0.475. The mean values for all probabilities for the geometric distribution, with the value of 0.499, against reverse geometric distribution, with the value of 1.934 are nearly four times lower. The highest factor for geometric distribution is observed for the probability of 0.1. The lowest value (0.270) is observed for the probability of 0.4. For the reverse geometric probability, the highest value is 3.083 for the probability of 0.1. On the other hand, the lowest value 1.081 is observed for the probability of 0.9.

The results presented in Fig. 4E show that the coverage at the level of 30% for the uniform propagation probability can be obtained with the propagation probability of 0.2. Better results can be obtained for the reverse proportional distribution, with an average probability of 0.1. The same result can be obtained for proportional distribution with average propagation probability at the level of 0.5. The geometric distribution requires a higher probability to obtain such coverage and average probability at the level of 0.6 was required. The highest average propagation probability to obtain 30% of coverage was required for the Gaussian distribution of probabilities with the average value at the level of 0.65. With the average propagation probability at the level of 0.5, an increase of coverage is observed for the proportional distribution from 29% to 61%, which is the highest increase observed, equal to 32%. With the same probability, uniform propagation has a gain of 75% and 68% with the 0.5 probability.

V. Results from Real Networks

The analysis of results for the presented real networks indicates the highest efficiency of reverse geometric distribution with a performance factor of 1.623. It is more than four times more than geometric distribution with the factor 0.473. For the proportional distribution, the average value of 0.664 was obtained, while for the reverse proportional the average value was 1.250. The lowest value of the performance factor was obtained for the Gaussian probability distribution at the level of 0.437. Fig. 5A shows distribution of simulation cases for all used simulation configurations. Fig. 5B shows results for each spraying strategy for all networks. The largest coverage was obtained for network R1 (Fig. 5C2). The reverse geometric and reverse proportional distributions of 50% coverage are already achieved with the probability of 0.05 and the distribution of proportional at 0.5 probability. Geometric and Gaussian distribution with a similar 0.65 probability. These are by far the best results compared to the four other networks. For R2 network (Fig. 5C3), when analyzing the increase in probability and uniformity, we see that the reverse geometric and reverse proportional values behave similarly to other networks. In turn, coverage for proportional networks reach the value of 50% only when the probability is close to 0.65, Gaussian with the probability close to 0.75. On the other hand, the geometric value of 50% achieves close to 0.9 with a high legal probability. Analyzing the increase in the probability and respect of the uniform (Fig. 5C4), we see that the reverse geometric and reverse proportional values also behave according to the general scheme. On the other hand, the coverage for proportional value of 50% is achieved only with a probability of close to 0.3 Gaussian with a probability of close to 0.65. On the other hand, the geometric distribution achieves 50% near 0.5 probability, which is a better result compared to the R2 network. In the R4 network (Fig. 5C5), analyzing the increase in the probability and the uniform results, we see that no distribution behaves according to the general scheme. Networks do not reach 50% coverage in a probability of 0.9. The reverse proportional and reverse geometric and proportional distributions reach a maximum of nearly 30% coverage only with a probability of 0.9. The Gaussian and geometric distributions reach a probability with nearly 20% coverage. In the R5 network (Fig. 5C6) coverage for the proportional distribution is 50% only when the probability is close to 0.4, Gaussian with a probability of close to 0.6. On the other hand, the geometric value of 50% is close to 0.55 probability, which is also a better result compared to the R2 network.

VI. Conclusions

The presented research was focused on supporting the spreading processes within networks with increasing propagation probabilities according to the proposed strategies

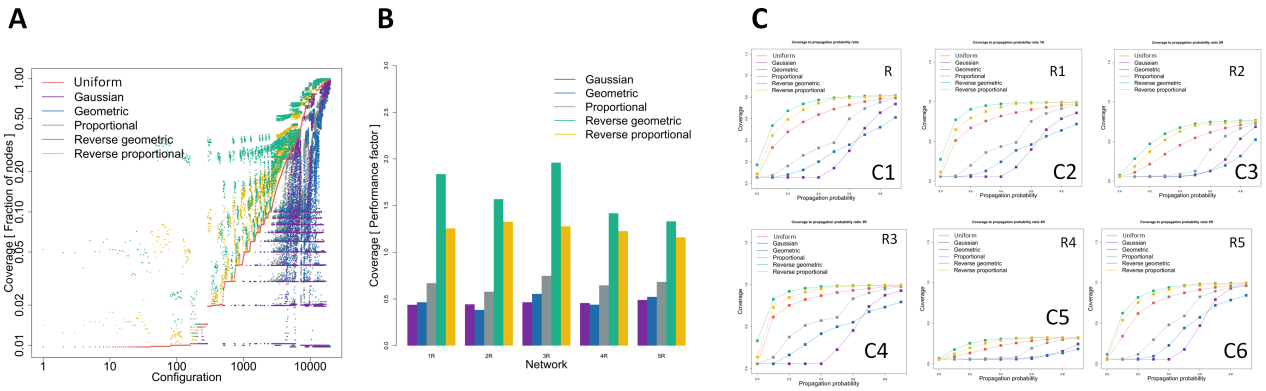


Fig. 5. **A** Results from real networks with distribution of simulation cases above and below results from uniform probability distribution; **B** Coverage for all spraying strategies within real networks; **C1** Aggregated results from all networks for all propagation probabilities; **C2** Results from network R1 for all propagation probabilities; **C3** Results from network R2 for all propagation probabilities; **C4** Results from network R3 for all propagation probabilities; **C5** Results from network R4 for all propagation probabilities; **C6** Results from network R5 for all propagation probabilities;

TABLE IV

PERFORMANCE FACTORS FOR USED SPRAYING STRATEGIES

Network	Prop	RevProp	Gaussian	Geom	RevGeom
BA1	0.531	1.872	0.352	0.336	2.649
BA2	0.559	1.397	0.420	0.377	1.798
BA3	0.586	1.275	0.444	0.444	1.536
BA4	0.628	1.203	0.442	0.501	1.392
BA5	0.698	1.186	0.445	0.570	1.347
ER5	0.750	1.223	0.437	0.598	1.386
ER1	0.586	1.794	0.484	0.417	2.412
ER2	0.604	1.445	0.458	0.448	1.825
ER3	0.681	1.377	0.462	0.552	1.713
ER4	0.721	1.209	0.446	0.567	1.445
WS1A	1.107	1.810	0.859	0.799	3.365
WS1B	0.615	1.898	0.523	0.460	2.562
WS1C	0.614	1.973	0.520	0.458	2.723
WS1D	0.578	1.877	0.462	0.405	2.659
WS1E	0.586	1.755	0.467	0.413	2.337
WS2A	0.983	1.916	0.578	0.547	2.486
WS2B	0.628	1.397	0.456	0.466	1.778
WS2C	0.617	1.444	0.434	0.453	1.814
WS2D	0.621	1.470	0.466	0.464	1.942
WS2E	0.621	1.415	0.462	0.462	1.833
WS3A	0.685	1.546	0.442	0.413	1.742
WS3B	0.724	1.422	0.469	0.577	1.727
WS3C	0.680	1.254	0.445	0.530	1.501
WS3D	0.666	1.304	0.443	0.521	1.562
WS3E	0.676	1.316	0.457	0.533	1.612

TABLE V

PERFORMANCE FACTORS FOR USED SIMULATION PARAMETERS

Parameter	Prop	RevProp	Gaussian	Geom	RevGeom
PP0.01	1.000	1.103	1.000	1.000	1.396
PP0.10	0.576	1.811	0.570	0.570	3.083
PP0.20	0.381	2.050	0.352	0.360	2.991
PP0.30	0.330	1.758	0.246	0.273	2.401
PP0.40	0.362	1.683	0.189	0.270	2.134
PP0.50	0.399	1.729	0.167	0.340	1.931
PP0.60	0.813	1.413	0.216	0.395	1.661
PP0.70	0.904	1.228	0.443	0.498	1.412
PP0.80	0.986	1.154	0.678	0.583	1.249
PP0.90	0.991	1.061	0.885	0.698	1.081
Betweenness	0.680	1.507	0.477	0.498	1.957
Closeness	0.680	1.507	0.477	0.498	1.957
Degree	0.672	1.464	0.464	0.485	1.871
Eigenvector	0.664	1.519	0.481	0.514	1.950
SF0.01	0.585	2.313	0.320	0.240	3.655
SF0.02	0.599	1.746	0.366	0.356	2.456
SF0.03	0.626	1.565	0.405	0.428	2.068
SF0.04	0.653	1.465	0.438	0.480	1.855
SF0.05	0.677	1.408	0.464	0.519	1.732
SF0.06	0.692	1.360	0.495	0.545	1.636
SF0.07	0.708	1.329	0.521	0.573	1.573
SF0.08	0.719	1.301	0.547	0.592	1.517
SF0.09	0.730	1.278	0.570	0.608	1.471
SF0.10	0.746	1.245	0.604	0.632	1.414

based on the propagation probability distribution. Two of the presented approaches were based on a reversed geometric and a reversed proportional distribution, with higher increase of probabilities for the high ranked nodes. They resulted in the highest increase of coverage within the network when compared to uniform distribution. Usage of geometric and proportional distribution of probabilities resulted in a decrease of coverage when compared to the same propagation probability, but the overall costs of the campaign can be lower. The Gaussian approach delivered the lowest coverage while it targeted mainly the nodes with medium ranks.

The highest increase of results with the use of reversed distributions was observed for the networks with lower average degree like R4, R5, BA1, ER1 or WS1A. WS2A, WS3A. Proportional approaches were working best for networks with the highest average degrees like R1, R3, BA5, ER5 and WS1E, WS2E, WS3E. While strategies with reversed geometric and proportional distributions can be considered as most effective in terms of network coverage, the cost associated with them can be relatively high, because of higher incentives required by the high ranked nodes. An over-representation of hubs can be increased for such strategies.

The experiment showed that the coverage can be increased even though low ranked nodes are targeted when the average propagation probability is increased.

The presented study opens several research questions for the future work. The proposed approach can be used for both increasing and decreasing the propagation probabilities. The probabilities can be assigned in a more dynamic manner, during the process, while more knowledge about the process is being acquired. Another direction can be based on the usage of a wider range of seeding methods and usage of adaptive approaches.

ACKNOWLEDGEMENTS

This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562.

REFERENCES

[1] K. Kandhway and J. Kuri, "How to run a campaign: Optimal control of sis and sir information epidemics," *Applied Mathematics and Computation*, vol. 231, pp. 79–92, 2014.

[2] O. Hinz, B. Skiera, C. Barrot, and J. U. Becker, "Seeding strategies for viral marketing: An empirical comparison," *Journal of Marketing*, vol. 75, no. 6, pp. 55–71, 2011.

[3] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[4] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1029–1038.

[5] M. Bampo, M. T. Ewing, D. R. Mather, D. Stewart, and M. Wallace, "The effects of the social structure of digital networks on viral marketing performance," *Information systems research*, vol. 19, no. 3, pp. 273–290, 2008.

[6] S. Bhagat, A. Goyal, and L. V. Lakshmanan, "Maximizing product adoption in social networks," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 603–612.

[7] J. Berger and K. L. Milkman, "What makes online content viral?" *Journal of marketing research*, vol. 49, no. 2, pp. 192–205, 2012.

[8] J. Y. Ho and M. Dempsey, "Viral marketing: Motivations to forward online content," *Journal of Business research*, vol. 63, no. 9-10, pp. 1000–1006, 2010.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[10] R. Michalski, J. Jankowski, and P. Kazienko, "Negative effects of incentivised viral campaigns for activity in social networks," in *Cloud and Green Computing (CGC), 2012 Second International Conference on*. IEEE, 2012, pp. 391–398.

[11] J. Jankowski, "Mixture seeding for sustainable information spreading in complex networks," in *Asian Conference on Intelligent Information and Database Systems*. Springer, 2017, pp. 191–201.

[12] J. Jankowski and R. Michalski, "Increasing coverage of information spreading in social networks with supporting seeding," in *International Conference on Data Mining and Big Data*. Springer, 2017, pp. 209–218.

[13] J. Jankowski, M. Zioło, A. Karczmarczyk, and J. Wątróbski, "Towards sustainability in viral marketing with user engaging supporting campaigns," *Sustainability*, vol. 10, no. 1, p. 15, 2017.

[14] E. Bulut, Z. Wang, and B. K. Szymanski, "Cost-effective multiperiod spraying for routing in delay-tolerant networks," *IEEE/ACM Transactions on Networking (ToN)*, vol. 18, no. 5, pp. 1530–1543, 2010.

[15] ——, "Time dependent message spraying for routing in intermittently connected networks," in *Global Telecommunications Conference, 2008. IEEE GLOBECOM 2008. IEEE*. IEEE, 2008, pp. 1–6.

[16] G. Bello-Orgaz, J. J. Jung, and D. Camacho, "Social big data: Recent achievements and new challenges," *Information Fusion*, vol. 28, pp. 45–59, 2016.

[17] J. L. Iribarren and E. Moro, "Impact of human activity patterns on the dynamics of information diffusion," *Physical review letters*, vol. 103, no. 3, p. 038702, 2009.

[18] M. Trusov, R. E. Bucklin, and K. Pauwels, "Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site," *Journal of marketing*, vol. 73, no. 5, pp. 90–102, 2009.

[19] A. Dobele, D. Toleman, and M. Beverland, "Controlled infection! spreading the brand message through viral marketing," *Business Horizons*, vol. 48, no. 2, pp. 143–149, 2005.

[20] X. Fan and V. O. Li, "The probabilistic maximum coverage problem in social networks," in *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*. IEEE, 2011, pp. 1–5.

[21] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1039–1048.

[22] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 199–208.

[23] M. De Domenico, C. Granell, M. A. Porter, and A. Arenas, "The physics of spreading processes in multilayer networks," *Nature Physics*, vol. 12, no. 10, p. 901, 2016.

[24] J. Jankowski, R. Michalski, and P. Bródka, "A multilayer network dataset of interaction and influence spreading in a virtual world," *Scientific data*, vol. 4, p. 170144, 2017.

[25] R. Lambiotte, L. Tabourier, and J.-C. Delvenne, "Burstiness and spreading on temporal networks," *The European Physical Journal B*, vol. 86, no. 7, p. 320, 2013.

[26] F. Stonedahl, W. Rand, and U. Wilensky, "Evolving viral marketing strategies," in *Proceedings of the 12th annual conference on Genetic and evolutionary computation*. ACM, 2010, pp. 1195–1202.

[27] L. Seeman and Y. Singer, "Adaptive seeding in social networks," in *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, 2013, pp. 459–468.

[28] J. Jankowski, P. Bródka, P. Kazienko, B. K. Szymanski, R. Michalski, and T. Kajdanowicz, "Balancing speed and coverage by sequential seeding in complex networks," *Scientific reports*, vol. 7, no. 1, p. 891, 2017.

[29] J.-L. He, Y. Fu, and D.-B. Chen, "A novel top-k strategy for influence maximization in complex networks with community structure," *PloS one*, vol. 10, no. 12, p. e0145283, 2015.

[30] J. Jankowski, "Dynamic rankings for seed selection in complex networks: Balancing costs and coverage," *Entropy*, vol. 19, no. 4, p. 170, 2017.

[31] J.-X. Zhang, D.-B. Chen, Q. Dong, and Z.-D. Zhao, "Identifying a set of influential spreaders in complex networks," *Scientific reports*, vol. 6, p. 27823, 2016.

[32] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature physics*, vol. 6, no. 11, p. 888, 2010.

[33] K. E. Read, "Cultures of the central highlands, new guinea," *Southwestern Journal of Anthropology*, pp. 1–43, 1954.

[34] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[35] D. J. Watts and S. H. Strogatz, "Collective dynamics of âĂŸsmall-worldâĂŹnetworks," *nature*, vol. 393, no. 6684, p. 440, 1998.

[36] P. Erdos, "On random graphs," *Publicationes mathematicae*, vol. 6, pp. 290–297, 1959.

[37] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social networks*, vol. 31, no. 2, pp. 155–163, 2009.

[38] M. E. Newman, "The structure of scientific collaboration networks," *Proceedings of the national academy of sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[39] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*. ACM, 2005, pp. 36–43.

[40] M. E. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical review E*, vol. 74, no. 3, p. 036104, 2006.

[41] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical review E*, vol. 68, no. 6, p. 065103, 2003.

A7.

Karczmarczyk, A., Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeting Multi-Attribute Nodes in Complex Networks. Symmetry, 13(4), 731.

symmetry

MDPI

Article

Multi-Criteria Seed Selection for Targeting Multi-Attribute Nodes in Complex Networks

Artur Karczmarczyk [1], Jarosław Jankowski [1] and Jarosław Wątrobski [2,*]

[1] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland; artur.karczmarczyk@zut.edu.pl (A.K.); jaroslaw.jankowski@zut.edu.pl (J.J.)

[2] The Faculty of Economics, Finance and Management of the University of Szczecin, Mickewicza 64, 71-101 Szczecin, Poland

* Correspondence: jaroslaw.watrobski@usz.edu.pl

Abstract: Online environments have evolved from the early-stage technical systems to social platforms with social communication mechanisms resembling the interactions which can be found in the real world. Online marketers are using the close relations between the users of social networks to more easily propagate the marketing contents in their advertising campaigns. Such viral marketing campaigns have proven to provide better results than traditional online marketing, hence the increasing research interest in the topic. While the majority of the up-to-date research focuses on maximizing the global coverage and influence in the complete network, some studies have been conducted in the area of budget-constrained conditions as well as in the area of targeting particular groups of nodes. In this paper, a novel approach to targeting multi-attribute nodes in complex networks is presented, in which an MCDA method with various preference weights for all criteria is used to select the initial seeds to best reach the targeted nodes in the network. The proposed approach shows some symmetric characteristics—while the global coverage in the network is decreased, the coverage amongst the targeted nodes grows.

Keywords: complex networks; social networks; viral marketing; information propagation; MCDA; TOPSIS

Citation: Karczmarczyk, A.; Jankowski, J.; Wątrobski, J. Multi-Criteria Seed Selection for Targeting Multi-Attribute Nodes in Complex Networks. *Symmetry* **2021**, *13*, 731. https://doi.org/10.3390/sym13040731

Academic Editor: José Carlos R. Alcantud

Received: 10 March 2021
Accepted: 16 April 2021
Published: 20 April 2021

1. Introduction

The analysis of social networks has evolved from early-stage sociograms based on small graphs into mainstream multi-billion node social networks with high business potential [1]. Social platforms let their users easily connect to their friends or acquaintances and easily maintain relationships. These close relations between social network users have been widely used by online marketers to improve the engagement of potential consumers to benefit from their services and products [2]. Viral marketing campaigns in social networks have proven to bring better effects in engaging potential consumers than traditional online advertising [3].

This performance of viral marketing resulted in increased research on information propagation in complex networks. While the majority of the research focuses exclusively on increasing the network coverage with information, as the only factor and performance measure, some works aim their attention at a targeted approach [4,5], also with a focus on user preferences [6]. From a different perspective, other approaches avoid repeated messages due to lowered performance causing a habituation effect [7], information overload [8] or the need for delays between messages for multi-product campaigns [9]. Efforts towards targeting specific users have mainly been focused on single attributes or network metrics for the seed selection [10]. The real-life applications of social networks in viral marketing campaigns are often based on selecting multiple attributes such as age, gender and localization of the target group [11].

To better address the aforementioned needs, the authors' main contribution in this paper is to provide an approach in which multi-attribute targeted groups of users can be reached in social networks by providing the initial seeding information to a limited number of selected network users. In the proposed approach, contrary to other studies, the selection of the seeded nodes of the social network is based on multiple, often conflicting, criteria and nodes' attributes. Moreover, by virtue of the MCDA (Multi-Criteria Decision Analysis) foundations of the proposed approach, the importance of each criterion considered in the selection process can be adjusted to meet the marketer's needs. MCDA tools, such as sensitivity analysis [12], also allow us to further study and understand the effect each seeded nodes' attribute has on the planned viral marketing campaign's capacity to reach the targeted group of the network nodes [13]. Some symmetric characteristics of the proposed approach are assumed—whilst the global coverage in the network can decrease, the proposed approach strives to maximize coverage amongst the targeted nodes.

The paper is comprised of five main sections. After this introduction, the state-of-the-art literature review is presented in Section 2. It is followed by the methodology presentation in Section 3 and the empirical study results in Section 4. Eventually, the paper is concluded in Section 5.

2. Literature Review

The early stage research in the area of information spreading assumed that all nodes within the network have the same interest in the product or the propagated content. The network coverage was the main assumed factor and performance measure for influence maximisation problem identified firstly in [14]. From this point of view, the most central nodes, having a high influence on others, had the highest potential to be selected as seeds. Most of the seed selection methods focused on node network characteristics and heuristics improving the performance [15]. Usually, only the whole network structures are taken into account for seed selection.

While real campaigns take into account various node characteristics, the problem was emphasized by [5] and a targeted approach to viral marketing was proposed. It was based on assigning nodes to a potential market and searching for a local centrality score during the seeding process. For each user, the average importance factor was calculated to determine the impact on target group. Another study focused on targeting with the use of costs assigned to users within the network, together with the benefits related to the user interests [4]. It extends the typical approaches focused on assumption that users are acquired at the same costs with same benefits for marketers. As a result, the authors proposed a cost-aware targeted viral marketing with an effective computational approach, making the seeds selection within billion-scale networks possible. From the perspective of practical applications the authors took into account the number of posts under specific topics are a representation of user interest and potential benefits. While the earlier methods focused on influence maximisation based solely on centralities and influence, the study in [16] distinguished two classes of methods, taking into account more complex structural relations like overlap, and other group focused on user features and social information. They use, among others, trust between the users and cost. The study emphasises the lack of methods taking into account the user interest. The approach is based on the interest in the message. The experimental study was based on randomly assigned interest vectors within well-known datasets, without nodes' attributes. An integrated marketing approach was proposed in [6] for combining targeted marketing with viral marketing. The approach took into account users with revealed preferences and users with potentially high utility scores for the marketer. One of the goals was the maximization of information awareness and constraints focused on reaching the targeted users. The study [17] explored Cost-aware Targeted Viral Marketing model, with focus on the cost of the nodes' acquisition and potential benefits. Integer programming was used with the potential to search for close to exact solutions within large scale networks. From other perspective, the authors of [18] introduced a Targeted Influence Maximization problem, using an objective function

and penalization parameter for adoption of non-target nodes. The proposed approaches focused on general target groups characterized by benefits or knowledge acquired from user posts.

While targeting can be based on various performance evaluation criteria and campaign goals it creates space for applications or multi-criteria decision support methods. In the recent years some preliminary research has began in the area of utilising multi-criteria decision analysis (MCDA) techniques in the social network studies. Zareie et al. [19] used the TOPSIS method (Technique for Order Preference by Similarity to Ideal Solution) to reduce overlap and maximize coverage while influencing social networks. Yang et al. [20] used TOPSIS in the Susceptible-Infected-Recovered (SIR) model to dynamically identify influential nodes in complex networks, and in [21] used entropy weighting for setting the weights values. Liu et al. [22] used TOPSIS to evaluate the importance of nodes in Shanxi water network and Beijing subway networks by comparing each node's close degree to an ideal object. Robles et al. [23] used multiobjective optimization algorithms to maximize the revenue of viral marketing campaigns while reducing the costs. Wang et al. [24] proposed a Similarity Matching-based weighted reverse influence sampling for influence maximization in geo-social location-aware networks. Gandhi and Muruganantham [25,26] used TOPSIS to provide a framework for Social Media Analytics for finding influencers in selected networks. Montazerolghaem [27] used separately AHP and TOPSIS to provide rankings of effective factors in network marketing success in Iran. In their prior research, Karczmarczyk et al. [28] used the PROMETHEE II method (Preference Ranking Organization METHod for Enrichment of Evaluations) for evaluation of performance of viral marketing campaigns in social networks, as well as for decision support in the planning of such campaigns.

The up-to-date literature studies show a multitude of available MCDA methods [29]. Some examples of known and widely used MCDA methods include AHP, TOPSIS [30,31], or methods from the ELECTRE and PROMETHEE families [32]. The methods can be divided into three groups, based on the used approach. The first group, also known as the American school of MCDA methods, use the axiom of full variants comparability and two basic relations are available—indifference and preference of variants. The resulting model is aggregated into a single criterion [33]. The methods from the second group, also known as the European school of MCDA methods, are based on the axiom of partial comparability of variants. The aggregation takes place using the outranking relation. The third group consists of methods based on the foundations from both the aforementioned groups. The current taxonomy of the available MCDA methods can be found, for example, in [29,32,34].

The analysis of the existing works shows that among the large number of studies related to the information propagation and influence maximization, only a small fraction is focused on the very common real-life problem of targeting users with specific characteristics. The discussed approaches focused on single attributes and node characteristics for the seed selection to reach the assumed audiences or communities. Nonetheless, the social media skyrocketing is usually based on selection of parameters of the target group with various values of the attributes such as age, gender or localization, with different importance from the perspective of the campaign performance. This forms an interesting research gap, which is addressed in this paper with the proposed new approach. The approach is based on the assumption that, in order to maximize reaching a multi-attribute target group in the network, the seed selection process is also based on a multi-criteria evaluation of nodes. The seed selection process is supported with MCDA methods, allowing us to assign weights to individual attributes of the network nodes and produce rankings of seeds with the potential to increase the coverage in the addressed multi-attribute target group.

3. Methodology

In this section, the methodological framework of the approach proposed in this paper is presented. In Section 3.1, the assumptions regarding the multi-attribute nature

of the targeted nodes are presented. Subsequently, in Section 3.2, the problem of multi-criteria seed selection for targeting heterogeneous multi-attribute nodes is explained. Then, in Section 3.3, the MCDA foundations of the proposed approach are presented and the selection of the TOPSIS method is justified. Finally, in Section 3.4 the TOPSIS foundations and its adaptation for seed selection for targeting multi-attribute nodes are presented. The conceptual framework of the proposed approach is also visually presented on Figure 1.

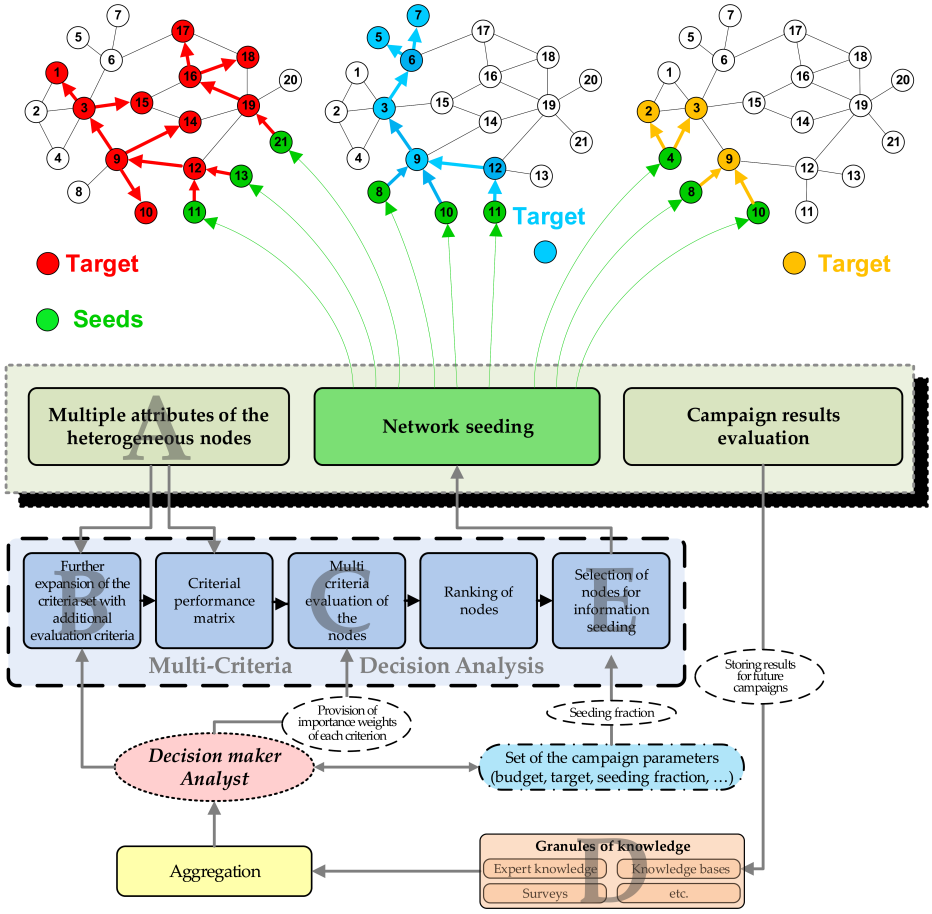


Figure 1. Conceptual framework of the proposed approach. Marks A–E provide anchors to be referred in the main text of the paper.

3.1. Multi-Attribute Nature of the Targeted Nodes

The proposed methodology complements the widely-used Independent Cascade (IC) model [14] for modeling the spread within the complex networks by taking into account the problem of reaching targeted multi-attribute nodes in social networks by the information propagation processes. In the proposed approach, it is assumed that the network nodes are characterized not only by the centrality relations between them and other nodes [35–37], but also by a set of custom attributes $C_1, C_2, \ldots, C_n$ (see Figure 1A).

The values of these attributes for individual vertices can be expressed as precise numerical values, such as age [years] or income [dollars]. Alternatively, if the attributes represent qualitative properties of the nodes, their values can be converted to numeric values with the use of 5-point Likert scale [38,39] (1—strongly disagree, 5—strongly agree) or enumerations (e.g., age: 1—young, 2—midle-aged, 3—old; or sex: 1—male, 2—female).

The nodes can also be characterized by the computed attributes derived from the network characteristics and measures. These include the centrality measures such as degree [35], closeness [40], betweenness [41] or eigenvector [36,37]. Additional attributes can also be derived as a composite of the two aforementioned types of attributes, by comput-

ing centrality measures based on limited subsets of the nodes' neighbors (see Figure 1B). For example, if attribute C_i represented the degree of a node, that is, the total count of its neighbors, the C_{i_1} could represent the count of its male neighbors, and C_{i_2} the count of its female neighbors.

The aim of the proposed methodological framework is to reach the targeted network nodes with multi-attribute characteristics, based on the multi-criteria process of selecting nodes for seeding in the process of information propagation.

3.2. Multi-Attribute Seed Selection

As was described in Section 3.1, in the proposed approach an attempt is made to reach the nodes with specific values of the selected attributes. For example, in preventive oncological social campaigns, an attempt is made to reach middle-aged women, that is, aged between 50 and 69.

In the independent cascade model [14], the information propagation process in a complex network is preceded by the selection of seeds. That means choosing a subset of network vertices, to which the information is provided at the beginning of the process, in order for them to pass the information further through the network. Normally, the seeds represent a given fraction of all network nodes. For example, the seeding fraction can be set to 5% of the network. There are numerous approaches to selecting the initial seeds, which generally result in producing a ranking of all network nodes and seeding information to the ones on top of the list.

Whilst other approaches focus on generating the ranking based on a single centrality measure, such as degree [35] or eigencentrality [36], in the authors' proposed approach, multiple attributes are considered in order to select the seeds with the highest potential to eventually propagate the information to the targeted nodes.

It is important to note, that in the proposed approach, the final coverage of the network, i.e., the fraction of nodes to which the information was eventually delivered, can be lower than in case of the traditional centrality-based approaches. However, the proposed method increases the chances to maximize the coverage within the targeted nodes' groups.

3.3. MCDA Foundations of the Proposed Approach and the Research Method Justification

The approach presented in this paper is based on the MCDA methodology foundations [42]. The adaptation of the MCDA methodology for the needs of seed selection resulted directly from the formal and practical assumptions of the research. First, the assumed modeling goal was an attempt to reach only the targeted set of multi-attribute nodes. Therefore, any attempt to obtain the optimal solution in a global sense (such as maximization of the global coverage) was disregarded in this research. Second, the fulfillment of the goals adopted in this research requires considering a number of attributes in the process of seed selection. Third, it was established that a compromise maximizing matching the required goals would be searched for, at the expense of the global network coverage.

The aforementioned premises of the multi-criteria modeling environment and goals, as well as the analysis of the formal components of the MCDA model at the stage of the model structuring and preference modeling, are the starting point for the selection of the appropriate MCDA method. It is worth noting that this is a significant problem, and an improper selection of the MCDA method can lead to incorrect results in the final decision model [29,32].

In this paper, the assumed effect of the construction and operation of the MCDA model is a ranking of variants [43]. The criterial performance of the variants will be expressed on a quantitative scale [44]. The expected result is a complete ranking of variants [45]. The deterministic simulation data environment present in this paper, shows the quantitative character of the input data. The research assumptions require that different weights of the individual criteria are taken into account, and their nature will also be quantitative. There is no need to use relative or absolute weighting criteria [46]. In the modeling process, it was also assumed that due to the deterministic nature of the simulation model being developed,

there is no natural uncertainty of the preferential information. In practice, this implies the use of the methods from the "American school" [45]. Based on [29,44], as well as the MCDA methods' set discussed in [32], using the expert system provided in [47], it is easy to show that aforementioned requirements are fully met only by the following set of MCDA methods: MAUT (Multi-Attribute Utility Theory), MAVT (Multi-Attribute Value Theory), SAW (Simple Additive Weighing), SMART (Simple Multi-Attribute Ranking Technique), TOPSIS (Technique for Order Preference by Similarity to an Ideal Solution), UTA (Utilites Additives), VIKOR (VIsekriterijumska optimizacija i KOmpromisno Resenje).

On the foundations of the aforementioned analysis, as well as based on the [32] formal recommendations, two groups of MCDA methods can be indicated as valid for solving the problem stated in this paper. The first one is based on an additive/multiplicative form of a utility/value function (MAUT, MAVT, SAW, SMART, UTA), and the second one is based on reference points (TOPSIS, VIKOR).

The former group of methods is founded on a very trivial mathematical principles—a simple aggregation of data and partial utilities. In practice, this results in transferring into the final models an undesirable effect of linear substitution of criteria. Consequently, this directly implies the possibility of obtaining incorrect rankings (failure to meet the level of individual criteria to a satisfactory degree).

Among the latter group, there is a significant level of similarity between both the TOPSIS and VIKOR methods. They both are based on the same assumptions and differ only in the chosen technique of normalization and aggregation of data. The TOPSIS method assumes minimizing the distance to the ideal solution and maximizing the distance to the anti-ideal solution, whereas in VIKOR only the distance to the ideal solution is minimized.

The principles of the TOPSIS and VIKOR methods, along with the fact that TOPSIS uses vector normalization (compared to linear normalization in VIKOR), expedite the selection of the TOPSIS method as the one which has the best potential in the considered problem of seeds' selection [48]. Consequently, it was the TOPSIS method that was chosen for the further stages of this research. Moreover, it is important to note that the chosen TOPSIS method does not require the attribute preferences to be independent [49–51]. This further strengthens the potential of using this method in the considered problem, in which, due to its preliminary character, we do not yet have full knowledge in the area of dependence or independence of the model attributes.

3.4. Multi-Criteria Seed Selection for Multi-Attribute Nodes Targeting

The Technique for Order Performance by Similarity to Ideal Solution (TOPSIS) is a widely-used MCDA method, originating from the American MCDA school. Originally formed by Hwang and Yoon [52], it is based on the concept that given a set of criteria and their possible values, a positive ideal solution (PIS), and negative ideal solution (NIS) can be indicated. These are a two hypothetical, non-existent, alternatives, whose all values for all criteria are either maximized (PIS) or minimized (NIS). When a set of alternatives are compared, in the TOPSIS method they are ranked based on their relative distance to the PIS and NIS. The best alternative should be as close as possible in terms of criteria values to the PIS, and as far as possible from NIS.

In the proposed approach, the TOPSIS method is used for multi-criteria evaluation of the nodes (see Figure 1C). First of all, the criteria for evaluation of the potential seeding nodes need to be chosen. Then, a decision matrix $D[x_{ij}]$ is built based on the criteria values of all vertices in the studied network, in which the m rows represent the vertices and n columns represent the criteria (see Equation (1)):

$$D[x_{ij}] = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} \end{pmatrix} \quad (1)$$

In the second step of the algorithm, the decision matrix is normalized. Different formulae are used for the benefit criteria (2) and different for the cost criteria (3):

$$\&r_{ij} = \frac{x_{ij} - min_i(x_{ij})}{max_i(x_{ij}) - min_i(x_{ij})} \tag{2}$$

$$\&r_{ij} = \frac{max_i(x_{ij}) - x_{ij}}{max_i(x_{ij}) - min_i(x_{ij})} \tag{3}$$

The MCDA-based approaches extend the traditional aggregating approaches by the fact that the weights of individual decision attributes can be adjusted to varying values. The analyst adjusts the weights of each decision criterion to the preferences of the decision maker. In the case of the considered problem of seed selection, the marketer adjusts the weights of individual criteria to increase as much as possible the potential to reach to the targeted network nodes through the seeded network nodes. The weights are chosen based on the analyst's knowledge, skills and experience (see Figure 1D). Therefore, in the third step of the TOPSIS algorithm used in the authors' proposed approach, the weights are imposed on the decision matrix and, consequently, a weighted normalized decision matrix is constructed:

$$v_{ij} = w_j \cdot r_{ij} \tag{4}$$

In the fourth step of the algorithm, the positive and negative ideal solutions (V_j^+ and V_j^+ respectively) are computed (Equations (5) and (6)). In the case of the studied seed selection problem, the positive ideal solution would represent a vertex, which for all criteria has the best possible values, whereas the negative ideal solution would be a vertex with the worst possible values for each criterion.

$$V_j^+ = \{v_1^+, v_2^+, v_3^+, \ldots, v_n^+\} \tag{5}$$

$$V_j^- = \{v_1^-, v_2^-, v_3^-, \ldots, v_n^-\} \tag{6}$$

In the penultimate, fifth, step of the TOPSIS method, the Euclidean distances between each network vertex and the positive and negative ideal solutions are computed:

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^+)^2} \tag{7}$$

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2} \tag{8}$$

Eventually, the relative closeness of each vertex to the ideal solution is computed:

$$CC_i = \frac{D_i^-}{D_i^- + D_i^+} \tag{9}$$

The obtained CC_i scores are then used to rank the vertices and build the final ranking, which then can be used for selecting the vertices for the initial network seeding (see Figure 1E).

All in all, the MCDA foundations of the proposed approach facilitate obtaining network nodes' rankings with the highest, according to the analyst, potential to reach the targeted nodes in the social network. Moreover, the use of MCDA allows us to study the stability of the obtained ranking with sensitivity analyses. This, in turn, allows us to study the effect of each individual criterion on the final ranking and, therefore, allows us to iteratively improve the obtained solution.

4. Empirical Study

4.1. Real-Life Usage Example

In this section, a brief real-life usage example of the proposed approach will be presented, explaining every step of the proposed framework on a small real network. In further sections, a more in-depth analysis is performed on a larger synthetic network.

The empirical example in this section will be performed on a real network. Enron emails network [53] was selected due to its limited size (143 nodes and 623 edges), which allows us to study in detail the status of every single node of the network. It is important to keep in mind that the proposed approach is intended for networks with nodes characterized by multiple attributes. Due to the fact that the publicly available network repositories principally provide only edge lists of networks, the attributes had to be overlaid on the network artificially. Therefore, artificial values for two attributes were generated for the network, based on [54]: gender (69 nodes male, and 74 nodes female), and age (0–29 years—62 nodes, 30–59 years—55 nodes, over 60 years—26 nodes).

For such a network, for illustrative purposes, two complete scenarios with two different targets will be presented. In both, a constant propagation probability (0.1) and seeding fraction (0.05, i.e., 7 vertices) is assumed.

4.1.1. Target 1: Male Aged 0–29

In this scenario, the aim of the viral marketing campaign is to reach men aged 0–29, that is, the targets are described by specific values of two criteria: gender (C2) and age (C5). The target group, therefore, consists of 28 nodes (see Figure 2). Apart from the two target-describing attributes, some other criteria are also available: degree (C1), degree male (C3), degree female (C4), degree aged 0–29 (C6), degree aged 30–59 (C7), degree aged 60+ (C8). The decision maker (DM)/analyst, based on their expertise, provide the preference weights for all criteria: C1: 8.20, C2: 25.40, C3: 12.60, C4: 3.80, C5: 28.40, C6: 14, C7: 3.80, C8: 3.80. These weights are provided by the DM as input data to the proposed approach, as the ones which, according to the DM, allow to rank the nodes in order to find the seeds potentially best for maximizing influence in the targeted group. In order to provide such weights, the analyst can refer to archival knowledge and use decision support systems or MCDA methods such as AHP [39].

Once the preference weights are known, the TOPSIS method is used to evaluate all vertices. The top seven (seeding fraction 0.05) are chosen as seeds and the campaign is started.

For this scenario, the simulations (see Figure A1 in Appendix A) have shown the campaign averagely reached 9/28 targeted nodes (32.14%), with global coverage 0.2224. A traditional degree-based approach for the same network results averagely in reaching 7.7/28 targeted nodes (27.5%), with global coverage 0.2881. The multi-criteria approach reached 4.64% more of the targeted nodes with global coverage lower by 0.0657.

4.1.2. Target 2: Female Aged 30–59

In this scenario, the aim of the viral marketing campaign is to reach women aged 30–59. The target group consists of 24 nodes (see Figure 2). Again, apart from the two target-describing attributes, some other criteria are also available: degree (C1), degree male (C3), degree female (C4), degree aged 0–29 (C6), degree aged 30–59 (C7), degree aged 60+ (C8). It is important to note that, contrary to other approaches [4], in the proposed approach the criteria values are reused and only the preference weights are adjusted. This time, the decision maker, based on their expertise, provide the following preference weights for the criteria: C1: 4.4, C2: 30.4, C3: 4, C4: 10.4, C5: 30.40, C6: 5.4, C7: 10.4, C8: 4.4.

Vertex		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Attribute	Sex	1	2	1	1	1	1	1	2	1	2	2	1	2	2	1	2	2	2	1	1	1	2	2	2	2	1	2	2	1	1	1	2	2	1	1	2
	Age	2	1	1	1	1	3	2	1	2	2	2	3	1	1	1	3	1	1	1	1	3	3	1	1	1	3	1	2	2	2	2	2	3	2	2	1
Target	1	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
Ranking	Deg.	132	30	33	37	122	66	98	128	34	77	76	91	42	59	139	55	2	48	17	11	90	28	95	72	103	45	47	52	43	105	8	12	120	115	49	89
	1	94	20	7	6	47	107	89	77	53	118	124	115	43	51	46	134	5	38	3	1	116	112	70	55	73	103	27	100	28	90	13	99	141	96	68	63
	2	102	41	117	125	137	129	92	83	58	13	12	124	44	56	143	53	4	49	110	109	126	40	68	62	65	115	60	14	79	95	39	2	75	98	70	66

Vertex		37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Attribute	Sex	1	1	1	2	2	2	2	2	1	1	2	1	1	1	2	2	1	2	2	2	2	1	2	2	2	2	2	1	1	1	2	1	1	2	1	2
	Age	2	3	3	1	1	1	3	2	2	2	1	2	2	1	2	1	2	1	2	1	2	3	2	2	2	1	1	2	2	1	2	1	2	2	1	2
Target	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0
	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	1	0	0	1	0	1
Ranking	Deg.	24	118	67	97	74	141	6	110	106	117	16	4	50	15	46	126	26	79	109	25	84	68	119	27	93	130	140	29	21	41	19	38	18	81	39	13
	1	62	117	106	71	59	82	102	128	83	91	26	15	60	12	92	80	45	40	129	33	122	108	132	97	126	74	79	44	25	18	52	23	16	119	8	66
	2	37	136	119	63	51	86	27	25	99	100	33	35	67	107	10	82	46	69	22	34	17	118	28	6	19	84	90	64	47	116	7	112	45	21	121	5

Vertex		73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108
Attribute	Sex	1	2	2	1	1	2	1	1	2	1	2	2	2	1	2	2	1	2	2	1	1	1	1	1	1	1	2	2	1	1	1	2	2	1	2	2
	Age	2	3	2	1	1	2	1	2	2	2	2	1	1	3	3	2	1	1	1	1	1	2	2	1	2	2	3	1	3	1	2	1	3	2	2	3
Target	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
	2	0	0	1	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Ranking	Deg.	75	9	22	88	32	83	102	143	134	82	101	104	63	92	121	134	123	61	7	138	124	100	3	124	131	142	71	44	36	57	99	60	1	108	70	112
	1	75	120	104	30	17	123	32	95	130	76	125	69	48	109	138	130	37	54	21	49	42	84	9	42	98	93	135	39	101	19	86	56	11	88	111	140
	2	81	24	3	130	114	16	134	104	30	91	26	73	54	132	80	30	141	52	31	142	140	94	23	140	101	105	57	43	113	128	93	48	1	96	11	71

Vertex		109	110	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143
Attribute	Sex	1	1	1	2	2	1	2	1	2	2	2	2	1	1	2	2	2	2	2	1	1	1	2	1	2	2	1	1	2	2	1	2	2	1	1
	Age	2	1	1	1	3	1	2	2	1	1	1	3	2	1	3	3	2	1	1	2	1	1	3	1	1	2	3	1	3	3	1	1	2	1	2
Target	1	0	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	1	0	0	1	0
	2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
Ranking	Deg.	64	94	78	62	136	113	53	108	80	20	73	86	35	23	31	14	51	96	114	116	88	58	111	5	129	85	133	10	54	136	56	127	69	40	65
	1	64	34	31	61	142	36	110	88	67	35	65	137	50	10	133	127	114	58	72	85	30	22	139	4	78	113	121	2	136	142	24	81	105	14	57
	2	89	133	127	42	88	138	9	96	61	32	50	55	59	111	36	20	8	74	76	103	130	122	72	106	78	18	135	108	38	88	123	77	15	120	85

Figure 2. Visual presentation of two real-life usage scenarios for targeting male aged 0–29 (target 1) or female aged 30–59 (target 2). The table contains: values of the sex and age attributes, information on targeted nodes for both scenarios, and the rankings of nodes for seeding.

Once the preference weights are known, the TOPSIS method is used to evaluate all vertices. The top seven (seeding fraction 0.05) are chosen as seeds, and the campaign is started.

For this scenario, the simulations (see Figure A2 in Appendix A) have shown the campaign on average reached 9.5/24 targeted nodes (39.58%), with global coverage 0.2552. A traditional degree-based approach for the same network results averagely in reaching 6.8/24 targeted nodes (28.33%), with global coverage 0.2881. The multi-criteria approach reached 11.25% more of the targeted nodes with global coverage lower by 0.0329.

4.1.3. Real-Life Example Discussion

In the real-life example, two complete scenarios with two different targets were presented. As expected, in both cases the proposed approach resulted in lowering the global coverage but increasing the influence in the targeted set of nodes. In both cases, it was the decision-maker (DM) who first determined the values for weights. This is a subjective assessment, based on the DM's knowledge, skills and experience. In case the weights would have been estimated improperly, the ranking of the nodes would be ordered differently, and, therefore, different 7 nodes would be selected as seeds (see Section 3.4). This, in turn, could result in reaching fewer targeted nodes in the network (see Section 4.8).

The actual participation of the decision-maker in the process of solving the task is very important in MCDA, and the actual performance of the obtained solution is dependent on both the quality of the attributes and the proper selection of the values of the vector of the relative importance of the decision model criteria. Attempting to obtain the maximum potential to reach through the seeded nodes to the targeted nodes requires searching for the most satisfying values of the vector of the relative importance of the decision model criteria.

4.2. Setup of the Comprehensive Experiment

The basic usage example presented above is followed by a set of three more in-depth analysis scenarios, performed on a larger synthetic network. In order to illustrate the proposed approach, the empirical study was performed on a Barabasi-Albert (BA) synthetic network [55]. The Barabasi-Albert network model was created as an outcome of a research of the structure of the WWW in the 90's. Two complementary mechanisms drive the construction of BA networks: network growth and preferential attachment. In the BA synthetic networks, several selected nodes (hubs) have an unusually high degree compared to the other vertices in the network.

Over the recent years, there has been an abundance of research showing that a vast number of social networks, both virtual and real, are scale-free in their nature [55–58]. Their degree k follows a power law $k^{-\lambda}$ and exponent λ is typically $2 < \lambda < 3$. The sample network was generated with exponent λ with value in the middle of this range $\lambda = 2.5$. Moreover, in order to allow clear visualisation of the network, the vertices count was set to 1000. The resulting network was characterized by the following the average values of its centrality metrics:

- Betweenness—1687.295;
- Degree—3.994;
- Closeness—0.0002310899;
- Eigen Centrality—0.03661858.

Since the proposed approach is intended for networks whose nodes are described with multiple attributes, the subsequent step was to assign a set of attributes to each of the vertices of the obtained network. The most of publicly available network datasets are based mainly on set of nodes and edges, without node attributes. To overcome this problem, we used node attributes following distributions from demographic data. It is similar to approach presented in [16]. The information on sex distribution from demographic data was overlaid on the network to obtain the first attribute [54]. This resulted in 470 network nodes marked as male and 530 marked as female. Subsequently, the age distribution information [54] was used to add to the network the second attribute, with three possible values:

- young, i.e., aged 0–49, 64.62% of the population;
- mid-aged, i.e., aged 50–69, 25.34% of the population;
- elderly, i.e., aged 70 and above, 10.04% of the population.

Finally, the goal of the information spreading campaign was chosen for the empirical research. For illustrative purposes, it was decided that a real-life example of social campaign for a breast cancer prevention program (mammography) would be used [59]. This campaign targets women aged 50–69, which in the case of the network generated for this experiment translated to 130 out of the total of 1000 nodes of the network.

4.3. Criteria for Seed Selection

As was described in Section 3, in the proposed approach the initial seeds were selected from the network based on multiple criteria. In the case of the studied synthetic network, apart from the sex and age attributes, the general degree of each node was also taken into account, as well as the degree measurements based on each value of the two attributes. This resulted in a total of eight evaluation criteria, presented in Table 1.

Table 1. Seed selection criteria.

No	Criterion	Preference
C1	Degree	max
C2	Sex (Match/Mismatch)	min
C3	Degree Male	max
C4	Degree Female	max
C5	Age (Match/Mismatch)	min
C6	Degree Young	max
C7	Degree Mid-Aged	max
C8	Degree Elderly	max

The criterion C1 represents the number of neighbors of each evaluated vertex. Criterion C2 is based on the sex attribute and is equal to 0 if there is a match between the targeted and actual sex or 1 in the case of a mismatch. Criterion C3 represents the count of male neighbors of a vertex, whereas criterion C4 represents female neighbors of a vertex. In turn, criterion C5 indicates the difference between the targeted and actual age group of a vertex. For example, if the targeted age group was young, vertices from age groups young, mid-aged and elderly would obtain the values of 0, 1 and 2 respectively. Since the targeted group in this experiment is in the middle, that is, mid-aged, vertices from this group would obtain value 0 and from other groups would obtain value 1 for criterion C5. Last, but not least, criteria C6, C7 and C8 represent the count of respectively young, mid-aged and elderly neighbors of a vertex. All criteria C1–C8 were then assembled to create a single decision matrix for the TOPSIS method. At this stage, it is important to note that during the research the authors decided to follow the degree-based criteria, as the degree is the most basic measure which can be used for benchmarking of the approach. If other measure, such as closeness, betweenness, eigencentrality, and so forth, was used as criterion C1, also the remaining criteria C3, C4, C6, C7, C8 would need to be modified to use the selected metric.

The last step required for the seed-selection setup was specifying the preference direction of all evaluation criteria C1–C8. Because criteria C2 and C5 represent difference between the targeted and actual values, the lowest possible values were preferred. On the other hand, since the remaining criteria are based on the degree network centrality measure, the preference direction for these criteria was maximum.

After the experiment was set up, three scenarios based on various weights of individual criteria were studied. Their description and results are presented in the following sections.

4.4. Scenario 1: Single Criterion

The first scenario studied was intended to be similar to the approaches that are based solely on a single centrality measure, here—the degree. Therefore, the preference weights for the TOPSIS ranking-generation method were set to a significant value of 100 for C1, and a negligible value of 1 for all other criteria. All vertices were evaluated and ordered by rank. It was decided, that in the simulations the seeding fraction of 0.05 and propagation of 0.3 will be used. Therefore, the 50 vertices with the highest CCi scores were selected as seeds (see Table 2).

The analysis of Table 2 allows us to observe that the best vertex, labelled *3* obtained significantly more score than any other vertex (0.9975 compared to 0.6800 and 0.6000 for vertices *4* and *2* ranked 2 and 3, respectively). It is also noticeable that the score of the best vertex *3* was over two-fold higher than the score of vertices *24* and *1* ranked 6/7, with an equal score of 0.4400. These scores can be confirmed, when the degree measure of each of the nodes is verified. The degree of the leading vertex *3* is equal to 52, followed by 36, 32, 29, 28 for vertices *4*, *2*, *12*, *5* respectively and 24 for vertices *1* and *24*. Last, but not least, it can be observed that because the degree was used as the main criteria for the selection

of seeds, multiple of the selected nodes are scored equally, for example all nodes ranked 40–45 are scored 0.1800 and all nodes ranked 46–50 are scored 0.1600.

Table 2. Seeds selected for Scenario 1, ordered by their rank and CCi score obtained in the applied TOPSIS method.

Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score
1	3	0.9975	11	49	0.4000	21	29	0.2800	31	18	0.2400	41	151	0.1800
2	4	0.6800	12	6	0.4000	22	170	0.2800	32	153	0.2400	42	97	0.1800
3	2	0.6000	13	11	0.3800	23	47	0.2800	33	57	0.2200	43	65	0.1800
4	12	0.5400	14	16	0.3400	24	21	0.2600	34	10	0.2200	44	59	0.1800
5	5	0.5200	15	26	0.3400	25	14	0.2600	35	40	0.2200	45	101	0.1800
6	24	0.4400	16	7	0.3400	26	45	0.2600	36	238	0.2200	46	36	0.1600
7	1	0.4400	17	113	0.3400	27	103	0.2600	37	56	0.2000	47	116	0.1600
8	30	0.4200	18	135	0.2800	28	82	0.2600	38	172	0.2000	48	37	0.1600
9	185	0.4200	19	17	0.2800	29	9	0.2400	39	20	0.1801	49	93	0.1600
10	19	0.4000	20	53	0.2800	30	42	0.2400	40	143	0.1800	50	55	0.1600

After the seeds were selected, the campaign was simulated over the same network, with the same seeds for 10 consecutive times. In order to allow repeatability of the simulation conditions, a set of 10 pre-drawn weights for each connection (edge) in the network was used. The outcomes of each simulation were stored and presented in the form of a visual graph (see Figure A3 in Appendix A). On average, the simulation took 8.6 iterations and resulted in 433.6 nodes being infected (0.4336 coverage). However, only 50.5 nodes of the 130 targeted nodes were infected (0.3885 target coverage).

4.5. Scenario 2: Two Criteria

In the second scenario, the preference weight of the degree measure was reduced in favor of the more accurate female degree (C4) and mid-aged degree (C7). Therefore the weights of C4 and C7 were set to 100 while the weights of the rest of the criteria was set to 1. All vertices were evaluated again, under the new conditions and their ranking was built. The correlation coefficient between the rankings for both scenarios is equal to 0.9022 for the scores and 0.7510 for the ranks of the vertices. The results of the top 50 vertices, selected as seeds, are presented in Table 3.

Table 3. Seeds selected for Scenario 2, ordered by their rank and CCi score obtained in the applied TOPSIS method.

Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score
1	3	0.9980	11	30	0.4075	21	20	0.3645	31	116	0.3073	41	34	0.2560
2	4	0.8142	12	9	0.4048	22	18	0.3606	32	26	0.3045	42	93	0.2476
3	2	0.7554	13	19	0.4036	23	7	0.3482	33	29	0.3045	43	464	0.2476
4	5	0.5836	14	11	0.3936	24	170	0.3482	34	152	0.3044	44	14	0.2445
5	12	0.5392	15	113	0.3857	25	153	0.3442	35	174	0.2913	45	48	0.2445
6	24	0.5178	16	17	0.3857	26	185	0.3260	36	82	0.2900	46	56	0.2354
7	6	0.4741	17	42	0.3856	27	53	0.3260	37	10	0.2840	47	69	0.2341
8	1	0.4452	18	21	0.3708	28	172	0.3250	38	238	0.2839	48	33	0.2341
9	135	0.4296	19	57	0.3658	29	16	0.3135	39	195	0.2839	49	97	0.2325
10	49	0.4164	20	143	0.3658	30	47	0.3135	40	122	0.2589	50	295	0.2325

When Table 3 is analyzed, it is clearly visible that the scores obtained by the best vertices are much more diversified than in case of the first scenario. The three leading vertices are still the ones labelled *3*, *4* and *2*; however, the order of the subsequent two has changed. The vertex *5* is now ranked 4 with the score of 0.5836 (previously 0.5200), followed by the vertex *12* now scored 0.5392 (previously 0.5400). The vertex *24* remained on position 6; however, it is now followed by vertex *6*, scored 0.4741, which in the previous scenario was ranked 12th with the score of 0.4000. A detailed analysis of the differences between ranks obtained by vertices in the rankings for scenarios 1 and 2 is presented on Figure 3A. The horizontal axis presents the consecutive ranks of all 1000 vertices of the studied network in scenario 1, whereas the vertical axis shows how these vertices were then ranked in scenario 2. The closer the point representing a vertex is to the diagonal line

on the chart, the smaller the change in the rank occurred. It can be observed, that while in case of the top-ranked vertices only small changes in rank occur, as it can be confirmed in Table 3, in the case of the vertices further down the list, changes of even hundreds of levels in rank can be observed.

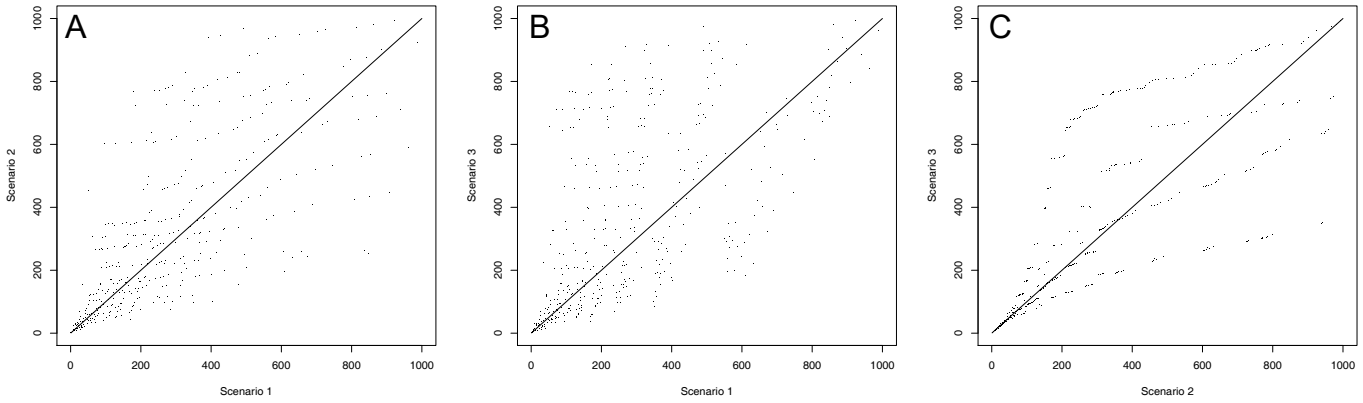


Figure 3. Visual comparison of ranks of nodes obtained in rankings for various scenarios: (**A**) scenarios 1 and 2; (**B**) scenarios 1 and 3; (**C**) scenarios 2 and 3.

Subsequent to the selection of the seeds, ten simulations were performed with the same conditions as in the first scenario. The visual representation of the outcomes of the simulations are presented in Figure A4 in Appendix A. In this scenario, the simulations averagely lasted 9.1 iterations, that is, longer by 0.5 iteration and resulted in 435.6 nodes infected (0.4356 coverage, 0.0020 more). What is interesting, the usage of two criteria allowed us to increase the coverage in the target group. Averagely 52 targeted nodes were infected, that is, 0.4 target coverage, which is 0.0115 more than in the first scenario.

4.6. Scenario 3: Four Criteria

In the third scenario, it was decided to focus on seeding information not only to vertices with high values of female degree (C4) and mid-aged degree (C7), but also to nodes which are already in the target group, that is, the right sex (C2, female) and age (C5, mid-aged). The seeds selected for this scenario are presented in Table 4.

Table 4. Seeds selected for Scenario 3, ordered by their rank and CCi score obtained in the applied TOPSIS method.

Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score	Rank	Vertex	Score
1	3	0.9069	11	9	0.4120	21	20	0.3750	31	29	0.3197	41	122	0.2782
2	4	0.7842	12	11	0.4023	22	153	0.3561	32	185	0.3197	42	34	0.2731
3	2	0.7191	13	30	0.3985	23	170	0.3535	33	116	0.3148	43	33	0.2717
4	5	0.5821	14	19	0.3950	24	18	0.3534	34	152	0.3125	44	93	0.2679
5	24	0.5291	15	143	0.3862	25	7	0.3412	35	174	0.3067	45	14	0.2660
6	12	0.5248	16	21	0.3810	26	53	0.3326	36	195	0.2934	46	130	0.2577
7	6	0.4782	17	113	0.3775	27	172	0.3315	37	82	0.2846	47	69	0.2566
8	1	0.4508	18	17	0.3774	28	16	0.3279	38	464	0.2822	48	97	0.2543
9	49	0.4236	19	42	0.3774	29	47	0.3278	39	10	0.2788	49	74	0.2474
10	135	0.4198	20	57	0.3757	30	26	0.3197	40	238	0.2788	50	104	0.2474

The analysis of Table 4 shows that the vertex *3* is still the leading one, however its score is much lower in case of this scenario (0.9069, compared to 0.9975 and 0.9980 in scenarios 1 and 2 respectively). Some minor changes in ranks can also be observed for the remaining seeds. Figure 3B visualizes the comparison of ranks between scenarios 1 and 3, whereas Figure 3C between scenarios 2 and 3. The analysis of these figures allows us to visually observe that the ranking obtained in scenario 3 is more similar to the one obtained in scenario 2 than to the one in scenario 1. This can be confirmed, indeed, by comparing the correlation coefficients between all scenarios (see Table 5).

Table 5. Correlation matrix between the three scenarios' ranks (**A**) and scores (**B**).

(A) RANKS	Scenario 1	Scenario 2	Scenario 3	(B) SCORE	Scenario 1	Scenario 2	Scenario 3
Scenario 1	x	0.7510	0.7099	Scenario 1	x	0.9022	0.8186
Scenario 2	0.7510	x	0.7308	Scenario 2	0.9022	x	0.8933
Scenario 3	0.7099	0.7308	x	Scenario 3	0.8186	0.8933	x

The results of the ten simulations performed for this scenario under the same conditions as used previously, are visually presented in Figure A5 in Appendix A. The average duration of the simulations was 8.7 iterations, which is slightly longer than in scenario 1 but shorter than that in scenario 2. On average, 435 nodes were infected (0.4350 coverage), which, similarly, is better than scenario 1 but worse than scenario 2. Finally, averagely 52.7 targeted nodes were infected, that is, 0.4054 targeted coverage, which is 0.0054 better than in scenario 2 and 0.0169 better than in the traditional approach, mimicked in scenario 1 (see Tables 6 and 7).

Table 6. Average simulation results for scenarios 1–3.

Scenario	Preferences	Avg. Last Iter.	Inf. Nodes	Coverage	Targeted Inf. Nodes	Targeted Coverage
1	100-1-1-1-1-1-1-1	8.60	433.60	0.4336	50.50	0.3885
2	1-1-1-100-1-1-100-1	9.10	435.60	0.4356	52.00	0.4000
3	1-100-1-100-100-1-100-1	8.70	435.00	0.4350	52.70	0.4054

Table 7. Comparison of differences between the average simulation results for scenarios 1–3.

Average Last Iteration				Average Coverage				Average Targeted Coverage			
Δ	**S1**	**S2**	**S3**	**Δ**	**S1**	**S2**	**S3**	**Δ**	**S1**	**S2**	**S3**
S1	x	−0.5	−0.1	**S1**	x	−0.0020	−0.0014	**S1**	x	−0.0115	−0.0169
S2	0.5	x	0.4	**S2**	0.0020	x	0.0006	**S2**	0.0115	x	−0.0054
S3	0.1	−0.4	x	**S3**	0.0014	−0.0006	x	**S3**	0.0169	0.0054	x

4.7. Sensitivity Analysis

As it was observed in Sections 4.4–4.6, depending on the preference weights regarding evaluation criteria, the evaluation score of each vertex varied, resulting in differences in the obtained rankings and diverse sets of initial seeds for performing the information propagation campaign. The MCDA methodological foundations of the proposed approach allow to perform sensitivity analysis of the obtained rankings, and thus recognize how changes in the criteria preference affect the final rankings and, in turn, the selected seeds.

In this section, a sensitivity analysis for the seed selection problem for the studied network is presented. For clarity, the subset of analyzed vertices was limited to the ones which were selected as seeds in any of the scenarios 1–3. This resulted in a subset comprising of a total of 63 vertices: 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 14, 16, 17, 18, 19, 20, 21, 24, 26, 29, 30, 33, 34, 36, 37, 40, 42, 45, 47, 48, 49, 53, 55, 56, 57, 59, 65, 69, 74, 82, 93, 97, 101, 103, 104, 113, 116, 122, 130, 135, 143, 151, 152, 153, 170, 172, 174, 185, 195, 238, 295, 464.

In order to perform the sensitivity analysis, at first the weights of all criteria were set to 1. Then, the weight of each criterion was gradually changed to 1, 25, 50, 75 and 100, while the rest of criteria remained at an unchanged level. Afterwards, the level of all criteria was increased to 25, and each criterion was tested again with the weight of 1, 25, 50, 75 and 100, while the rest of the criteria remained at an unchanged level. The same was then repeated for the levels of 50 and 75. At each combination of weights, the TOPSIS method was used to compute a ranking. The score and ranks of each of the 63 studied vertices was stored, and plotted afterwards. The plots representing the changes of score of each vertex is presented in Figure 4. The changes of ranks are presented in Figure 5.

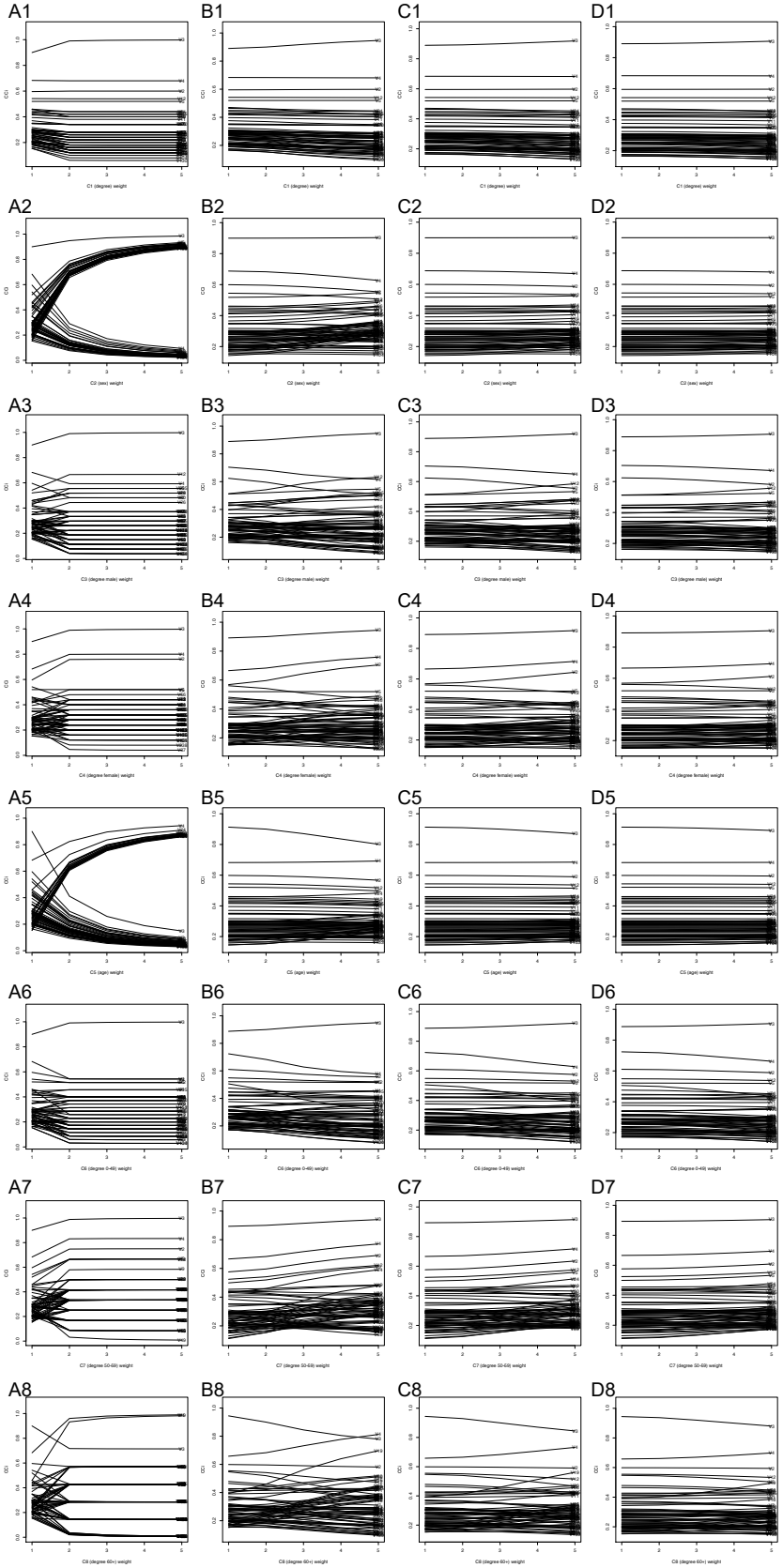


Figure 4. Sensitivity analysis on the subset of 63 network vertices. The charts represent how changes in a single criterion (1–8) affect the score obtained by the analysed vertices, when the weights of the other criteria are set to 1 (**A**), 25 (**B**), 50 (**C**) or 75 (**D**).

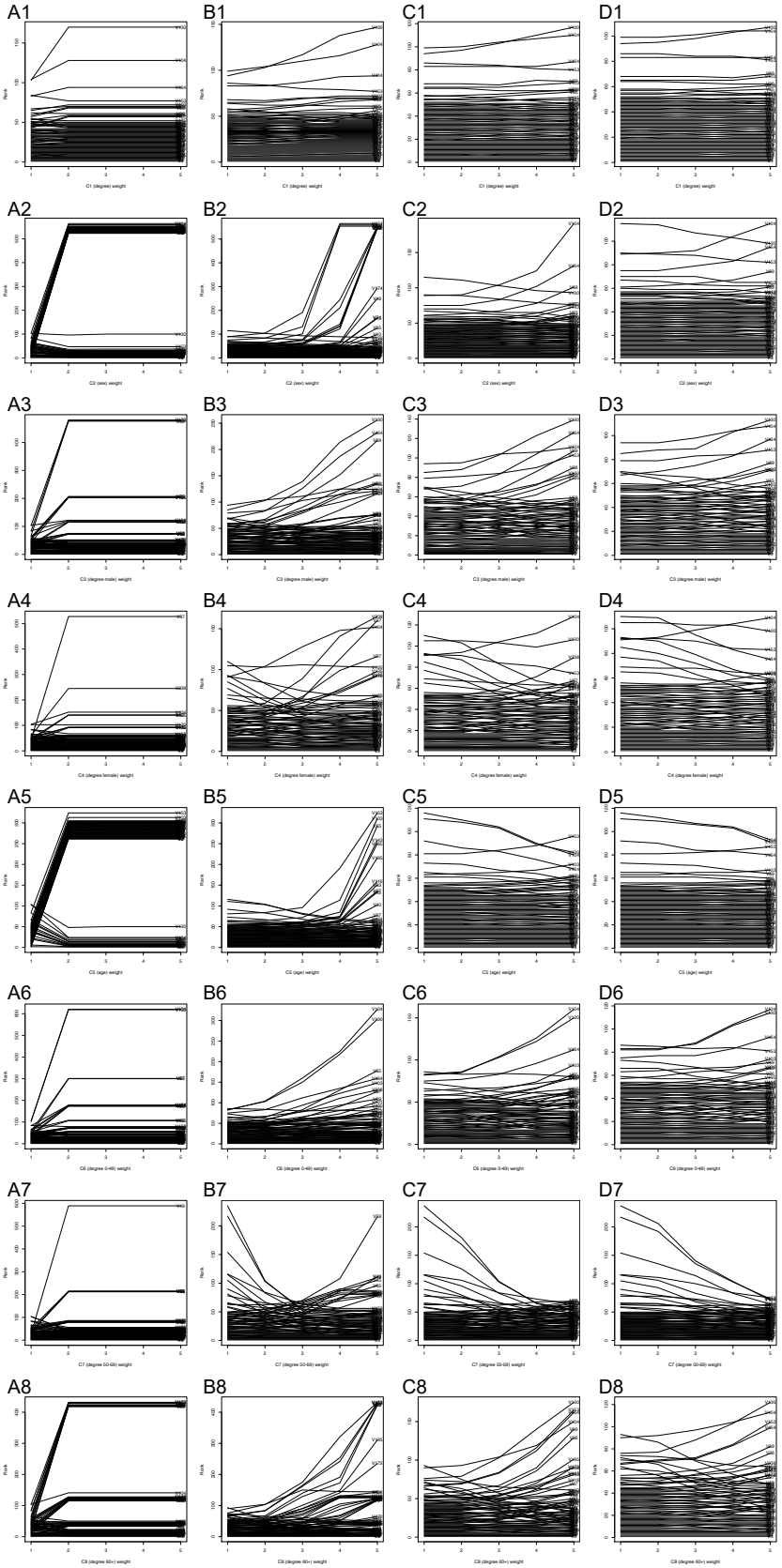


Figure 5. Sensitivity analysis on the subset of 63 network vertices. The charts represent how changes in a single criterion (1–8) affect the ranks obtained by the analysed vertices, when the weights of the other criteria are set to 1 (**A**), 25 (**B**), 50 (**C**) or 75 (**D**).

The analysis of Figure 4A shows how each of the criteria support or conflict with individual vertices. It is particularly clear because, while the weight of each criterion is increased in the range 1–100, the weights of the remaining criteria are locked at the level of 1. The chart A8 demonstrates that, in some cases, the vertex *3*, which was the leading one in all three exemplary scenarios, in some cases can be outran by other vertices. If the weight of criterion C8 (elderly degree) was increased to 25, while the weights of the other criteria remained negligible at the value of 1, the score of vertex *3* would drop below 0.8 and it would be ranked 3rd. However, if the weights of the other criteria were levelled at 25, the vertex would be the leader again, unless the weight of criterion C8 was increased close to 100. Then the vertex *3* would be ranked second.

Similarly, as can be observed in chart A5, if the weight of criterion C5 (age) was increasing, yet the other weights remained at 1, the vertex *3* would lose score very fast, down to a level of approximately 0.2. However, if the weights of the other criteria were increasing, the downfall of the score would be reduced to 0.8 (B5) or even 0.9 (C5, D5).

An interesting observation can be made looking at charts A1–A8. As was seen in Table 2 in Section 4.4, many vertices obtained the same score, and therefore their rank could vary. During the sensitivity analysis, this resulted in plots for multiple vertices being superimposed one on another. For example, on chart A1, only vertices *3*, *4*, *2*, *12* and *5* can be located easily, while the remaining vertices are stacked together on the chart.

Because criterion C1 is based on the degree centrality measure, the vertices' plots cluster in multiple score-groups, based on a plentiful, yet enumerable set of possible degree values, in the case of the studied network. On the other hand, due to the fact that the criteria C7 and C8 are based on the degrees of less numerous social groups (mid-aged and elderly), the possible values of the degree measure are more limited in this case and, therefore, there are less possible score values, which can be observed on the charts A7 and A8. In case of the chart A2, it can be observed that if the vertices are appraised based on the criterion C2 (sex), where only two values are possible, the vertices cluster in two groups. Since both sexes are distributed in the studied network at a roughly even probability level, it can be observed on the chart that both groups of vertices' plots are similar in size. On the other hand, however, in case of criterion C5, also only two values are possible, so the vertices are plotted in two groups too. However, because only about a quarter of the studied network is in the targeted middle-aged group, a clear disproportion between the groups of plots can be observed on the chart A5.

Whilst in the case of Figure 4, the values on the vertical axis were limited to the range from 0 to 1, and multiple vertices were allowed to have the same value, in case of Figure 5 each value can be assigned only to a single vertex at a time. As was mentioned earlier, the set of analyzed vertices is limited to 63 for readability. The charts on Figure 5 are scaled to show ranks from 1 (best) to the worst one obtained by any of the 63 studied vertices. It is important to reiterate, that each of the 63 studied nodes was in the group of 50 best vertices in one of the scenarios described above. Therefore it is very unforeseen to observe that the chart C1 ends at about rank 120, obtained by the worst vertex *130*, and the chart A6 ends around rank 600 for vertices *104* and *130*. These observations emphasize the importance of proper selection of seeds for information spreading campaigns in social networks.

4.8. Full Range Analysis

The empirical study was concluded by performing a comprehensive set of 65,610 simulations based on the full range of the seed selection preference weights. For each of the eight decision criteria, the weights of 1, 50 and 100 were assigned. That resulted in 3^8 possible sets of criteria preference weights and, consequently, 6561 sets of seeds, for each of which ten simulations under invariable conditions were performed. The results of the performed 65,610 simulations were then stored and aggregated for further analysis.

For the studied synthetic network, the highest number of infected vertices was reached for the seeds indicated by rankings based on high weights of the C5 (age) criterion, and negligible weights of the other criteria. It was equal to 459.7 infected nodes, that is, 0.4597

coverage. For such scenarios, averagely 61.3 targeted nodes were infected, that is, 0.4715 coverage of the targets.

On the other hand, the highest coverage within the targeted nodes was achieved in the simulations originating from the rankings produced by the scenarios in which high weight values were assigned to criteria C2 (sex) and C5 (age). On average 75.8 targeted nodes were infected in these simulations, that is, 0.5831 targets' coverage. For these scenarios, on average 458.6 vertices were infected, that is, 0.4586 coverage. This substantial increase in the count of the infected targets might be caused by the fact, that for this scenario, all seeds were part of the target group themselves (resulting in on average 25.8 non-seed targets infected, i.e., 0.1985), whereas in the scenario described in Section 4.6, only 5 of the initial seeds were from the target group (resulting in, on average, 47.7 non-seed targets infected, i.e., 0.3669 of the targets).

All in all, the simulation results have shown that the use of a multi-attribute seed selection approach, proposed in this paper, at the cost of reducing the coverage on the studied network by 0.0011, allowed us to increase the coverage within the targeted nodes by 0.1116 compared to the approach oriented on maximizing the global network coverage.

5. Conclusions

Large-scale networks used daily by billions of users [60] create a medium for transmitting information and content. While most influence maximisation methods focus on increasing coverage, it is also important to reach users interested in content or services to avoid the distribution of unwanted messages, decrease information overload and habituation effect and, as a result, increase campaign performance. Earlier research in the area of information spreading focused mainly on influence maximisation. Only limited number of studies discussed targeting nodes with specific characteristics with main focus on their single attributes.

This paper proposes a novel approach to seeding information in multi-attribute social networks, in order to target multi-attribute groups of nodes. In the proposed approach, the seeds for initializing the campaign are chosen based on the ranking obtained with an MCDA method. During information spreading initialization, it is possible to adjust the weights assigned to each attribute. This, in turn, allows to manipulate the symmetry between the global coverage and coverage within the targeted group of nodes. Particularly, the coverage within the targeted multi-attribute nodes' group can be increased, at the cost of potentially reducing the global coverage. The experimental research has shown a superior performance of the proposed approach, compared to traditional approaches focused on the degree centrality measure.

Although the empirical research has shown that the multi-attribute approach to the seed selection allowed us to significantly increase the coverage within the targeted group of nodes, the full-scope study has shown that even higher increase could be obtained if the higher weights were assigned to the criteria which were not initially selected for research in the empirical study. Therefore, grasping this experimental domain knowledge, especially in form of creation of an ontology for selection of criteria for targeting particular types of targets, is a very promising possible future field of research. Such ontology could provide guidelines for the marketer, for assigning weights to the multi-attribute seed rank generation.

Moreover, during the research, finding a multi-attribute model of a real network proved to be very problematic and it was necessary to perform the empirical study on networks with attributes superimposed artificially, based on the known distributions of these attributes in population. This allowed us to study the efficiency of the proposed approach, but comparing to other similar works in this field was not possible. It would be beneficial to include in future work the collection of knowledge about a real multi-attribute social network, in order to allow benchmarking of the proposed approach on a real model. This, in turn, implies additional methodical challenges, as proper reflecting of the non-deterministic nature of performance data in complex networks requires proper adjusting

of the MCDA-based decision models and methods used. In practice, the usage of fuzzy extensions of MCDA methods (which proved to be powerful tools for dealing with data uncertainty) seems to be very promising.

Last, but not least, this research focused only on the multiple values of the network attributes. Future work should include a more profound look into the main aspects of the multi-attributed complex network itself.

Author Contributions: Conceptualization, A.K., J.J. and J.W.; Data curation, A.K. and J.J.; Formal analysis, A.K., J.J. and J.W.; Funding acquisition, J.J. and J.W.; Investigation, J.J.; Methodology, A.K., J.J. and J.W.; Project administration, J.J.; Resources, A.K.; Software, A.K.; Supervision, J.J., J.W.; Validation, J.J.; Visualization, A.K. and J.J.; Writing—original draft, A.K., J.J. and J.W.; Writing—review and editing, J.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Science Centre of Poland, the decision no. 2017/27/B/HS4/01216 (A.K., J.J.) and within the framework of the program of the Minister of Science and Higher Education under the name "Regional Excellence Initiative" in the years 2019–2022, project number 001/RID/2018/19, the amount of financing PLN 10,684,000.00 (J.W.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The final steps of each of the 10 simulations from various scenarios are presented below. The blue "s" vertices represent the seeds. The green "i" nodes represent the non-targeted vertices which were infected. The empty vertices with red outline represent the targets of the campaign. The fully-colored red vertices represent the targets which were successfully reached in the campaign.

Figure A1 presents the target 1, and Figure A2 the target 2 of the real-life usage example from Section 4.1. Subsequently, Figures A3–A5 present scenarios on the synthetic network simulations from Sections 4.4–4.6 respectively.

Figure A1. Visual representation of the real-life usage example—target 1.

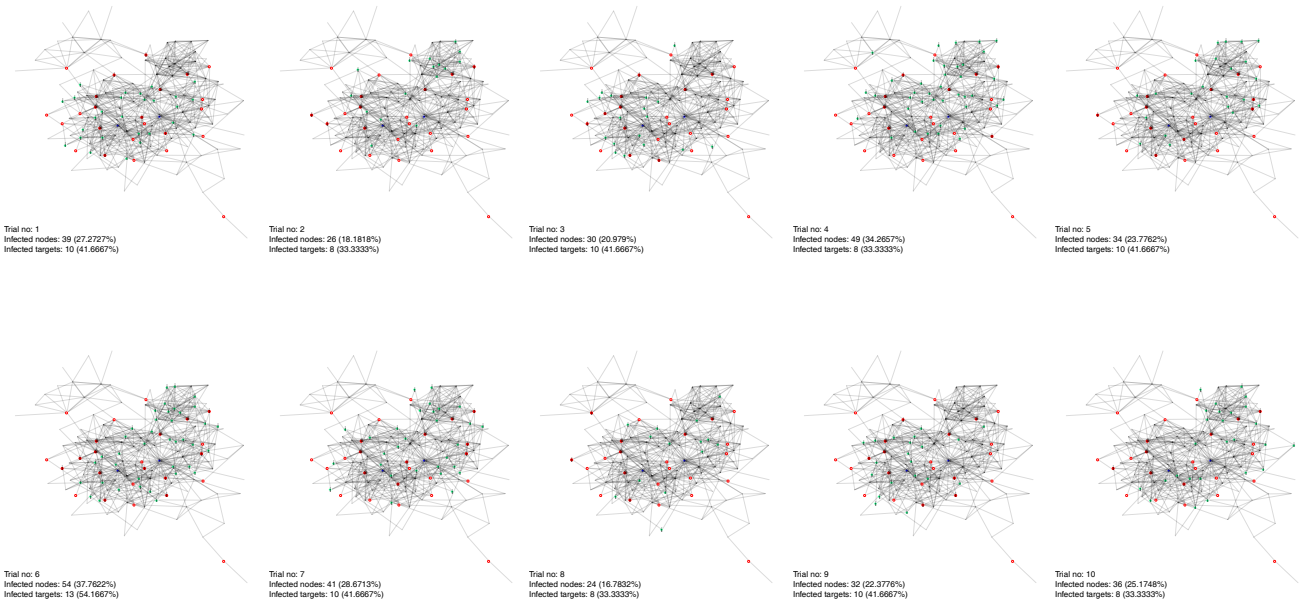


Figure A2. Visual representation of the real-life usage example—target 2.

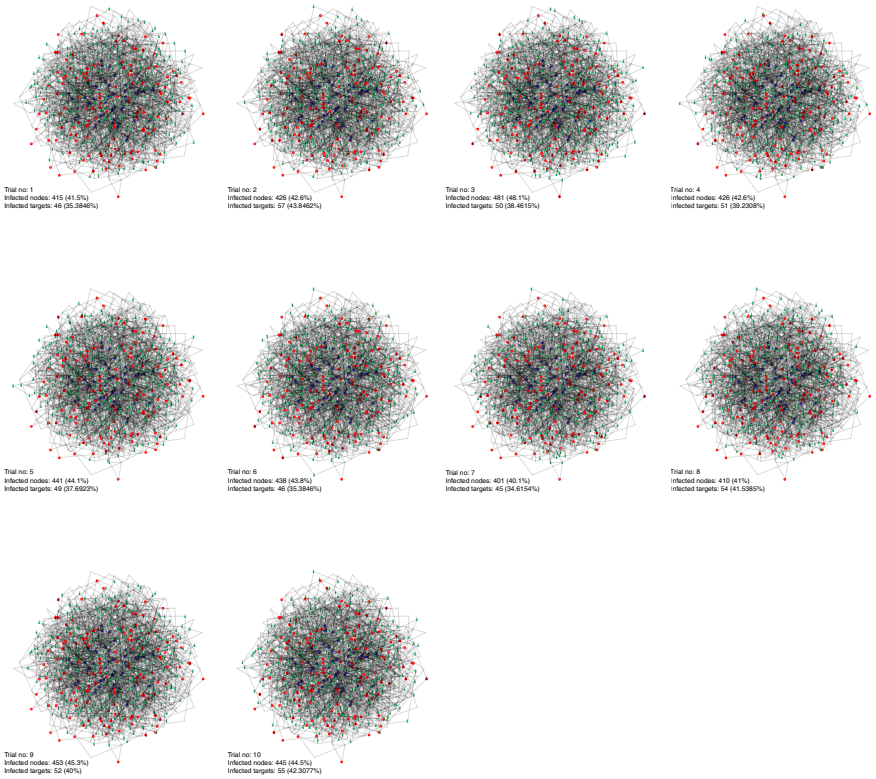


Figure A3. Visual representation of 10 trials for Scenario 1.

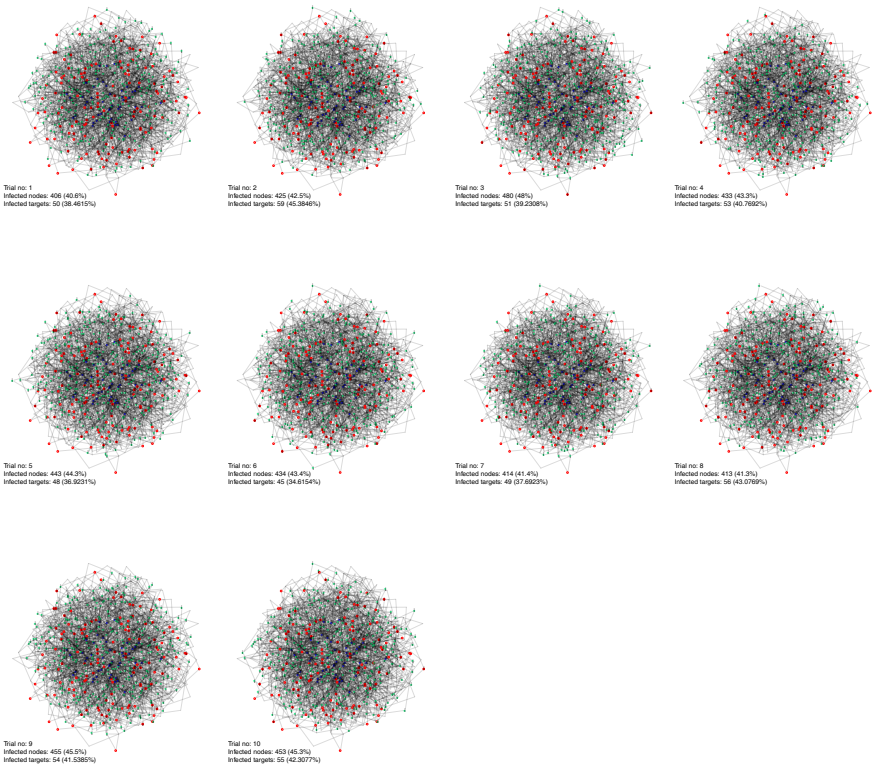


Figure A4. Visual representation of 10 trials for Scenario 2.

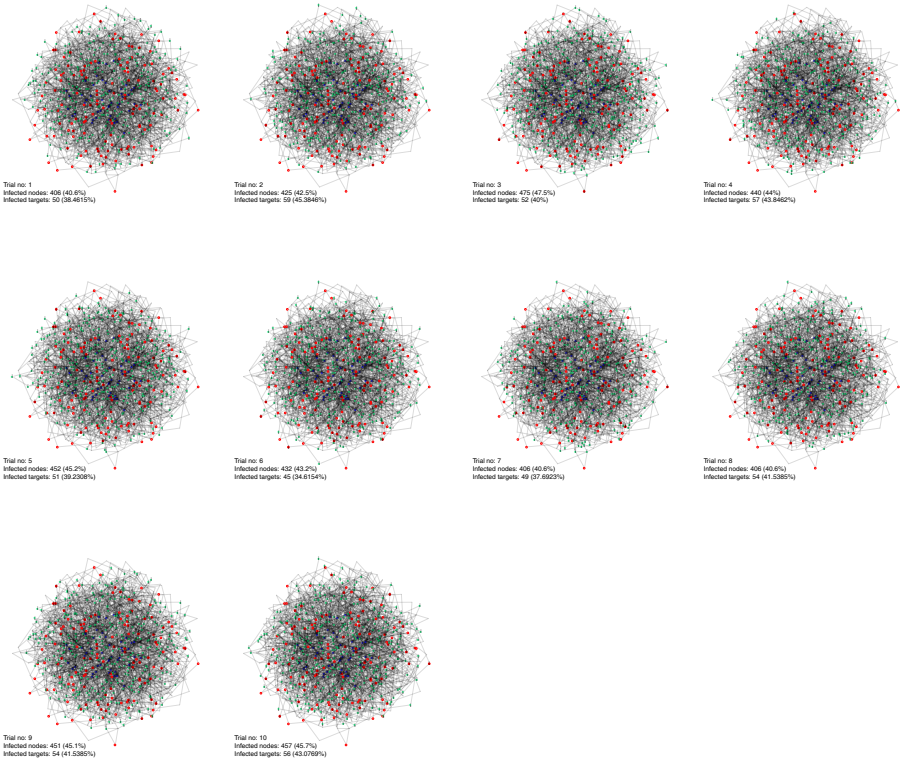


Figure A5. Visual representation of 10 trials for Scenario 3.

References

1. Dunbar, R.I. Do online social media cut through the constraints that limit the size of offline social networks? *R. Soc. Open Sci.* **2016**, *3*, 150292. [CrossRef] [PubMed]
2. Vinerean, S. Importance of strategic social media marketing. *Expert J. Mark.* **2017**, *5*. Available online: http://hdl.handle.net/11159/1381 (accessed on 27 June 2020).
3. Iribarren, J.L.; Moro, E. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Phys. Rev. Lett.* **2009**, *103*, 038702. [CrossRef] [PubMed]
4. Nguyen, H.T.; Dinh, T.N.; Thai, M.T. Cost-aware targeted viral marketing in billion-scale networks. In Proceedings of the IEEE INFOCOM 2016—The 35th Annual IEEE International Conference on Computer Communications, San Francisco, CA, USA, 10–15 April 2016; pp. 1–9.
5. Mochalova, A.; Nanopoulos, A. A targeted approach to viral marketing. *Electron. Commer. Res. Appl.* **2014**, *13*, 283–294. [CrossRef]
6. Liu, Q.; Dong, Z.; Liu, C.; Xie, X.; Chen, E.; Xiong, H. Social marketing meets targeted customers: A typical user selection and coverage perspective. In Proceedings of the 2014 IEEE International Conference on Data Mining, Shenzhen, China, 14–17 December 2014; pp. 350–359.
7. Voss, G.; Godfrey, A.; Seiders, K. Do satisfied customers always buy more? The roles of satiation and habituation in customer repurchase. In *Marketing Science Institute Working Paper Series 2010*; Marketing Science Institute: Cambridge, MA, USA, 2010; pp. 10–101.
8. Luo, C.; Lan, Y.; Wang, C.; Ma, L. The Effect of Information Consistency and Information Aggregation on eWOM Readers' Perception of Information Overload. In Proceedings of the PACIS 2013, Jeju Island, Korea, 18–22 June 2013; p. 180.
9. Datta, S.; Majumder, A.; Shrivastava, N. Viral marketing for multiple products. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 118–127.
10. Thakur, N.; Han, C.Y. An approach to analyze the social acceptance of virtual assistants by elderly people. In Proceedings of the 8th International Conference on the Internet of Things, Santa Barbara, CA, USA, 15–18 October 2018; pp. 1–6.
11. Thakur, N.; Han, C.Y. Framework for an intelligent affect aware smart home environment for elderly people. *Int. J. Recent Trends Hum. Comput. Interact. (IJHCI)* **2019**, *9*, 23–43.
12. Pamučar, D.S.; Božanić, D.; Randđelović, A. Multi-criteria decision making: An example of sensitivity analysis. *Serbian J. Manag.* **2017**, *12*, 1–27. [CrossRef]
13. Mukhametzyanov, I.; Pamucar, D. A sensitivity analysis in MCDM problems: A statistical approach. *Decis. Mak. Appl. Manag. Eng.* **2018**, *1*, 51–80. [CrossRef]
14. Kempe, D.; Kleinberg, J.; Tardos, É. Maximizing the spread of influence through a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 24–27 August 2003; pp. 137–146.
15. Hinz, O.; Skiera, B.; Barrot, C.; Becker, J.U. Seeding strategies for viral marketing: An empirical comparison. *J. Mark.* **2011**, *75*, 55–71. [CrossRef]
16. Zareie, A.; Sheikhahmadi, A.; Jalili, M. Identification of influential users in social networks based on users' interest. *Inf. Sci.* **2019**, *493*, 217–231. [CrossRef]
17. Li, X.; Smith, J.D.; Dinh, T.N.; Thai, M.T. Why approximate when you can get the exact? Optimal targeted viral marketing at scale. In Proceedings of the IEEE INFOCOM 2017—IEEE Conference on Computer Communications, Atlanta, GA, USA, 1–4 May 2017; pp. 1–9.
18. Pasumarthi, R.; Narayanam, R.; Ravindran, B. Near optimal strategies for targeted marketing in social networks. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, Paris, France, 10–15 July 2015; pp. 1679–1680.
19. Zareie, A.; Sheikhahmadi, A.; Khamforoosh, K. Influence maximization in social networks based on TOPSIS. *Expert Syst. Appl.* **2018**, *108*, 96–107. [CrossRef]
20. Yang, P.; Liu, X.; Xu, G. A dynamic weighted TOPSIS method for identifying influential nodes in complex networks. *Mod. Phys. Lett. B* **2018**, *32*, 1850216. [CrossRef]
21. Yang, Y.; Yu, L.; Zhou, Z.; Chen, Y.; Kou, T. Node Importance Ranking in Complex Networks Based on Multicriteria Decision Making. *Math. Probl. Eng.* **2019**, *2019*. [CrossRef]
22. Liu, Z.; Jiang, C.; Wang, J.; Yu, H. The node importance in actual complex networks based on a multi-attribute ranking method. *Knowl. Based Syst.* **2015**, *84*, 56–66. [CrossRef]
23. Robles, J.F.; Chica, M.; Cordon, O. Evolutionary Multiobjective Optimization to Target Social Network Influentials in Viral Marketing. *Expert Syst. Appl.* **2020**, 113183. [CrossRef]
24. Wang, L.; Yu, Z.; Xiong, F.; Yang, D.; Pan, S.; Yan, Z. Influence Spread in Geo-Social Networks: A Multiobjective Optimization Perspective. *IEEE Trans. Cybern.* **2019**. [CrossRef]
25. Gandhi, M.; Muruganantham, A. Potential influencers identification using multi-criteria decision making (MCDM) methods. *Procedia Comput. Sci.* **2015**, *57*, 1179–1188. [CrossRef]

26. Muruganantham, A.; Gandhi, G.M. Framework for Social Media Analytics based on Multi-Criteria Decision Making (MCDM) Model. *Multimed. Tools Appl.* **2020**, *79*, 3913–3927. [CrossRef]
27. Montazerolghaem, M. Effective Factors in Network Marketing Success and Ranking Using Multi-criteria Decision Making Techniques. *Int. J. Appl. Optim. Stud.* **2019**, *2*, 73–89.
28. Karczmarczyk, A.; Jankowski, J.; Wątróbski, J. Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. *PLoS ONE* **2018**, *13*, e0209372. [CrossRef] [PubMed]
29. Cinelli, M.; Kadziński, M.; Gonzalez, M.; Słowiński, R. How to Support the Application of Multiple Criteria Decision Analysis? Let Us Start with a Comprehensive Taxonomy. *Omega* **2020**, 102261. [CrossRef] [PubMed]
30. Pamučar, D.S.; Božanić, D.I.; Kurtov, D.V. Fuzzification of the Saaty's scale and a presentation of the hybrid fuzzy AHP-TOPSIS model: An example of the selection of a brigade artillery group firing position in a defensive operation. *Vojnoteh. Glas.* **2016**, *64*, 966–986. [CrossRef]
31. Chatterjee, P.; Stević, Ž. A two-phase fuzzy AHP-fuzzy TOPSIS model for supplier evaluation in manufacturing environment. *Oper. Res. Eng. Sci. Theory Appl.* **2019**, *2*, 72–90. [CrossRef]
32. Wątróbski, J.; Jankowski, J.; Ziemba, P.; Karczmarczyk, A.; Zioło, M. Generalised framework for multi-criteria method selection. *Omega* **2019**, *86*, 107–124. [CrossRef]
33. Stević, Ž.; Tanackov, I.; Vasiljević, M.; Novarlić, B.; Stojić, G. An integrated fuzzy AHP and TOPSIS model for supplier evaluation. *Serbian J. Manag.* **2016**, *11*, 15–27. [CrossRef]
34. Mardani, A.; Jusoh, A.; Zavadskas, E.K. Fuzzy multiple criteria decision-making techniques and applications—Two decades review from 1994 to 2014. *Expert Syst. Appl.* **2015**, *42*, 4126–4148. [CrossRef]
35. Freeman, L.C. Centrality in social networks conceptual clarification. *Soc. Netw.* **1978**, *1*, 215–239. [CrossRef]
36. Bonacich, P. Technique for analyzing overlapping memberships. *Sociol. Methodol.* **1972**, *4*, 176–185. [CrossRef]
37. Valente, T.W.; Coronges, K.; Lakon, C.; Costenbader, E. How correlated are network centrality measures? *Connections* **2008**, *28*, 16.
38. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **1932**, *22*, 55.
39. Saaty, T.L.; Vargas, L.G. The legitimacy of rank reversal. *Omega* **1984**, *12*, 513–516. [CrossRef]
40. Sabidussi, G. The centrality index of a graph. *Psychometrika* **1966**, *31*, 581–603. [CrossRef]
41. Freeman, L.C. A set of measures of centrality based on betweenness. *Sociometry* **1977**, 35–41. [CrossRef]
42. Roy, B.; Vanderpooten, D. The European school of MCDA: Emergence, basic features and current works. *J. Multi Criteria Decis. Anal.* ***1996***, *5*, 22–38. [CrossRef]
43. Roy, B. Paradigms and challenges. In *Multiple Criteria Decision Analysis: State of the Art Surveys*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 3–24.
44. Guitouni, A.; Martel, J.M. Tentative guidelines to help choosing an appropriate MCDA method. *Eur. J. Oper. Res.* **1998**, *109*, 501–521. [CrossRef]
45. Roy, B.; Słowiński, R. Questions guiding the choice of a multicriteria decision aiding method. *EURO J. Decis. Process.* **2013**, *1*, 69–97. [CrossRef]
46. Vansnick, J.C. On the problem of weights in multiple criteria decision making (the noncompensatory approach). *Eur. J. Oper. Res.* **1986**, *24*, 288–294. [CrossRef]
47. Wątróbski, J.; Jankowski, J.; Ziemba, P.; Karczmarczyk, A.; Zioło, M. Generalised framework for multi-criteria method selection: Rule set database and exemplary decision support system implementation blueprints. *Data Brief* **2019**, *22*, 639. [CrossRef] [PubMed]
48. Stević, Ž.; Alihodžić, A.; Božičković, Z.; Vasiljević, M.; Vasiljević, Đ. Application of combined AHP-TOPSIS model for decision making in management. In Proceedings of the 5th International Conference Economics and Management-Based on New Technologies "EMONT", Vrnjačka Banja, Serbia, 18–21 June 2015; pp. 33–40.
49. Behzadian, M.; Otaghsara, S.K.; Yazdani, M.; Ignatius, J. A state-of the-art survey of TOPSIS applications. *Expert Syst. Appl.* **2012**, *39*, 13051–13069. [CrossRef]
50. Chen, S.J.; Hwang, C.L. Fuzzy multiple attribute decision making methods. In *Fuzzy Multiple Attribute Decision Making*; Springer: Berlin/Heidelberg, Germany, 1992; pp. 289–486.
51. Yoon, K.P.; Hwang, C.L. *Multiple Attribute Decision Making: An Introduction*; Sage Publications: Thousand Oaks, CA, USA, 1995; Volume 104.
52. Hwang, C.L.; Yoon, K. Multiple criteria decision making. In *Lecture Notes in Economics and Mathematical Systems*; Springer: Berlin/Heidelberg, Germany, 1981; Volume 186, pp. 58–191.
53. Rossi, R.A.; Ahmed, N.K. The Network Data Repository with Interactive Graph Analytics and Visualization. In Proceedings of the AAAI 2015, Austin, TX, USA, 25–30 January 2015.
54. Gus. *Ludność. Stan i Struktura Ludności Oraz Ruch Naturalny w Przekroju Terytorialnym*; Statistics Poland: Warsaw, Poland, 2016.
55. Barabási, A.L.; Bonabeau, E. Scale-free networks. *Sci. Am.* **2003**, *288*, 60–69. [CrossRef] [PubMed]
56. Chiasserini, C.F.; Garetto, M.; Leonardi, E. Social network de-anonymization under scale-free user relations. *IEEE ACM Trans. Netw.* **2016**, *24*, 3756–3769. [CrossRef]
57. Luo, Y.; Ma, J. The influence of positive news on rumor spreading in social networks with scale-free characteristics. *Int. J. Mod. Phys. C* **2018**, *29*, 1850078. [CrossRef]

58. Liu, W.; Li, T.; Liu, X.; Xu, H. Spreading dynamics of a word-of-mouth model on scale-free networks. *IEEE Access* **2018**, *6*, 65563–65572. [CrossRef]
59. Ministerstwo Zdrowia. *Program Profilaktyki Raka Piersi (Mammografia)*; www.gov.pl; Ministerstwo Zdrowia: Warsaw, Poland, 2018.
60. WeAreSocial Digital 2020. Available online: https://wearesocial.com/digital-2020 (accessed on 27 June 2020).

A8.

Karczmarczyk, A., Jankowski, J., Wątróbski, J. (2021). Multi-Criteria Seed Selection for Targeted Influence Maximization within Social Networks – in proceedings of International Conference on Computational Science: ICCS 2021

Temat: ICCS2021 notfication for paper 173
Nadawca: "Computational Collective Intelligence" <cci_2021@easychair.org>
Data: 16.03.2021, 16:51
Adresat: Artur Karczmarczyk<artur@piruety.net>

Dear Artur Karczmarczyk,

Thank you for your submission to the International Conference on Computational Science (ICCS).
Please read below for the decision on your paper.

Due to the current global health conditions, ICCS 2021 will be held in a fully virtual format.
(further news and details will be announced in the coming weeks)

As always, submission to ICCS was very competitive, with over 650 submissions.
The program committee and conference chairs had to carefully select papers of the highest quality and relevance to the field of Computational Science.

We are pleased to inform you that your paper (see below) has been accepted for publication as a 7-page short paper in Springer's LNCS Series and oral presentation (20 minutes incl. questions) at the conference. Please take note to revise your paper to the accepted length.

173 Multi-Criteria Seed Selection for Targeted In uence Maximization within Social Networks

Acceptance is still conditional on the following:

1. That you revise your paper taking into account the reviewers' comments, strictly following the guidelines for camera-ready submission for 7-page papers, and submit your final version before the deadline of 5 April 2021.
Detailed instructions on how to submit your camera-ready paper will be sent in a few days, in a separate e-mail.

2. That you register for the conference before the author registration deadline (5 April 2021):
https://www.iccs-meeting.org/iccs2021/registration-accommodation/

Please note that the registration deadline is strict. Only the papers registered until this date will be included in the proceedings.

We wish you good health in these troubled times.

With best regards,
The ICCS Organizing Committee

Multi-Criteria Seed Selection for Targeted Influence Maximization within Social Networks

Artur Karczmarczyk[1][0000−0002−1135−7510], Jarosław Jankowski[1][0000−0002−3658−3039], and Jarosław Watrobski[2][0000−0002−4415−9414]

[1] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin
{artur.karczmarczyk,jaroslaw.jankowski}@zut.edu.pl
[2] The Faculty of Economics, Finance and Management of the University of Szczecin, Mickewicza 64, 71-101 Szczecin, Poland jwatrobski@usz.edu.pl

Abstract. Information spreading and influence maximization in social networks attracts attention from researchers from various disciplines. Majority of the existing studies focus on maximizing global coverage in the social network through initial seeds selection. In reality, networks are heterogeneous and different nodes can be a goal depending on campaign objectives. In this paper a novel approach with multi-attribute targeted influence maximization is proposed. The approach uses the multi-attribute nature of the network nodes (age, gender etc.) to better target specified groups of users. The proposed approach is verified on a real network and compared to the classic approaches delivers 7.14% coverage increase.

Keywords: seed selection · targeted influence maximization · MCDA.

1 Introduction

Social media are used for maintaining connections with relatives, friends and to access information sources. Virtual marketing within social media is strategized to reach people with specific interests. It results in a better engagement of the potential client thereof [4] and makes possible to avoid targeting users not interested in products or services. While most of the research focused on influence maximization and global coverage, social networking platforms deliver the ability to pick multiple choice parameters for an exact target class. The need to better address the real specifics of campaigns is visible, but the targeted approaches are introduced in a limited number of studies and are focused mainly on single node attributes [7] [15].

The approach presented in this paper deals with the selection of nodes for seeding the social platform on the basis of manifold criteria, as well as diverse attributes within agent based computational environment. The MCDA foundations of the proposed approach enable to adjust the gravity of each touchstone to be computed for selection purpose, in order to meet the requirements of the advertiser. Moreover, the relevant MCDA tools and computations enable to gauge

the impact of nodes seeding individually on the viral marketing strategy to hit the target groups. The paper comprises of five sections. The Introduction is followed by the Literature review section 2. Next, the methodology discussion is presented 3. After that, experimental results are showed 4 and followed by concluding statements 5.

2 Literature Review

In the area of information spreading within social platforms, it was supposed in the early stages of research that all the nodes of a network carry the same level of inclination towards a promulgated product or service or any other content [6]. However, in reality more result-oriented campaigns allow multiple node behaviors to be taken into consideration and better nodes allocation [7]. Recent studies used the cost assignment to the user of the network combined with the user interest benefits [8]. The goal of nodes selection can be also avoidance of intense campaign with unnecessarily repeated messages [1]. Pasumarthi et al. identified a targeted influence maximization problem, introducing an objective functionality and a penalizing criterion for adopting non targeted nodes [9].

Recently, initial studies are held discussing the application of MCDA techniques in the areas related to social networking. TOPSIS [3] method is used by Yang et al., in SIR (Susceptible Infected Recovered) model for identification of influential nodes in complex network [13]. Entropy weight method is used to measure and set up the weight values [14]. For maximizing the coverage and reducing the overlap, TOPSIS method is used by Zarei et al., while a social network is being influenced [16]. PROMETHEE [4] method was used by Karczmarczyk et al., to evaluate the responsiveness of viral marketing campaigns within social networking portals and also for providing decision support in order to plan these campaigns [5].

Review of studies in the area of information spreading and influence maximization has shown that among large number of studies only a small chunk is targeting the most common problem such as reaching out the specific user with multiple characteristics. Most of the existing approaches behave mono-trait by addressing nodes as a single attribute. However, social networks generally identify the target groups relying on multiple parameters, such as gender, localization or age. This identifies a research gap for seed selection based on a multi-characteristic computation in order to target specific multi-attribute network nodes, which this paper addresses.

3 Methodology

The proposed methodology complements the widely-used Independent Cascade (IC) model for modeling the spread within the complex networks [6], by taking

[3] Technique for Order Preference by Similarity to Ideal Solution
[4] Preferences Ranking Organization METHod for Enrichment of Evaluations

into account the problem of reaching targeted multi-attribute nodes in social networks by the information propagation processes. In the proposed approach, it is assumed that the network nodes are characterized not only by the centrality relations between them and other nodes, but also by a set of custom attributes $C_1, C_2, \ldots, C_n$. The nodes can also be characterized by the computed attributes derived from the network characteristics and measures, such as degree. Last, but not least, additional attributes can be derived as a composite of the two aforementioned types of attributes, by computing centrality measures based on limited subsets of the nodes' neighbors. For example, if attribute C_i represented the degree of a node, i.e. the total count of its neighbors, the C_{i_1} could represent the count of its male neighbors.

The aim of the proposed methodological framework is to maximize the influence within the targeted group of multi-attribute network nodes. While other approaches focus on generating the ranking of seeds based on a single centrality measure, in the authors' proposed methodological framework, the seeds are selected based on multiple attributes. This allows to select seeds which might be worse at maximizing global influence in the network, but which are better at maximizing influence in the targeted group of multi-attribute network nodes.

The approach presented in this paper is based on the MCDA methodology foundations [11]. The assumed modeling goal is to reach only the targeted set of multi-attribute nodes, instead of maximixing global influence in the network. Based on the guidelines provided by [3], it was decided that the PROMETHEE II method is most suitable for the proposed approach. It is an MCDA method that uses pairwise comparison and outranking flows to produce a ranking of the best decision variants. In the proposed approach, PROMETHEE II is used to produce a multi-criteria ranking of the nodes in the network with the aim to shortlist the ones which have the best chances to maximize influence in the targeted group of multi-attribute nodes. A detailed description of the PROMETHEE methods can be found in [2]. The MCDA foundations of the proposed approach help maximizing influence in the targeted group of multi-attribute nodes by selecting the seeds which have the highest, according to the marketer, potential to reach the targeted nodes in the social network. Moreover, the use of tools such as GAIA visual aid allows to understand the preferences backing the actual seed selection, and provide feedback which allows to further iteratively improve the obtained solution.

4 Empirical Study

In order to illustrate the proposed approach, the empirical study with the use of agent based simulations was performed on a relatively small real network [10] with 143 vertices and 623 edges giving the ability of detailed multi-criteria analysis. The proposed approach is intended for networks whose nodes are described with multiple attributes. However, the publicly available network datasets predominantly consist only of information on their nodes and edges, without information on the node attributes. To overcome this problem, the node attributes

Table 1. Criteria used in the empirical research.

Criterion	Values	Criterion	Values
C1 degree	integer [1-42]	C5 age	1: 0-29, 2: 30-59, 3: over 60
C2 gender	1: male, 2: female	C6 deg. younger	integer [0-18]
C3 deg. male	integer [0-20]	C7 deg. medium	integer [0-15]
C4 deg. female	integer [0-22]	C8 deg. older	integer [0-9]

Table 2. Top 7 network nodes used as seeds in the empirical research. A - degree; B - betweenness, C - closeness, D - eigen centrality, E-G - the proposed multi-attribute approach

A	105	17	95	48	132	43	91	E	17	95	48	132	50	105	20
B	107	17	48	91	32	95	141	F	19	95	48	50	132	91	105
C	105	17	95	37	74	48	91	G	132	20	136	19	50	122	3
D	105	31	136	132	20	19	69								

were artificially overlaid over the network, following the attributes' distribution from demographic data. Two demographic attributes were overlaid on the network – gender and age. For illustrative purposes, the target for the viral marketing campaign was chosen for the empirical research. In this experiment, the male users from the youngest age group were targeted, which translates to 28 of all the 143 users of the network. In the proposed approach, the seeds are selected from the network based on multiple criteria. In the empirical research, apart from the two aforementioned demographic attributes, also the degree measure was taken into account, as well as 5 criteria based on a mix of the degree and the demographic measures. This resulted in a total of 8 seed evaluation criteria, which are presented in Table 1.

Initially, the classic single-metric approaches were tested on the network, to provide a benchmark for the proposed approach. Four centrality metrics (degree, closeness, betweenness and eigen centrality) were used individually to first rank all vertices in the network, and then select the top nodes as seeds. It was decided for the seeding fraction to be set to 0.05 (seven seeding nodes) and propagation probability to 0.10. Moreover, in order to allow repeatability of the experiment for seeds selected by each approach, 10 pre-defined scenarios were created, in which each node was assigned a pre-drawn weight. The seeds obtained from rankings based on each centrality measure, i.e. degree, betweenness, closeness and eigen centrality, are presented in Table 2A - 2D respectively. The averaged simulation results are presented in Table 3A - 3D.

In the next step of the empirical study, the authors' proposed approach was used to choose the seeds based on a multi-criteria ranking produced by the PROMETHEE II method. All eight criteria were taken into account. Initially, the usual preference function was used for comparing each vertex under all criteria. Also, all criteria were given an equal preference weight (see Table 4E). As a result, seven seeds were selected (see Table 2E). It can be noticed that the

Table 3. Aggregated results from the empirical study simulations

	Iterations	Infected	Coverage	Infected targeted	Coverage
A	6.6	41.2	0.2881	7.7	0.2750
B	6.1	33.7	0.2357	5.5	0.1964
C	6.2	39.2	0.2741	6.2	0.2214
D	6.5	34.3	0.2399	9.0	0.3214
E	5.9	40.6	0.2839	9.2	0.3286
F	6.0	40.7	0.2846	9.5	0.3393
G	6.4	30.1	0.2105	9.7	0.3464

Table 4. Utilized PROMETHEE II parameters

	Criteria	C1	C2	C3	C4	C5	C6	C7	C8
E	Weight	1	1	1	1	1	1	1	1
	Preference function	Usual	Usual	Usual	Usual	Usual	Usual	Usual	Usual
F	Weight	1	1	1	1	1	1	1	1
	Preference function	Linear	Usual	Linear	Linear	Usual	Linear	Linear	Linear
	q; p	3; 9	1; 2	1; 4	1; 2	1; 2	1; 4	1; 3	1; 2
G	Weight	8.2	25.4	12.6	3.8	28.4	14	3.8	3.8
	Preference function	Linear	Usual	Linear	Linear	Usual	Linear	Linear	Linear
	q; p	3; 9	1; 2	1; 4	1; 2	1; 2	1; 4	1; 3	1; 2

produced seed set is considerably different than the ones produced by the classic approaches (compare with Table 2A-2D).

After the simulations were executed with the newly selected seeds, it was observed that averagely 40.6 network nodes were infected (0.2839 coverage, see Table 3E). It is a worse result than for the degree-based approach. What is important to note, however, is that averagely 9.2 targeted nodes were infected, i.e. 0.3286 targeted coverage, which was the best result so far.

One of the benefits of using the PROMETHEE methods is the possibility to adjust the preference function used in pairwise comparisons of the nodes under individual criteria. While the usual preference function provides a simple boolean answer for the pairwise comparison of gender (C2) and age (C5) criteria, in case of the criteria based on degree, usage of a linear preference function with indifference and preference thresholds can yield better results. Therefore, in the subsequent step of the empirical research, a linear preference function was applied to all degree-based criteria (see Table 4F). The change in the preference function resulted in a different set of seeds selected for simulations (see Table 2F). The averaged results from the simulations are presented in Table 3F. It can be observed, that both global and targeted coverage values improved slightly.

Depending on the target group, the marketer can decide that some criteria can better help to reach the target group than the other criteria. Therefore, the marketer can adjust the preference weights of each criterion. Before the last set of simulations in this empirical research, an expert knowledge was elicited from the marketer with the use of the Analytical Hierarchy Process (AHP) [12], to adjust the preference weights of all criteria. The elicited weights used in the final set of

Fig. 1. GAIA Visual Analysis

simulations is presented in Table 4G. The adjusted preference weights resulted in a significantly different set of nodes used as seeds in the campaign (see Table 2G). The averaged simulation results are presented in Table 3G. The approach resulted in the best coverage in the targeted group (0.3464, compared to 0.2750 for the degree-based approach, 0.0714 difference). In the final step of the research, the GAIA visual analysis aid was used to study the criteria preference relations in the seed selection decision model (see Fig. 1). The analysis of Fig. 1 allows to observe that criteria C2 and C5 are not related to each other in terms of preference. This is quite straightforward, because these criteria represent the gender and age respectively. On the other hand, the remaining criteria are similar in terms of preference, possibly because they are all partially based on the degree measure.

5 Conclusions

The existing research in the area of information spreading focuses mainly on influence maximisation. Only limited number of studies discuss targeting nodes with specific characteristics with main focus on their single attributes. This paper proposes a novel approach to multi-attribute targeted influence maximization in social networks, focused on a multi-attribute seed selection. In the proposed approach, the seeds for initializing the campaign are chosen based on a ranking obtained with an MCDA method. The weights of individual criteria can be adjusted, as well as criteria values' comparison preference functions can be chosen to best fit the marketer's needs. In the experimental research, the proposed approach resulted in target nodes' coverage superior by as much as 7.14% compared to traditional degree-based approaches. The research opens some possible future directions. It would be beneficial to further broaden the research scope by studying how the changes in seeding fraction and propagation probability affect the efficiency of the proposed approach. Moreover, this research was performed on a network with attributes superimposed artificially. A research project can be run in order to collect knowledge about a real multi-attribute social network.

6 Acknowledgments

This work was supported by the National Science Centre of Poland, the decision no. 2017/27/B/HS4/01216 (AK, JJ) and within the framework of the program of the Minister of Science and Higher Education under the name "Regional Excellence Initiative" in the years 2019-2022, project number 001/RID/2018/19, the amount of financing PLN 10,684,000.00 (JW).

References

1. Abebe, R., Adamic, L., Kleinberg, J.: Mitigating overexposure in viral marketing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
2. Brans, J.P., Mareschal, B.: Promethee methods. In: Multiple criteria decision analysis: state of the art surveys, pp. 163–186. Springer (2005)
3. Cinelli, M., Kadziński, M., Gonzalez, M., Słowiński, R.: How to support the application of multiple criteria decision analysis? let us start with a comprehensive taxonomy. Omega p. 102261 (2020)
4. Iribarren, J.L., Moro, E.: Impact of human activity patterns on the dynamics of information diffusion. Phys. Rev. Lett. **103**, 038702 (Jul 2009)
5. Karczmarczyk, A., Jankowski, J., Watróbski, J.: Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PloS one **13**(12) (2018)
6. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 137–146 (2003)
7. Mochalova, A., Nanopoulos, A.: A targeted approach to viral marketing. Electronic Commerce Research and Applications **13**(4), 283–294 (2014)
8. Nguyen, H.T., Dinh, T.N., Thai, M.T.: Cost-aware targeted viral marketing in billion-scale networks. In: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications. pp. 1–9. IEEE (2016)
9. Pasumarthi, R., Narayanam, R., Ravindran, B.: Near optimal strategies for targeted marketing in social networks. In: Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems. pp. 1679–1680 (2015)
10. Rossi, R.A., Ahmed, N.K.: The network data repository with interactive graph analytics and visualization. In: AAAI (2015), http://networkrepository.com/email-enron-only.php
11. Roy, B., Vanderpooten, D.: The european school of MCDA: Emergence, basic features and current works **5**(1), 22–38
12. Saaty, T.L.: Decision-making with the ahp: Why is the principal eigenvector necessary. European journal of operational research **145**(1), 85–91 (2003)
13. Yang, P., Liu, X., Xu, G.: A dynamic weighted topsis method for identifying influential nodes in complex networks. Modern Physics Letters B **32**(19), 1850216 (2018)
14. Yang, Y., Yu, L., Zhou, Z., Chen, Y., Kou, T.: Node importance ranking in complex networks based on multicriteria decision making. Mathematical Problems in Engineering **2019** (2019)
15. Zareie, A., Sheikhahmadi, A., Jalili, M.: Identification of influential users in social networks based on users' interest. Information Sciences **493**, 217–231 (2019)
16. Zareie, A., Sheikhahmadi, A., Khamforoosh, K.: Influence maximization in social networks based on topsis. Expert Systems with Applications **108**, 96–107 (2018)

A9.

Karczmarczyk, A., Wątróbski, J., Jankowski, J. (2021). Seeding for Complementary Campaign Objectives in Social Networks - in proceedings of The Americas Conference on Information Systems: AMCIS 2021

Temat: Decision on submission 1235 of AMCIS 2021
Nadawca: "Guy Pare, Wynne Chin, and Benoit Aubert" <amcis21c@precisionconference.com>
Data: 14.04.2021, 21:05
Adresat: Artur Karczmarczyk <artur.karczmarczyk@zut.edu.pl>

Subject: AMCIS 2021 Submission Decision, Manuscript 1235

ID: 1235
Title: Seeding for Complementary Campaign Objectives in Social Networks
Contact Author: Artur Karczmarczyk

Dear Artur Karczmarczyk,

It is a pleasure to conditionally accept your manuscript entitled "Seeding for Complementary Campaign Objectives in Social Networks" for inclusion at AMCIS 2021. Please review the comments from those who reviewed your paper at the bottom of this message and make revisions accordingly.

Please submit your final, camera-ready version of the paper by April 23, 2021 at 11:59 PM Eastern time. Instructions for preparing your manuscript in camera-ready form, can be found on the AMCIS 2021 website: https://amcis2021.aisconferences.org/ Please be sure to use the CAMERA READY TEMPLATE for your submission type, which can be found here. https://amcis2021.aisconferences.org/submissions/types-of-submissions/ .

**** IMPORTANT: Papers that do not comply with the template may still be rejected. ****

Please be sure to follow the steps below:

- Make as many of the requested revisions as possible before April 23, 2021.
- Download the CAMERA READY submission template for your submission type.
- Include all author information, abstract, and keywords per the template.
- Upload your revised manuscript through your Author Dashboard in our conference submission site: https://new.precisionconference.com/ais
- The deadline for camera-ready submissions is April 23, 2021 at 11:59 PM Eastern time.

Additional notes for completed research:
- Completed research papers are limited to 10 pages.

Additional notes for Emergent Research Forum papers:
- Emergent Research Papers are limited to 5 pages.
- You may choose to either include (1) the revised ERF paper for publication in the proceedings or (2) only the abstract for publication in the proceedings.

Please note:
1) For all papers accepted into AMCIS 2021, authors of accepted papers will retain copyright. However, by submitting a paper, authors do agree that AIS can publish and reproduce any accepted papers in the AMCIS 2021 proceedings or through other AIS communications' vehicles (i.e. www.aisnet.org) in the format of AIS' choosing (CD, USB, eLibrary, other electronic reproduction and printed proceedings) under an established ISBN number for AMCIS 2021.
2) AIS has a strict registration policy for accepted papers at AMCIS. Failure to follow this policy may result in your paper being pulled from the conference program or proceedings.
3) With AMCIS 2021 being held virtually, instructions for presenting completed and ERF papers will be forthcoming.

Once again, thank you for submitting your manuscript to AMCIS 2021 and we look forward to receiving your camera-ready version of your paper.

Seeding for Complementary Campaign Objectives in Social Networks

Emergent Research Forum (ERF)

Artur Karczmarczyk
Westpomeranian University of Technology in Szczecin, Poland
artur.karczmarczyk@zut.edu.pl

Jarosław Wątróbski
University of Szczecin, Poland
jaroslaw.watrobski@usz.edu.pl

Jarosław Jankowski
Westpomeranian University of Technology in Szczecin, Poland
jjankowski@wi.zut.edu.pl

Abstract

Various theoretical models are used in research into the dissemination of information in social networks. The assumed goals include selecting seeds in order to maximize the influence or reach the target subset of the network users (nodes). On the other hand, the multi-attribute nature of individual network nodes indicates the possibility of analyzing their multi-criteria nature, as well as the consequent use of MCDA methods in the process of seed selection. The state of art shows, however, that this contribution is still missing. This paper presents an attempt to use the multi-criteria decision analysis (MCDA) in the seed selection process and thus a methodological framework supporting reaching many varied sets of target nodes. As a result, a single viral marketing campaign in a social network can be performed to target multiple (often separate) sets of target nodes, thus fulfilling the objectives normally achieved with multiple campaigns.

Keywords

influence maximization, social networks, viral marketing, TOPSIS, AHP, MCDA, independent cascade model

Introduction

Viral marketing campaigns in social networks are based on the concept that users to whom a product is advertised will further spread the information on the product within the network. The research in the network area is interdisciplinary and attracts sociologists, physicists, computer scientists and marketers with a wide range of approaches and research goals. The independent cascades (IC) model is often used to simulate campaigns in networks, based on varying count of nodes to which the information would be advertised (seeding fraction, SF) and on varying propagation probability (PP) – the probability that the users would pass information to their neighbors (Kempe et al. 2003). Due to the hardships in obtaining detailed mappings of real networks, theoretical models are often used in research, which can easily be adjusted to accommodate for various studied phenomena. These models include Barabasi-Albert (BA), Watts-Strogatz (WS), and Erdos-Renyi (ER) networks (Barabási and Bonabeau 2003).

Whilst majority of the existing research focus on selecting seeds for maximizing the influence globally in the network (maximizing coverage), some researchers focus on adjusting the node selection methods for reaching a targeted subset of the network users. The prior research, however, assumes the network nodes to be marked by a single flag or two cost & benefit criteria (Mochalova and Nanopoulos 2014; Nguyen et al. 2016). Nonetheless, in real-life campaigns, the ordering party is able to choose from a wide set of attributes describing the users to which the product will be advertised, such as age, gender, localization. Whilst creating the rankings of nodes based on a single centrality measure such as degree is simple, when multiple user attributes need to be considered at once – the selection process is more complex. Multi-criteria decision analysis (MCDA) methods can be used to facilitate the selection process. The TOPSIS[1] method, due to its efficiency and automation capability, can be used to aggregate the criterial performance of all multi-

[1] Technique for Order of Preference by Similarity to Ideal Solution (Hwang et al. 1993)

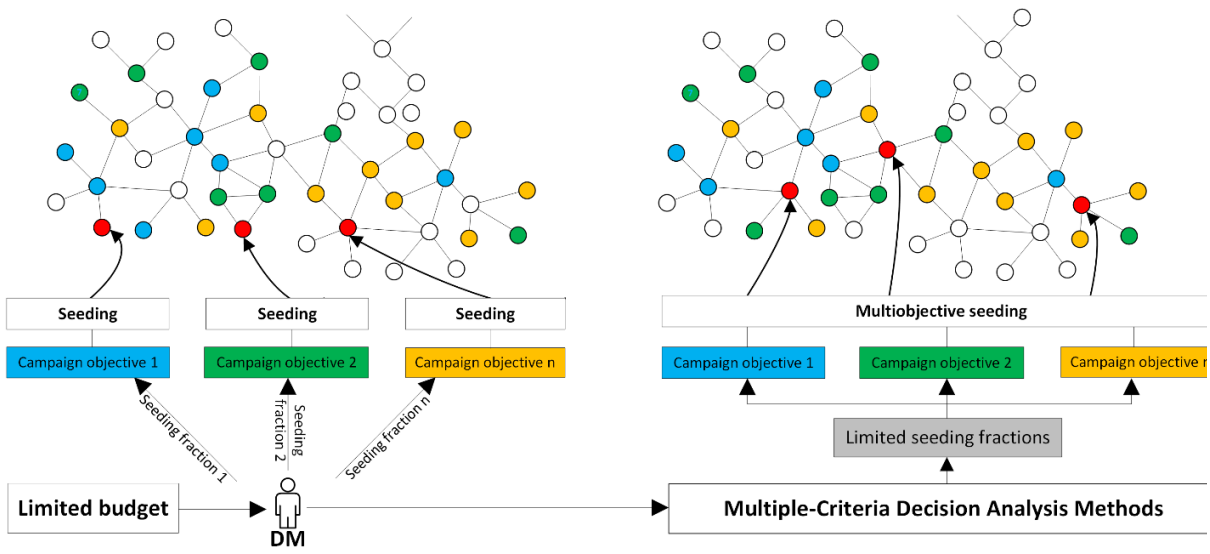
attribute nodes into a single ranking. The AHP[2] method, on the other hand, can be used to analyze the preferences of the ordering party, which can help selecting seeds allowing to obtain the campaigns' objectives.

The authors' main contribution in this study is to provide a sustainable methodological framework, in which a single viral marketing campaign in social network can be executed to reach multiple (often separate) sets of targeted nodes, thus fulfilling objectives normally achieved with multiple campaigns. This effort to aim multiple campaigns' objectives with a single campaign constitutes the research gap between the authors' previous research and this study.

Methodology

When a viral advertising campaign in social network is executed, the ordering party can choose from a vast set of attributes, such as age and gender, characterizing the users who will be displayed the advertised information. In the research it was assumed that the network nodes are described with multiple attributes. These attributes can store information such as mentioned above, i.e. age, gender, location etc. On the other hand, the attributes can include centrality measures such as degree, closeness etc. Also, some additional artificial attributes can be created as a cartesian product of the former and latter.

Fig. 1. Methodological framework of the proposed approach



In the proposed approach (see Fig. 1), it is assumed that the party ordering viral marketing campaigns tries to achieve n objectives, i.e. wants to maximize influence within n sets of targeted nodes. Traditionally, n campaigns could be executed, one for each target group. In the proposed approach, a limited budget, and therefore, a limited set of network users to seed the information to is assumed. Therefore, a sustainable approach is proposed, in which the decision maker (DM) uses an MCDA method and based on the campaign objectives $1, 2, \ldots, n$, expert judgment and the analyst's experience, selects a limited fraction of network nodes to initialize a single campaign in order to reach nodes matching the ordering party's multiple complementary objectives. Due to the sustainable nature of the proposed approach, it is assumed that the global coverage in the network might be reduced, but it is aimed to increase the coverage in the targeted network nodes.

Results and Discussion

The empirical research was initiated on a real network comprising of 143 nodes and 623 edges, with average degree 8.7133, betweenness 139.6573, closeness 0.0024 and eigen centrality 0.2426 (Ryan A. Rossi, Nesreen K. Ahmed 2015). Artificial values of two attributes were generated for the network, based on

[2] Analytic Hierarchy Process (Saaty 1988)

demographic structure data: gender (69 nodes male, and 74 nodes female), and age (0-29 years – 62 nodes, 30-59 years – 55 nodes, over 60 years – 26 nodes).

In order to demonstrate the proposed approach, two objectives were defined:

- O^{M}_{0-29} male users, aged 0-29;
- O^{F}_{30-59} female users, aged 30-29.

Initially, the seeding fraction of 0.05 was assumed, resulting in 7 nodes to provide the initial information to propagate over the network. Six scenarios were then simulated on the network, based on the independent cascade (IC) model. For each scenario, 10 simulations were performed on pre-drawn network weights, in order to ensure the same conditions for all scenarios. The propagation probability was set to 0.1.

The studied scenarios are presented in Table 1. The first three scenarios are based on selecting as seeds the nodes with the highest degree, regardless of the objectives of the campaign. The first and second scenario are representing separate single-objective campaigns. The third scenario represents a single campaign with two objectives, yet initialized with the same seeds as the two previous scenarios.

Scenario	Objectives	Seed Selection Strategy	Weights for TOPSIS							
			C1	C2	C3	C4	C5	C6	C7	C8
S°_{M}	O^{M}_{0-29}	Nodes with the highest degree	100							
S°_{F}	O^{F}_{30-59}		100							
S°_{MF}	$O^{M}_{0-29}, O^{F}_{30-59}$		100							
S^{*}_{M}	O^{M}_{0-29}	Ranking based on multiple attributes, obtained with the use of a MCDA method	8.2	25.4	12.6	3.8	28.4	14	3.8	3.8
S^{*}_{F}	O^{F}_{30-59}		4.4	30.4	4	10.4	30.4	5.4	10.4	4.4
S^{*}_{MF}	$O^{M}_{0-29}, O^{F}_{30-59}$		5.4	1.4	4.2	4.3	32.3	24.9	24.9	2.8

Table 1. Description of the six scenarios studied in the research.

On the other hand, in the subsequent three scenarios, the TOPSIS method was used each time before running the campaign simulation, to provide a ranking of nodes based on multiple criteria. The criteria used were as follows:

- C1 – degree of the node,
- C2 – gender of the node (male / female),
- C3 – degree male – the count of male neighbors of the node,
- C4 – degree female – the count of female neighbors of the node,
- C5 – age (0-29, 30-59, 60+),
- C6 – degree young – the count of neighbors of the node aged 0-29,
- C7 – degree mid-aged – the count of neighbors of the node aged 30-59,
- C8 – degree elderly – the count of neighbors of the node aged 60+.

The weights for the TOPSIS method were obtained separately for each scenario, with the use of expert judgment and the AHP method(Saaty 1988). The weights are presented in Table 1. After the rankings of nodes for each scenario were obtained, the simulations were performed.

The averaged results from all ten simulations for each scenario are presented in Table 2. Its analysis allows to observe that the use of the proposed approach resulted in slightly lower global coverage (-0.0210 decrease), however, the coverage in the targeted nodes' group increased by 0.0174.

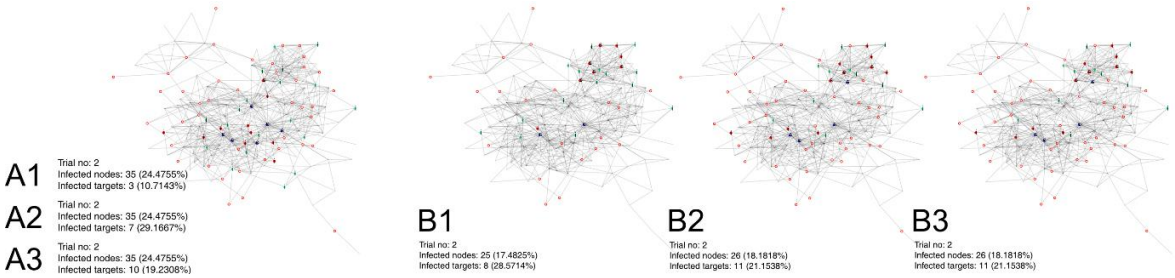
It might be beneficial to compare the approaches in detail. For this purpose, Fig 2. presents the results for a single simulation for all 6 scenarios. It can be observed that for A1-A3 the same nodes are selected as seeds, regardless of which nodes are targeted. On the other hand, for B1-B3 it can be observed, that depending on the campaign objectives, different set of seeds is selected, which results in different nodes being reached. Moreover, it can be observed that for B1-B3 many of the nodes selected as seeds are already a part of the target group.

Scenario	Average Iterations Count	Average Coverage	Decrease	Number of targeted nodes	Average coverage in targeted nodes	Increase

S_M°	6.6	0.2881	x	28	0.2750	x
S_F°	6.6	0.2881	x	24	0.2833	x
S_{MF}°	6.6	0.2881	x	52	0.2788	x
S_M^*	6.3	0.2476	-0.0405	28	0.3357	0.0607
S_F^*	6.2	0.2552	-0.0329	24	0.3958	0.1125
S_{MF}^*	5.9	0.2671	-0.0210	52	0.2962	0.0174

Table 2. Comparison of the averaged results on the real network.

Fig. 2. Visual representation of campaign results for a single simulation for scenarios S_M° (A1), S_F° (A2), S_{MF}° (A3), S_M^* (B1), S_F^* (B2), S_{MF}^* (B3)



In order to further confirm the validity of the proposed approach, the same study was performed on a synthetic Barabasi-Albert network. The BA model represents a multitude of social networks, both virtual and real, which are scale-free in their nature (Barabási and Bonabeau 2003). The synthetic network used in this research comprised of 1000 nodes, with average betweenness – 1687.295, degree – 3.994, closeness – 0.0002310899, and eigen centrality – 0.03661858.

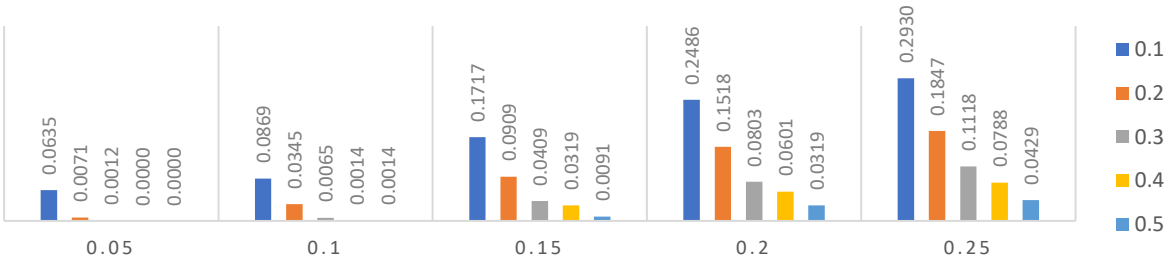
The averaged results of the simulations are presented in Table 3. Again, for the S_{MF}^* scenario, an increase in the average coverage in the targeted group can be observed at the cost of a slight decrease in the general coverage reached. This is a relevant research finding to practice, as this allows to reduce the marketing content overload, thus improving the multi-objective campaigns' quality.

Scenario	Average Iterations Count	Average Coverage	Decrease	Number of targeted nodes	Avg. coverage in targeted nodes	Increase
S_M°	4.9	0.1279	x	288	0.1257	x
S_F°	4.7	0.1274	x	130	0.0992	x
S_{MF}°	4.7	0.1282	x	418	0.1151	x
S_M^*	4.7	0.1266	-0.0013	288	0.1510	0.0253
S_F^*	4.4	0.1171	-0.0103	130	0.1862	0.087
S_{MF}^*	4.9	0.1254	-0.0028	418	0.1163	0.0012

Table 3. Comparison of the averaged results on the synthetic BA network.

Last, but not least, in order to further confirm the validity of the proposed approach, the same set of scenarios was tested for both the real and the synthetic networks with various values of the seeding fraction (0.05, 0.10, 0.15, 0.20 and 0.25) and the average propagation probability (0.1, 0.2, 0.3, 0.4, 0.5). The aggregate results for the S_{MF}^* scenario compared to the S_{MF}° scenario for the real network are presented on Fig. 3. The x-axis represents various seeding fractions and the bar colors represent various propagation probability values used in the simulations. The analysis of Fig. 3 allows to observe that the efficiency of the proposed approach increases with the growth of the seeding fraction, yet decreases with the growth of the propagation probability. Similar aggregate results were obtained for the studied synthetic network.

Fig. 3. Average increase in the coverage in the target group for the real network.



Conclusions

This study focuses on proposing a sustainable approach to fulfilling multiple campaign objectives by executing a reduced number of actual campaigns. MCDA methods are used to evaluate and rank multi-attribute nodes in social network for their adequacy for seeding information in the viral marketing campaign with multiple objectives.

The initial research performed on both real and synthetic networks has shown a promising increase in the coverage in the groups of targeted network nodes, especially for higher values of seeding fractions and lower propagation probabilities in the networks. From the practical point of view, the proposed approach can help to reduce the marketing content overload and the target users' irritation from high number of repeated messages from different campaigns, thus lowering such campaigns performance.

The future works in this study include further analysis of the proposed approach on various theoretic network models, and on vast range of real networks. Moreover, it would be beneficial to study the framework on networks with nodes characterized by a wider range of multiple attributes.

Acknowledgments

This work was supported by the National Science Centre of Poland, the decision no. 2017/27/B/HS4/01216 (AK, JJ) and within the framework of the program of the Minister of Science and Higher Education under the name "Regional Excellence Initiative" in the years 2019-2022, project number 001/RID/2018/19, the amount of financing PLN 10,684,000.00 (JW).

References

Barabási, A.-L., and Bonabeau, E. 2003. "Scale-Free Networks," *Scientific American* (288:5), pp. 60–69. (https://doi.org/10.1038/scientificamerican0503-60).

Hwang, C.-L., Lai, Y.-J., and Liu, T.-Y. 1993. "A New Approach for Multiple Objective Decision Making," *Computers & Operations Research* (20:8), pp. 889–899.

Kempe, D., Kleinberg, J., and Tardos, É. 2003. "Maximizing the Spread of Influence through a Social Network," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 137–146.

Mochalova, A., and Nanopoulos, A. 2014. "A Targeted Approach to Viral Marketing," *Electronic Commerce Research and Applications* (13:4), pp. 283–294. (https://doi.org/10.1016/j.elerap.2014.06.002).

Nguyen, H. T., Dinh, T. N., and Thai±, M. T. 2016. "Cost-Aware Targeted Viral Marketing in Billion-Scale Networks," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, , April, pp. 1–9. (https://doi.org/10.1109/INFOCOM.2016.7524377).

Ryan A. Rossi, Nesreen K. Ahmed. 2015. "Enron-Only | Email Networks | Network Data Repository," *Network Repository*. (http://networkrepository.com/email-enron-only.php, accessed July 11, 2020).

Saaty, T. L. 1988. "What Is the Analytic Hierarchy Process?," in *Mathematical Models for Decision Support*, Springer, pp. 109–121. (https://link.springer.com/content/pdf/10.1007/978-3-642-83555-1_5.pdf).

A10.

Karczmarczyk, A., Jankowski, J., Wątróbski, J. (2021). OONIS—Object-Oriented Network Infection Simulator. SoftwareX, 14, 100675.

SoftwareX 14 (2021) 100675

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Original software publication

OONIS — Object-Oriented Network Infection Simulator

Artur Karczmarczyk [a,*], Jarosław Jankowski [a], Jarosław Wątróbski [b]

[a] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology in Szczecin, Żołnierska 49, 71-210 Szczecin, Poland

[b] University of Szczecin, Mickewicza 64, 71-101 Szczecin, Poland

ARTICLE INFO

Article history:
Received 12 February 2020
Received in revised form 4 August 2020
Accepted 2 February 2021

Keywords:
Complex networks
Social networks
Independent cascades model
Seeding
Sequential seeding
Information spreading

ABSTRACT

Online systems with the highest global audiences take form of widely-used social platforms. Their immense traffic resulted in increased attention from researchers into various phenomena including information propagation in social networks. Although there exist some libraries, such as igraph and netdep, which allow representation of graphs in the R language, due to continual appearance of new models and information spreading approaches, the researchers are usually forced to write their own scripts to perform actual simulations and study their results.

In this paper, the authors propose an object-oriented library and environment in R, for running simulation experiments focused on information spreading within complex networks. Object-oriented programming paradigms such as encapsulation, separation of concerns and modularity were used in the proposed software, to provide researchers with a scalable framework allowing quick and easy creation of experimental scenarios for studying information propagation in complex networks. It also supports new approaches, not available in other libraries, related to spreading seeds over the time in a form of sequential seeding, as well as coordinated execution, making it possible to compare algorithms in invariable experimental conditions.

Code metadata

Current Code version	v1.0.0
Permanent link to code / repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX_2020_39
Legal Code License	GNU GPLv3
Code Versioning system used	git
Software Code Language used	R
Compilation requirements, Operating environments & dependencies	RStudio, 1.2.*
If available Link to developer documentation / manual	
Support email for questions	oonis@ideaspot.pl

Software metadata

Current software version	1.0.0
Permanent link to executables of this version	https://github.com/IdeaSpotPL/oonis/releases/tag/v1.0.0
Legal Software License	GNU GPLv3
Computing platform / Operating System	Linux, OS X, Microsoft Windows
Installation requirements & dependencies	R version 3.0.1+ https://cran.rstudio.com/
If available Link to user manual — if formally published include a reference to the publication in the reference list	
Support email for questions	oonis@ideaspot.pl

1. Introduction

Complex networks evolved from early-stage technical systems into popular social-media platforms [1]. It was observed that

* Corresponding author.
E-mail address: artur.karczmarczyk@zut.edu.pl (Artur Karczmarczyk).

https://doi.org/10.1016/j.softx.2021.100675

advertising in social networks can bring better results within limited budgets compared to traditional electronic marketing campaigns [2]. This advertising potential makes the information propagation in complex networks an interesting research problem [3].

Graphs and complex networks, as well as information spreading processes within them, is an interdisciplinary research topic studied in disciplines such as computer science, physics, medicine, epidemiology [4–6]. The independent cascades (IC) model and agent simulations allow to study the processes of information propagation in complex networks [7].

There exist some libraries in the R language, such as *igraph*[1] and *netdep*,[2] which allow and facilitate representation of graphs and complex networks in the R scripting environment [8,9]. However, in order to perform experiments in the IC model, the researchers are forced to write their own scripts, especially if they focus on adaptive or sequential seeding. Mixing the IC model logic with the actual information spreading logic, requires the researchers to reimplement the complete script each time a new study is done. Moreover, it ignores the achievements and benefits of the object-oriented programming paradigm [10]. This created an interesting software gap, which the authors of this paper decided to fill with an object-oriented library and surrounding environment for simulations and studying the processes of information propagation in complex networks.

The application of object-oriented programming paradigms allowed to encapsulate the layer of complex and repeatable logic of network information spreading processes simulations into library class methods. Moreover, the followed modular approach and separation of concerns, allowed to build an environment in which creation of simulation scenarios is easy, scalable and fast, with the use of interchangeable, extensible modules for seeding information, contamination of nodes and results printing.

To sum up, the authors propose an innovative object-oriented library in R, along with extensible environment for simulations and studying of information propagation in complex networks. The paper is divided into sections. Section 2 presents the background of the problem solved by the proposed software. Section 3 exhibits the foundations of the proposed software architecture. This is followed by three illustrative examples in Section 4 and conclusions in Section 5.

2. Problems and background

Information propagation in complex networks is an interesting research topic, studied in marketing, social media, medicine and physics, to name just a few [5,6]. One of the most popular approaches to studying it is the independent cascades (IC) model [11]. The model is founded on the following assumptions. Initially, none of the nodes in a network are aware of the information. The information is then seeded to a fraction of the network nodes. In the IC model, the nodes aware of the information try to infect all neighboring nodes. A single trial for infecting the neighbor is possible for each of the infected nodes. A random value is drawn and if it exceeds the propagation probability of the to-be-infected node, the information is passed to the latter. Otherwise, the information is not passed, and the former will not be able to try to pass the information to the latter anymore.

Although there exist some libraries in the R language, such as igraph, for storing and representing graphs and complex networks [8], in order to execute the IC model agent simulations, researchers need to develop their own scripts. The implementation of the IC model simulations is time consuming and repeatable.

[1] https://igraph.org/.

[2] https://cran.r-project.org/web/packages/netdep/index.html.

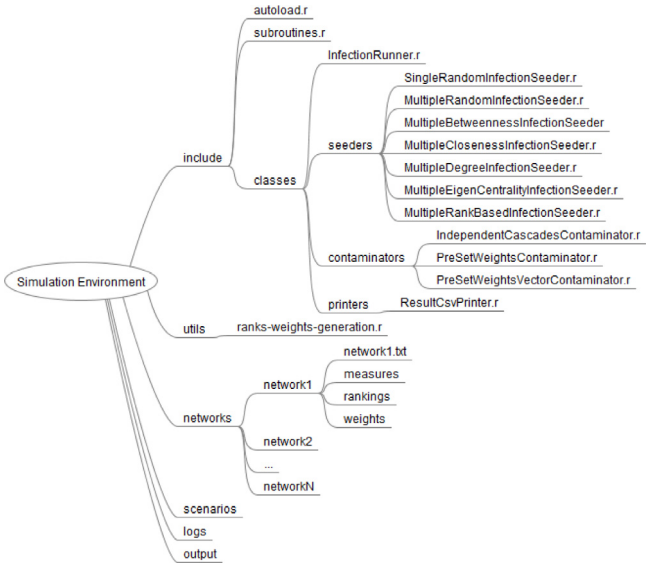


Fig. 1. The structure of the proposed environment for simulation experiments.

If the IC model logic is mixed by the researchers with the actual experiment logic, creation of each new research scenario requires copying and maintaining the IC model logic. Apart from the basic version of IC, further extensions towards adaptive seeding are required in many applications. It was the motivation for implementing sequential seeding based on spreading seeds over time with given parameters.

The main objective of the proposed software was to apply encapsulation and separation of concerns from the object-oriented programming paradigm [12], in order to separate the IC model logic layer from the layer of logic for studying the information spreading processes in complex networks. This allowed creation of experiment scenarios by replacing information spreading algorithms and approaches, without the need to make effort for maintaining the simulation logic layer. As a result, a scalable, flexible and time-efficient software solution to simulation research on information spreading in complex networks was produced.

3. Software framework

3.1. Software architecture

The proposed software is composed of two parts. The first one is the InfectionRunner, an object-oriented library class in R for simulations of information propagation in complex networks. The second one is a proposed environment for conducting experiments with the aforementioned library embedded inside.

The library comprises of the InfectionRunner class and three sets of pluggable modules for the main class. Three types of pluggable modules are available for this library: seeder, contaminator and result printer. The seeder is the component responsible for specifying which nodes and during which simulation iterations should be seeded with the information to pass through to the network. The contaminator is the component which determines how the network nodes which are already aware of the information would pass it to its neighboring nodes. Eventually, the result printer module allows to log the simulation results in a demanded format. The actual manner in which the three replaceable modules are cooperating is presented in Section 3.3.

The structure of the proposed environment for simulation experiments is presented on Fig. 1. The environment in its basic form comprises of 6 directories:

- **include** — the proposed library is embedded in this directory (classes), as well as some utility functions (subroutines.r) and list of all classes for the ease of library importing to other scripts (autoload.r);
- **utils** — it is proposed to store utility scripts in this directory; the software is shipped with *ranks-weights-generation.r* script which is used for computing networks' nodes' centrality measures such as degree, betweenness, closeness, eigen centrality, and ranking nodes based on these measures. Moreover, this utility generates sets of random weights assigned to each node of a network, which allow to test various algorithms on the network in immutable, repeatable conditions;
- **networks** — it is suggested to store networks for experiments in this directory; the networks should be in edge-list text format. The *ranks-weights-generation.r* util can be used to compute the network measures, rankings and weights for repeatability of research;
- **scenarios** — directory for storing individual simulation experiments;
- **logs** — directory for storing logs;
- **output** — directory for storing simulation results printed by the *<result printer>* module of the simulation engine.

3.2. Software functionalities

The proposed software provides an innovative object-oriented framework for running simulations of information spreading in complex networks. The encapsulation of the logic of independent cascades' infection model allows an easy reuse and scalability of experiment scenarios and scripts. The software provides three slots for interchangeable modules required for the actual simulation scenarios creation: seeder module, contaminator module and result printer module. Such encapsulation and separation of concerns, allows to easily perform experiments based on various approaches including, but not limited to:

- seeding variable fraction of initial nodes;
- seeding initial nodes in single or multiple iterations [13];
- influencing the information spreading process by probability spraying over the network nodes [14];
- evaluation and planning of viral marketing campaigns in social networks based on parametrized values of seeding fraction, propagation probability, nodes' centrality measures and rankings [15].

3.3. Sample code snippets analysis

3.3.1. Running the simulation

The InfectionRunner class is the core of the proposed library. It contains all the code and logic required for performing simulations based on the independent cascades model, with the *run()* method being the heart of the simulation. The method is presented on Fig. 2. The figure presents the *InfectionRunner* class definition with all methods but *run()* omitted for brevity.

In order to execute the *run()* method, the maximum number of iterations to simulate needs to be provided. The implementation of the *run()* method starts with a loop for each iteration of the simulation (lines 9–32). The current iteration number is stored in the *InfectionRunner* class at the beginning of the loop (line 10), in order to be accessible to all pluggable seeder, contaminator and printer modules.

Each iteration starts with an optional seeding. A call is made to the currently plugged-in seeder module with information on the current graph structure and current iteration number (line 11). As a result, the vector of vertices to seed the information to is returned. In order to keep the software easily extensible, it was decided that both seeder and contaminator modules will only provide the vector of nodes. The actual infection is performed by the *InfectionRunner* core class (lines 14–16).

In line 19, a call is made to the currently plugged-in contaminator module with information on the current graph structure and the current iteration number. Then, in lines 20–27, the vector of nodes obtained from the contaminator module is used to propagate the infection over the network. One of the assumptions of the IC model is that each infected node has only a single chance to infect the neighboring nodes. Therefore, in lines 24–27 the vertices used in current iteration are stored in order to abstain from using them again for infections.

If at least a single node was seeded or infected in the iteration, the process resumes with next iterations. The *run()* method can be supplied with optional parameter *minIterations*, with default value set to 1. In case when no nodes are infected during the current iteration and at least *minIterations* were already simulated, the process stops (lines 29–31).

Eventually, a call to the result printer is made with the current graph structure and information on the current information in order to store the simulation results (line 33).

3.3.2. OONIS library usage

The usage of the *InfectionRunner* class is very simple. In order to use it in one's experiments, a scenario script should be created. The actual contents of the scenario script depends on the researcher's needs, but should be similar to the code presented on Fig. 3.

Initially, a decision should be made which seeder, contaminator and result printer modules should be used in the scenario. The proposed software is shipped with 7 sample seeders:

- **SingleRandomInfectionSeeder** — this seeder module infects initial nodes based on random. Only a single seeding iteration is possible;
- **MultipleRandomInfectionSeeder** — this seeder module infects initial nodes based on random. Multiple seeding iterations are possible in order to allow studying sequential seeding approaches. This is a very simple seeder and works best when multitude of simulations is run in order to obtain statistical data. Due to lack of repeatability of the random results, it is not suggested to use it for experiments with low quantity of trials.
- **Multiple{Betweenness | Closeness | Degree | EigenCentrality}InfectionSeeder** — These seeder modules infect initial nodes based on their betweenness/closeness/degree/eigen centrality rank. Multiple seeding iterations are possible in order to allow studying sequential seeding approaches.
- **MultipleRankBasedInfectionSeeder** — This seeder module infects initial nodes based on their rank provided in one of the input parameters. Multiple seeding iterations are possible in order to allow studying sequential seeding approaches.

Moreover, the software is shipped with 3 sample contaminators:

- **IndependentCascadesContaminator** — This contaminator module tries to contaminate the nodes that are neighbors to already infected nodes. This is a very simple module, which draws the random value <0;1> to compare with the propagation probability value on each infection trial. Therefore, this contaminator is useful for some simple tests, but is inadequate if repetitive results are required. If using this contaminator, please consider running multitude of repeated simulations to obtain average statistic values for your further research.

```
InfectionRunner <- setRefClass(
  "InfectionRunner",
  fields = c('graph', 'vertices', 'currentIteration', 'infectionSeeder', 'contaminator', 'resultPrinter'),
  methods = list(

    # ...

    run = function(numIterations, minIterations = 1) {
      for (i in 1: numIterations) {
        currentIteration <<- i;

        # seeding phase
        verticesToSeed <- infectionSeeder$seed(graph, currentIteration);
        if (length(verticesToSeed) > 0) {
          infect(verticesToSeed);
        }

        # contamination phase
        contaminationResultList <- contaminator$contaminate(graph, currentIteration);
        verticesToInfect <- contaminationResultList$verticesToInfect;
        if (length(verticesToInfect) > 0) {
          infect(verticesToInfect)
        }
        verticesUsed <- contaminationResultList$verticesUsed;
        if (length(verticesUsed) > 0) {
          markAsUsed(verticesUsed)
        }

        if (length(verticesToSeed) == 0 & length(verticesToInfect) == 0 && currentIteration > minIterations) {
          break;
        }
      }
      resultPrinter$print(graph, currentIteration);

    # ...

    }
  )
)
```

Fig. 2. Fragment of the source code for the InfectionRunner class, representing the run() method.

```
# SEEDER
# choose and configure the simulation seeder component
seeder <- SomeDefaultOrCustomSeeder(...);

# CONTAMINATOR
# choose and configure the simulation contamination component
contaminator <- SomeDefaultOrCustomContaminator(...)

# RESULT PRINTER
resultPrinter <- SomeDefaultOrCustomResultPrinter(...);

# SIMULATION
maxIterations <- 100; minIterations <- 1;
ir <- InfectionRunner(infectionSeeder = seeder, contaminator = contaminator, resultPrinter = resultPrinter)
ir$readFromEdgesTxt('/path/to/network.txt', FALSE)
ir$run(maxIterations, minIterations)
```

Fig. 3. A possible source code of a simple complex network simulation experiment scenario with the use of the proposed environment and the *InfectionRunner* class.

- **PreSetWeightsContaminator** — This contaminator module tries to contaminate the nodes that are neighbors to already infected nodes. Each node is assigned a pre-drawn <0;1> value for comparing with the configured propagation probability in order to verify whether or not the contamination should occur, according to the coordinated execution method proposed in [16]. This contaminator uses a single average value of propagation probability for all nodes. If there is a need to simulate scenarios in which some nodes are provided more incentives than the others, in order to increase their propagation probability, a more advanced version of *PreSetWeightsVectorContaminator* can be used.
- **PreSetWeightsVectorContaminator** — This contaminator module tries to contaminate the nodes that are neighbors to already infected nodes. Each node is assigned a pre-drawn <0;1> value for comparing with the configured propagation probability in order to verify whether or not the contamination should occur. This contaminator allows to adjust the probability of contamination to particular nodes based on their rank. This allows to simulate scenarios in which some nodes are provided more incentives than the others, in order to increase their propagation probability.

Eventually, the software is shipped with a single results printer, **ResultCsvPrinter**. This module saves the results in a CSV format, which is easily readable by majority of analytic software. The output is split into 6 columns: (1) label of the experiment run, (2) total number of iterations, (3) total count of vertices in the network, (4) total count of infected vertices, (5) the iteration of last infection, (6) obtained coverage.

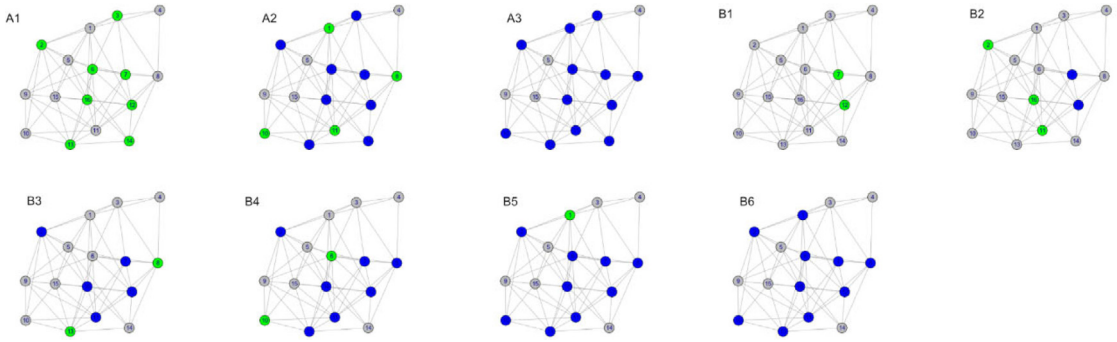


Fig. 4. Visualization of the information propagation process for the tribes [17] network.

Table 1
Single simulation run scenario parameters.

Scenario file	/oonis/scenarios/example/01-tribes-sinle-run.r
Network	Tribes [17]. 16 nodes, 58 edges, average degree: 7.
Iterations	Min: 1; max: 120
Seeding fraction	0.5, all seeded at iteration 1
Propagation probability	0.2, same for all nodes of the network
Seeder	MultipleRankBasedInfectionSeeder; nodes ordered by degree rank
Contaminator	PreSetWeightsVectorContaminator; single set of pre-drawn weights

4. Illustrative examples

In this section, four illustrative examples will be presented, demonstrating the major functions of the proposed library and simulation environment.

4.1. Single simulation run

In this scenario, a simple scenario of single infection simulation is presented. The simulation is performed on a network [17] built on 16 nodes and 58 edges (see Table 1). Fig. 4 A1–A3 present each iteration of the simulation. Green color represents nodes infected in the current iteration, whereas blue color indicates that the node was already used for infecting other nodes and cannot be reused. In this scenario, 8 nodes are seeded with the information. In the second iteration, the information is passed to four more nodes. No infections occur in the third iteration and thus the process ends, with 12 nodes infected, i.e. 75% coverage.

4.2. Sequential seeding, single simulation run

In this scenario, the same network [17] is seeded with information. However, considerably smaller number of nodes is seeded, and the seeding is performed sequentially — two nodes in the first iteration and two nodes in the second iteration (see Table 2). Fig. 4 B1–B6 present each iteration of the simulation.

In the first iteration, nodes 7 and 12 are infected. In the second iteration, node 11 is infected by both of the nodes. Moreover, nodes 2 and 16 are further seeded with information. In the third iteration additional two nodes are infected, followed by two nodes in iteration 4 and three nodes in iteration 5. The process ends in iteration 6 with a total of 10 nodes infected, i.e. 62.5% coverage.

4.3. Multiple parameters, multiple scenarios

In this scenario, in contrast to the previous ones, a total of 40 simulations is performed (see Table 3). A total of four sets of parameters is studied, resulting from a cartesian product of the sets of seeding fractions and propagation probabilities. Moreover, for each set of parameters, not a single but a set of 10 simulations for various pre-drawn weights is performed. The results of the simulations are presented in Table 4. The results from all 10 simulations for pre-drawn weights, need to be grouped and averaged for each set of parameters, because depending on the weights, slightly different results can be obtained.

This can be observed on Fig. 5, where all iterations for two sets of weights for the case with seeding fraction and propagation probability equal to 0.1 are presented. Because both simulations are executed on the same network and with identical parameters, both simulations start with seeding the same nodes (see C1.1 and C2.1). However, because of different pre-drawn weights of the nodes, differences occur already in iteration 2 (see C1.2 with 14 newly-infected nodes, compared to C2.2 with only 7 ones). Last infection occurs in the 5th iteration for the first set of weights and in the 8th iteration for the second one, with the coverage of the former being two-fold value of the latter.

4.4. Influence maximization problem

While the prior 3 scenarios were some basic examples explaining the mechanisms of the OONIS library, in this section a more real-life usage example for practitioners and researchers is provided. The problem of influence maximization in a network is studied.

The OONIS library allows to study influence maximization on any undirected network which can be fed to the library in an edge list format. Therefore, both real and synthetic networks can be studied. While the prior 3 scenarios focused on small real networks, in this section the focus is shifted to synthetic networks, which are often used by researchers to better study processes occurring in complex networks.

Although the OONIS library does not provide mechanisms for synthetic networks generation, the *igraph* library can be used in conjunction with OONIS to study information propagation and influence maximization in synthetic networks. The most-commonly used synthetic networks are based on the free-scale model proposed by Barabasi–Albert (BA) [19], small world model proposed by Watts–Strogatz (WS) [20] and random graph model introduced by Erdos–Renyi (ER) [21]. The *igraph* library provides tools for generating each of them: *barabasi.game()*, *watts.strogatz.game()* and *erdos.renyi.game()* respectively.

Table 2
Parameters for single simulation run scenario with sequential seeding.

Scenario file	/oonis/scenarios/example/02-tribes-sequential-seeding.r
Network	Tribes [17]. 16 nodes, 58 edges, average degree: 7.
Iterations	Min: 1; max: 120
Seeding fraction	0.125, two nodes seeded in iteration 1 and another two in iteration 2
Propagation probability	0.2, same for all nodes of the network
Seeder	MultipleRankBasedInfectionSeeder; nodes ordered by degree rank
Contaminator	PreSetWeightsVectorContaminator; single set of pre-drawn weights

Table 3
Parameters for the scenario with multiple parameters and multiple simulations.

Scenario file	/oonis/scenarios/example/03-email-multiple-simulations.r
Network	Emails [18]. 143 nodes, 623 edges, average degree: 8.
Iterations	Min: 1; max: 120
Seeding fractions	(1) 0.1; (2) 0.2; all seeded at iteration 1
Propagation probabilities	(1) 0.1; (2) 0.2; same for all nodes of the network
Seeder	MultipleRankBasedInfectionSeeder; nodes ordered by degree rank
Contaminator	PreSetWeightsVectorContaminator; 10 sets of pre-drawn weights

Table 4
Simulation results for the scenario with multiple parameters and multiple simulations. Abbreviations: W — weight, Inf — infected nodes, LI — last infection iteration, C — coverage, SF — seeding fraction, PP — propagation probability.

(A) SF: 0.1, PP: 0.1				(B) SF: 0.1, PP: 0.2				(C) SF: 0.2, PP: 0.1				(D) SF: 0.2, PP: 0.2			
W	Inf	LI	C	W	Inf	LI	C	W	Inf	LI	C	W	Inf	LI	C
1	65	5	0.45	1	101	8	0.71	1	76	5	0.53	1	105	8	0.73
2	31	8	0.22	2	94	9	0.66	2	63	5	0.44	2	98	6	0.69
3	38	8	0.27	3	107	8	0.75	3	64	5	0.45	3	109	5	0.76
4	57	6	0.40	4	82	7	0.57	4	67	4	0.47	4	85	5	0.59
5	44	7	0.31	5	101	7	0.71	5	69	7	0.48	5	107	5	0.75
6	53	11	0.37	6	98	6	0.69	6	68	6	0.48	6	102	6	0.71
7	24	5	0.17	7	89	6	0.62	7	44	5	0.31	7	100	5	0.70
8	44	7	0.31	8	101	10	0.71	8	71	6	0.50	8	106	5	0.74
9	52	8	0.36	9	86	7	0.60	9	78	7	0.55	9	95	5	0.66
10	29	4	0.20	10	99	7	0.69	10	64	5	0.45	10	104	5	0.73
Avg.	43.7	6.9	0.31		95.8	7.5	0.67		66.4	5.5	0.46		101.1	5.5	0.71

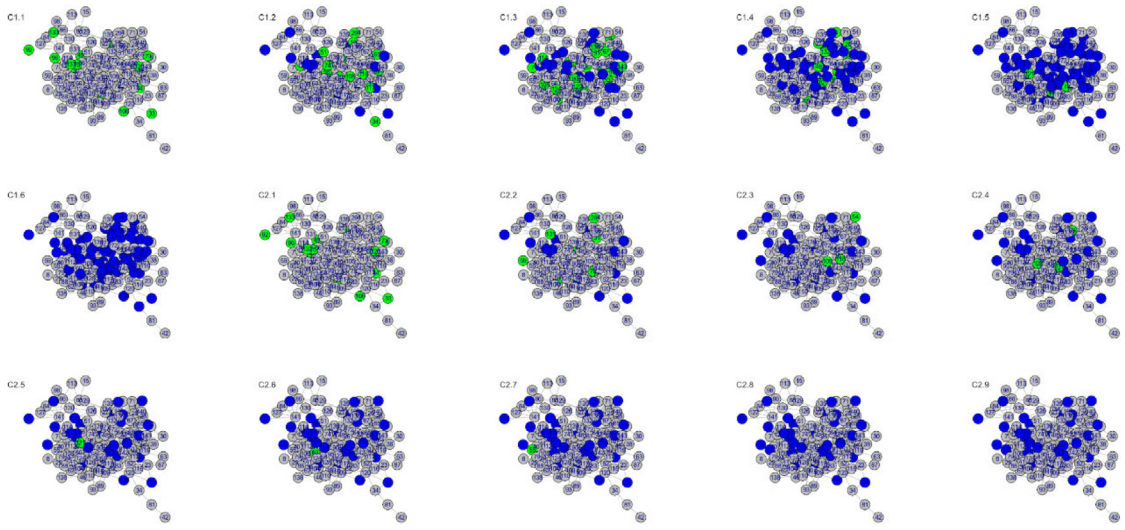


Fig. 5. Visual representation of the progress of the infection for seeding fraction and propagation probability equal to 0.1 and pre-drawn weights' sets 1 (C1.1–C1.6) and 2 (C2.1–C2.9).

For the purpose of this scenario, a 1000-node BA synthetic network with the exponent $\lambda = 2.5$ was generated with the use of the *barabasi.game()* function. The network contained 1000 vertices, 1997 edges. The minimum degree was 2, the maximum degree was 45 and the average degree was 3.994. The network is presented on Fig. 6.

In this scenario, the effect of the seeding fraction, average propagation probability and the measure used for selecting seeds on the final influence coverage was studied. The investigated values were as follows:

- seeding fraction: 0.01, 0.02, …, 0.19, 0.20;
- propagation probability: 0.05, 0.10, 0.15, 0.20;

Table 5
Influence coverage in the BA synthetic network.

SF	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	0.1	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.19	0.2
PP	Eigencentrality																			
0.05	0.0205	0.0349	0.0481	0.0604	0.0722	0.0826	0.0938	0.106	0.1187	0.1294	0.1411	0.153	0.1639	0.1743	0.1849	0.1957	0.2061	0.2164	0.2265	0.2366
0.1	0.0411	0.0627	0.0793	0.095	0.1102	0.1239	0.1353	0.1493	0.1658	0.1775	0.1907	0.2026	0.2136	0.2238	0.2353	0.2466	0.2567	0.2671	0.2764	0.2851
0.15	0.0781	0.1078	0.1278	0.1477	0.1609	0.178	0.1918	0.2067	0.2225	0.2345	0.247	0.2582	0.2706	0.2816	0.2935	0.3049	0.3158	0.3268	0.3371	0.346
0.2	0.1406	0.1723	0.1938	0.2144	0.2273	0.2456	0.2591	0.2733	0.2911	0.3029	0.3158	0.3268	0.3383	0.3496	0.3642	0.3746	0.3865	0.3963	0.4063	0.4141
	Betweenness																			
0.05	0.0208	0.034	0.0485	0.0606	0.0733	0.0847	0.0961	0.108	0.1193	0.1311	0.1411	0.1527	0.1644	0.175	0.1856	0.1963	0.2073	0.2186	0.229	0.2402
0.1	0.042	0.0632	0.0835	0.0991	0.1151	0.1288	0.1427	0.157	0.1699	0.1833	0.1945	0.2067	0.2198	0.231	0.243	0.253	0.2631	0.2741	0.2837	0.2947
0.15	0.0803	0.1078	0.1302	0.1494	0.1677	0.184	0.1989	0.2136	0.2276	0.2416	0.2544	0.2666	0.2806	0.2927	0.3049	0.3151	0.3255	0.3361	0.3458	0.3563
0.2	0.1439	0.1795	0.2048	0.2229	0.2399	0.2552	0.2709	0.2829	0.2973	0.3118	0.3244	0.3358	0.3493	0.3616	0.3728	0.3817	0.3914	0.3996	0.4093	0.4186
	Degree																			
0.05	0.0208	0.0343	0.0486	0.0612	0.073	0.084	0.0947	0.106	0.1161	0.1287	0.1407	0.1521	0.1636	0.1742	0.1857	0.1951	0.2057	0.217	0.2275	0.2384
0.1	0.042	0.0621	0.0819	0.0975	0.1131	0.1265	0.1401	0.1536	0.1633	0.1781	0.193	0.2041	0.2173	0.2288	0.2417	0.2507	0.2609	0.2716	0.2818	0.2916
0.15	0.0793	0.1041	0.1272	0.145	0.1629	0.1787	0.1944	0.2089	0.2194	0.2354	0.2509	0.2632	0.2769	0.2886	0.3019	0.3106	0.3211	0.3306	0.3403	0.3493
0.2	0.1403	0.1723	0.1968	0.2174	0.2342	0.2498	0.263	0.2801	0.2902	0.3054	0.3197	0.3323	0.3469	0.3559	0.3683	0.3767	0.3869	0.396	0.4053	0.4135

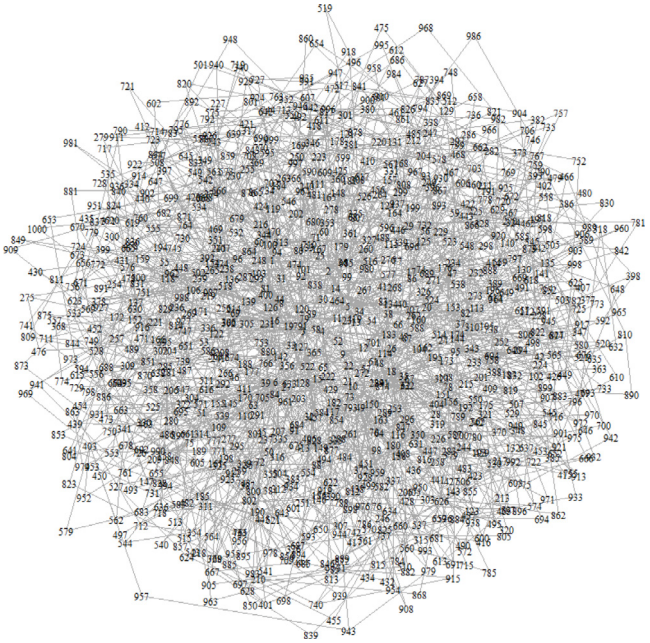


Fig. 6. Visual representation of the BA synthetic network used to illustrate the influence maximization problem.

- measures: degree, betweenness, eigencentrality.

After performing 2400 simulations (Cartesian product of each of the parameter values above, each executed on a set on 10 predrawn weights), the results were aggregated and are presented on Fig. 7 and in Table 5.

The analysis of Table 5 allows to observe how the increase in the values of seeding fraction and of propagation probability result in increase of the final coverage in the network. If the value of 0.2 is considered for both the seeding fraction and average propagation probability, it can be noted for the studied network that the best results are obtained if the seeds are chosen based on the betweenness measure (0.4186 coverage, compared to 0.4141 for eigencentrality and 0.4135 for the degree). On the other hand, if the seeding fraction is reduced to 0.01, and the propagation probability is reduced to 0.05, the degree measure is as good as the betweenness measure (0.0208, compared to 0.0205 for eigencentrality).

5. Possible extensions

Due to the fact that the OONIS library was built following the object-oriented programming paradigms, it is easily extensible. Information spreading processes can be modeled with the use of various approaches like models SIS and SIR, derived from epidemiology, with their further extensions [4], Independent Cascade Model [7], branching processes [22], social influence modeling with Linear Threshold Model [7] or scalable influence maximization [3]. Such diversity of models was the motivation for flexibility at the level of module covering spreading mechanics. Currently the library supports the Independent Cascade Model, however, the implemented mechanics can be treated as an example and used by researchers to extend it to support other models, such as the linear threshold model. That would require replacing the default **InfectionRunner** class provided in the library with a custom one, supporting the linear threshold model. R's Reference Classes inheritance can be used, to reuse some of the existing code, common for both linear threshold and independent cascade model. However, at least the *init()* and *run()* methods of the **InfectionRunner** class should be overridden. On the other hand, however, some components such as seeders and printers, could be reused in an unmodified form.

Moreover, the OONIS library can be easily complemented by other R libraries and modules. This was demonstrated in Section 4.4, where an *igraph* function was used to produce a synthetic network for further research with the use of the OONIS library. Such interoperability of the library with other general purpose network analysis libraries provides a great potential to research areas including, but not limited to information diffusion in networks with unknown community organization, seeding nodes from different communities, identifying inter-community links involved in the activation of nodes, propagation probability spraying in information diffusion processes.

6. Conclusions

In this paper, an innovative object-oriented library and simulation environment in R was presented. The software allows researchers to study the process of information spreading in complex networks under various network characteristics and campaign parameters. Because of the implemented separation of concerns, as well as encapsulation and interchangeability of seeding, contaminating and result printing modules, the framework can be easily extended to accommodate custom research requirements while, at the same time, the IC information propagation model logic remains intact.

During the research, areas of possible improvement and future works were identified. The *InfectionRunner* class interface could be extended in order to allow storing additional information about the nodes. This, in turn, would allow to further extend the environment by providing new seeder modules to allow additional research, such as multi-criteria seed selection.

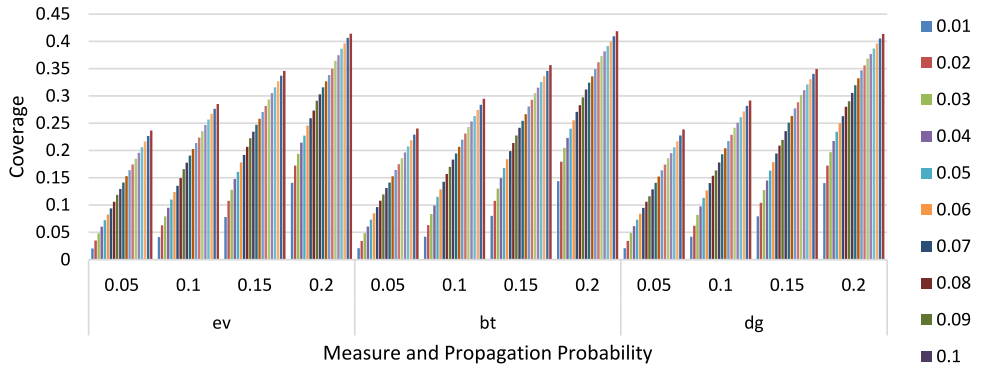


Fig. 7. Influence coverage in the BA synthetic network, grouped by measure used: ev — eigencentrality, bt — betweenness, dg — degree; and propagation probability — 0.05, 0.10, 0.15, 0.20. Individual bars in the chart legend represent the seeding fraction.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Science Centre, Poland, grant no. 2016/21/B/HS4/01562 (JJ, AK) and the Regional Excellence Initiative programme of the Minister of Science and Higher Education of Poland, years 2019–2022, project no. 001/RID/2018/19, financing 10 684 000,00 PLN (JW).

References

[1] Bello-Orgaz G, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Inf Fusion 2016;28:45–59. http://dx.doi.org/10.1016/j.inffus.2015.08.005.

[2] Watts DJ, Peretti J, Frumin M. Viral marketing for the real world. Harvard Business School Pub.; 2007.

[3] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. 2010. p. 1029–38.

[4] Kandhway K, Kuri J. How to run a campaign: Optimal control of SIS and SIR information epidemics. Appl Math Comput 2014;231:79–92. http://dx.doi.org/10.1016/j.amc.2013.12.164.

[5] Hinz O, Skiera B, Barrot C, Becker JU. Seeding strategies for viral marketing: An empirical comparison. J Mark 2011;75(6):55–71. http://dx.doi.org/10.1509/jm.10.0088.

[6] Iribarren JL, Moro E. Impact of human activity patterns on the dynamics of information diffusion. Phys Rev Lett 2009;103(3):038702. http://dx.doi.org/10.1103/PhysRevLett.103.038702.

[7] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. p. 137–46.

[8] Ju W, Li J, Yu W, Zhang R. iGraph: an incremental data processing system for dynamic graph. Front Comput Sci 2016;10(3):462–76. http://dx.doi.org/10.1007/s11704-016-5485-7.

[9] Csárdi G, Nepusz T, Airoldi EM. Statistical network analysis with igraph. New York, NY: Springer; 2016.

[10] Cox BJ. Object-oriented programming : An evolutionary approach. 1986, Accessed: Feb. 02, 2020. [Online]. Available: http://cumincad.scix.net/cgi-bin/works/2010%20+dave&hits=2:/Show?2a91.

[11] Shakarian P, Bhatnagar A, Aleali A, Shaabani E, Guo R. The independent cascade and linear threshold models. 2015, p. 35–48.

[12] Ober I, Ober I. On patterns of multi-domain interaction for scientific software development focused on separation of concerns. Procedia Comput Sci 2017;108:2298–302. http://dx.doi.org/10.1016/j.procs.2017.05.288.

[13] Jankowski J, Zioło M, Karczmarczyk A, Wątróbski J. Towards sustainability in viral marketing with user engaging supporting campaigns. Sustainability 2017;10(2):15. http://dx.doi.org/10.3390/su10010015.

[14] Karczmarczyk A, Bortko K, Bartków P, Pazura P, Jankowski J. Influencing information spreading processes in complex networks with probability spraying. In: 2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). 2018, p. 1038–46. http://dx.doi.org/10.1109/ASONAM.2018.8508637.

[15] Karczmarczyk A, Jankowski J, Wątróbski J. Multi-criteria decision support for planning and evaluation of performance of viral marketing campaigns in social networks. PLOS ONE 2018;13(12):e0209372. http://dx.doi.org/10.1371/journal.pone.0209372.

[16] Jankowski J, Szymanski BK, Kazienko P, Michalski R, Bródka P. Probing limits of information spread with sequential seeding. Sci Rep 2018;8(1):1–9. http://dx.doi.org/10.1038/s41598-018-32081-2.

[17] Read KE. Cultures of the central highlands, New Guinea, Southwest. J Anthropol 1954;10(1):1–43. http://dx.doi.org/10.1086/soutjanth.10.1.3629074.

[18] Rossi R, Ahmed N. The network data repository with interactive graph analytics and visualization. In: Presented at the twenty-ninth AAAI conference on artificial intelligence, Mar. 2015, Accessed: Feb. 02, 2020. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9553.

[19] Barabási A-L, Albert R. Emergence of scaling in random networks. Science 1999;286(5439):509–12. http://dx.doi.org/10.1126/science.286.5439.509.

[20] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393(6684):440–2. http://dx.doi.org/10.1038/30918.

[21] Erdös P, Rényi A. On random graphs' math. Debrecen 1959;6:290–7.

[22] van der Lans R, van Bruggen G, Eliashberg J, Wierenga B. A viral branching model for predicting the spread of electronic word of mouth. Mark Sci 2009;29(2):348–65. http://dx.doi.org/10.1287/mksc.1090.0520.