Knowledge-Based Systems 193 (2020) 105430

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Diversified viral marketing: The power of sharing over multiple online social networks☆

Dawood Al Abri [a,*,1], Shahrokh Valaee [b]

[a] Depertament of Electrical and Computer Engineering, Sultan Qaboos University, Oman
[b] Edward S. Rogers Sr. Department of Electrical & Computer Engineering, University of Toronto, Canada

ARTICLE INFO

Article history:
Received 21 June 2019
Received in revised form 20 November 2019
Accepted 22 December 2019
Available online 28 December 2019

Keywords:
Viral marketing
Diffusion model
Social networks
Advertisement
Influence maximization

ABSTRACT

The popularity of online social networks (OSNs) makes them attractive platforms to advertise products. Previous work on marketing in OSNs utilized older diffusion models that do not capture the interactions of modern OSNs and hence there is a need to develop a model that accounts for the interactions that occur in current OSNs. In this paper, we introduce a new model for information flow in online social networks that captures the sharing behavior exercised by users when they pass information from one online social network to their social circles in another network. We, then, formulate a problem of maximizing the marketing reach where the diversity of users' other social networks is taken as a constraint. We also propose a greedy algorithm to solve the aforementioned optimization problem. Numerical results show that the proposed algorithm achieves better results than algorithms that are based on classical degree centrality metric and with comparable running time.

1. Introduction

Social networks impact many aspects of individuals life such as whom they have a discussion with and consequently, whom may influence their opinion about different issues. Beyond the human influence, these social networks have been exploited to design better algorithm for various applications such as: routing algorithms that guide data forwarding based on the social ties (e.g. [1]), proactively reducing the peak load of the cellular network based on information dissimulation on the social network (e.g. [2]), etc. Recently, the popularity of online social networks (OSNs) and their increase influence in shaping and swaying people opinion has driven advertisers to utilize them as a marketing platform. This can be done directly where an advertiser pays the OSN to place an ad on a popular content (e.g. popular YouTube video) or indirectly via giving a freebie to some influential users of the targeted social network in the hope that they make a good review of the product or recommend it to their online social circles. Therefore, the ability to identify influential users is crucial for effective viral marketing strategy. The basic idea behind viral marketing is to market a product to few influential individuals who can endorse the product to their social networks. The hope is to create a cascade effect to spread the marketing of the product further down the social network; from one person to his acquaintance circles and so forth.

The use of social relationship to influence other people's opinion is something that had been studied before using diffusion models, which are used to model how ideas, innovation, influence, and diseases spread through social networks. Most of the works that attempt to determine the influential nodes are based on two diffusion models: the threshold model and cascade model [3]. In the threshold model, a weight is used to quantify the ability of a user to influence his friend. If the sum of weights of a user's active friends (i.e. those who bought the product) exceeds certain threshold, then that user becomes active. In the simplest form of the cascade model, each active user is given one chance to convince its inactive neighbors to adopt the product according to some probability distribution.

The basic forms of the current diffusion models were developed long before the widespread use of Internet and the explosive increase in using OSNs (Early forms of these models appear in the sociology literature in the 1970's [4,5]). Hence, their development were mainly based on the social interaction between individuals and as a result, they do not account for the specific interactions of modern OSNs (e.g. share, re-tweet, etc.) that surge in popularity in recent years. As result, they suffer from several problems, as we shall discuss in detail in Section 3, that make their use not appropriate for modern OSNs. With this in mind, this paper

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.105430.

* Corresponding author.
E-mail addresses: alabrid@squ.edu.om (D. Al Abri), valaee@comm.utoronto.ca (S. Valaee).

[1] Dawood Al Abri was a visiting scholar at the University of Toronto when this work was commenced.

https://doi.org/10.1016/j.knosys.2019.105430
0950-7051/

presents a first attempt at creating a new model to better capture the influence of user in modern OSNs. Our model differs from previous ones in that it links the influence of user to the typical actions that occur in OSNs, which is not the case in the previous models, where the user influence is typically related to influence probability between different members. Moreover, our model accounts for the possibility that a user may have presence in other social networks.

The main contribution of this paper can be summarized as follows:

- The paper highlights several limitations with current diffusion models that need to be addressed to better evaluate the influence spread in modern OSNs.
- The paper presents a new model that better captures the nature of interactions on modern OSNs and hence it provides a more realistic assessment of user's influence in OSNs.
- With aim of maximizing the impact of a viral marketing campaign across several OSNs, the paper introduces a new *Diversity-Constrained Influence Maximization (DCIM)* problem that aims at identifying the most influential users in a target network while incorporating a diversity constraint that aims at ensuring that information is spread across several social networks.
- The paper describes several greedy algorithms to solve the proposed DCIM problem, analyzes their complexity analytically, and studies their performance under various network conditions using extensive simulations.

The rest of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we discuss the limitations of the current diffusion models. Section 4 presents the proposed model as well as the related optimization problem. In Section 5, we describe few greedy algorithms for tackling the optimization problem and present the complexity analysis for these algorithms. Simulation results are discussed in Section 6. We conclude the paper in Section 7.

2. Related work

The popularity of social networks had motivated researchers to explore different aspects of these platforms (e.g. [6–13]). Morente-Molinera et al. [6] present a novel Group Decision Making process that extracts opinions expressed by expert on online social network to reach the final decision. Muller and Peres [7] looked into the effect of social network structure on innovation growth. In [8], a method to create fuzzy ontologies from the posts of users on online social networks is presented. Feng et al. [14] propose a recommendation system that aims at improving the personalized information presented to the user by exploiting information gathered from the social factors and social circles. Ren et al. [9] investigate the spreading and vanishing of information using a competition model with intervention in online social networks. In [10], the impact of merging social networks with IoT on the spread of malware is investigated. Molano et al. [11] propose a framework to connect IoT object through social networks in order applications for industry 4.0. Another important research area is utilizing the social networks as a medium for digital marketing [12] and an enabler for market-oriented organization [13]. Of particular interest here is exploiting the social networks to market products by propagating the recommendations of product through these social networks. Analysis of such propagation requires the use of diffusion models that are used to model how recommendations and opinions about a product spread in a social network.

The current diffusion models have roots in early sociology studies such as those by Schelling [4] and Granovetter [5] in the 1970's. In Schelling's work [4], a resident is classified as either content or discontent based on his neighborhood color. Simliarly, Granovetter [5] developed a model for persons with two mutually exclusive behaviors. Models based on these ideas were used to study the propagation of influence in social networks. The maximum influence problem was introduced first in [15,16] by Domingos and Richardson where a probabilistic model was used to determine the customers that have large impact on the network. In [17], Kempe et al. formulate the problem as discrete optimization problem and focus on determining an initial set of nodes to maximize the final set of nodes that become active (adapt the idea being spread). They provide an approximation guarantee for efficient algorithms to deal with the influence maximization problem considering several diffusion models. They showed a natural greedy strategy to obtain a solution that is provably within 63% of the optimal for several classes of models. A faster approach to compute the influence function of [17] was proposed in [18], which provides a general framework for selecting nodes, from a graph, to optimize the detection of outbreaks. The authors of [18] have shown that the model used in [17] is a special case of their proposed network outbreak detection problem. Chen et al. [19] proposed speed-up improvement over the original greedy algorithm [17] for the independent cascade and weight cascade models. Borgs et al. [20] proposed an algorithm for the influence maximization problem in nearly optimal time assuming independent cascade model of network diffusion.

In [3], the authors tackle the problem of minimizing the seeding set required to achieve a coverage that reaches a certain given percentage of the total size of the network (number of nodes). They assume that the propagation of information is local (i.e. it reaches within certain d hops from the source). They have shown that the seeding set size is proportional to the size of the network. They have used a simplified version of the linear threshold model. Their model uses an influence factor ρ that determines whether a node becomes active or not. If ρ of a node neighbors become active, then that node will be active in the next round. The number of rounds d (number of propagation hops) is a constraint in their problem setup.

In [21], the authors use a game theoretic approach to develop a heuristic to tackle the influence maximization problem using the linear threshold model. Their approach gives comparable results to other existing solutions but it has a faster running time. The authors in [22] focus on the problem of selecting k users to maximize the influence on a targeted set of users. They used a slightly modified version of the independent cascade model. In [23], Bhattacharya et al. develop and analyze a epidemiological-based mathematical model to study digital marketing in a social network. However, it is not clear how to expand this model when dealing several social networks that have an overlapping presence of users.

In [24], the authors analyze various influence mechanisms to determine the number of buyers of a given product. They model the network as a scale-free graph. They found out that given free product samples to small set of users can influence a larger set of users to purchase the product. In [25], the authors proposed a time-constrained influence maximization problem that seeks to find a seed set that influences the maximum number of users within a given time constrain. They use a modified independent cascade model that accounts for the delay to propagate the influence in the network.

The influence spread when user is present in multiple networks have been considered in [26–30]. In [26], the authors study the information diffusion between physical network and online network using the conventional SIR epidemic model. In [27], Liu et al. focus on studying the characteristics of networks that

include both online and offline interactions such as community structure. They also develop algorithms for event recommendation system and found out that the considering both online and offline interactions provide better metric to recommend users to events. Shen et al. [28] study the problem of identifying the minimum set of users that influence the largest number of common-interest users. They combined the different networks into a single network by creating a single node to represent the user presence in all OSNs in which he participates. Nguyen et al. [29] study the influence maximization in multiple networks by using a coupling scheme to represent the multiple network as single network, which is used to study the influence propagation in these networks. Using a similar coupling approach as [29], Zhang et al. [30] tackle the least cost influence maximization problem when users are present in multiple social networks. The analysis in [28–30] is based on linear threshold model, which suffers from the problems that we have highlighted in the introduction. Moreover, the goal of [30] is to find the minimum seeding set required to achieve a coverage that reaches a certain given percentage of the total size of the combined network ([3] can be considered as special case of [30]). In our work, we seek to find a set of k users that maximizes the marketing reach, which is used as an indicator for the user influence. In our formulation, we also have diversity constraint to ensure that information is spread over different OSNs.

All of these works use the two basic models described in the introduction with some modifications to study a particular variation of the influence maximization (e.g. limit propagation to a certain number of hops). Consequently, the obtained results may not be the best reflection of what actually occurs in OSNs due to problems that are described next.

3. Limitations of current diffusion models

As mentioned in the previous section, the current models have sociology origins. Consequently, using these models to analyze online social networks may not capture the interactions that occur on the online space accurately as these models were developed based the interactions that occur on ordinary face-to-face interactions. In this section, we discuss some of the limitations of the current models when it comes to analyzing online social networks.

Looking at the current models in the context of product marketing, they generally use binary classification of users: either the user will buy the product (active) or he will reject it (inactive). However, such binary classification is problematic to adopt for online social networks. To clarify this, let us consider human interaction and assume that person A is attempting to sell something to person B. By the end of their conversation, it may be easy to conclude whether person B will buy the product or not. Note, however, that it is certainly possible that B is still hesitant and a simple binary classification is not suitable (nonetheless, in current models, B must be classified to one category). On the other hand, if person A is disseminating his marketing message through an online social media network, then classification of followers (i.e., users who receive this message) as either active or inactive may not be as simple as the previous direct interaction. There is no telltale sign that a follower will buy the product or not.

Moreover, in ordinary human society, the influence of a person is related to the size of his community and how well is he known in his community. However, nowadays, due to the presence of people in multiple online social networks, it is misleading to judge the influence of a person based on his presence in one OSN. For example, consider a user A with small number of friends on Facebook (see Fig. 1). If A posts an opinion piece in his Facebook page and his friends share a link to his post via their other OSN accounts, then the number of users who read this post could potentially be very large if A's friends have strong presence in other OSNs. Hence, based only on the user's Facebook presence, A is not influential but taken into consideration the sharing to other OSNs (via friends), A may be more influential than other users in Facebook. Add to that, the number of friends in ordinary social network is typically small, while in OSNs, having followers or subscribers in the hundreds is not something uncommon. In addition to that, in ordinary human social network, the friendship relationship is mostly symmetric (A is a friend of B and B is a friend of A). However, in OSNs, that is not the case; A may subscribe (or follow) B but B may not follow back. Even if we try to adopt the current models to OSNs by, say, using a directed graph for each OSN that the user is present in, we will run into difficulties. For example, because there is no mechanism in some OSNs to propagate the content further in the same OSNs (e.g. if you receive a notification about new YouTube video, you cannot re-post it to your subscriber directly). Hence, if you use the cascade model, the propagation will be limited to one-hop from the source. Similarly, what does it mean to use the threshold model in this case? Does it mean that user receives certain number of posts about the product from the group to which he subscribed? How practical is that if he subscribes to a large number of users? We argue here that it would be more natural to capture the interactions in the modern OSNs using a new model rather than attempting to adopt the existing models.

Another problem that we see with current models is that they have the notion of influence probability, which we believe is quite challenging to estimate practically. Estimating the influence probability that user A have over B is not trivial as it may require knowing the history of interaction between these users and how many times did A manage to "convince" B about certain issue. Based on influence probability, the user will be classified as active or inactive. However, in our model, we are focusing on marketing reach, which is how many users will eventually read (or see or view) the post. We are not classifying the user as active or inactive. Our view is that in OSNs, it is not realistic to classify the user in a binary fashion. In marketing context, how is it possible to say a follower of A is an active (e.g. will buy the product) or inactive (will not buy the product). It is more feasible, through the current tracking technique, to determine whether the followers view the post and/or share that post. Note also, that in the IC model, there is one probability that is related to the interaction of two users A and B which is the probability that A will activate B. In our case, we have two probabilities: (i) the probability that B reads A's post and (ii) the probability the B shares the post of A with his own followers in other OSNs. The main advantage of our model is that it relates more realistically to the actions that a typical user may do in an OSN, which can be easily estimated given the ability of current OSNs to track users' activities when they are logged into the OSNs. These tracking techniques are becoming so sophisticated to the degree that Facebook recently announced that it will track users even if they do not have a Facebook account [31]. A description of tracking approaches as well as tracking statistics in the top one million sites can be found in [32]. In the next section, we describe our proposed model, which aims to better capture the interactions that occur on modern online social networks and address the limitations of the current models.

4. Model and optimization problem

4.1. Model

Online social networks (OSNs) provide platforms for interaction between users. Typically, a user creates an account before

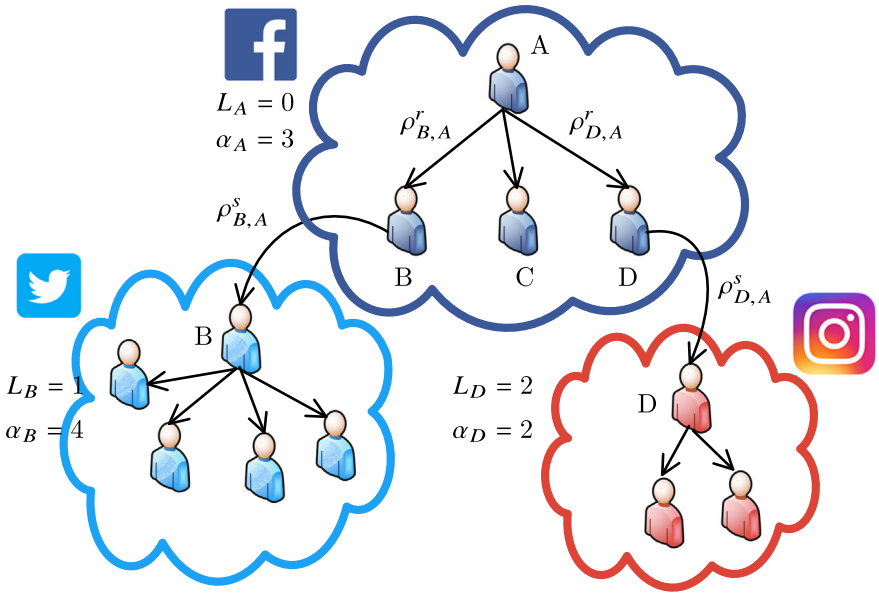

Fig. 1. Illustration of information propagation via sharing to other social networks. User *A* makes a post in Facebook. His friends *B* and *D* read it and share it to their other social networks (Twitter and Instagram, respectively).

s/he starts disseminating his or her own content, which is site specific (e.g. videos, short messages, photos, etc.). Majority of social networks allow users to follow (or subscribe to) other users' accounts. Whenever a user produces a new content, followers are notified or receive the new content (e.g., a new tweet is received by all of the user followers). Moreover, an increasing number of these OSNs allow users to share content with their social circles in other OSNs via some communication means (e.g. email) or posting directly to other OSNs.

Basic notation and terminology: Let $\mathbb{I} = \{0, 1, 2, \ldots, m\}$ be the set of identifiers for the online social networks. An identifier is an integer that corresponds to a particular online social network (e.g. 0 = Facebook, 1 = Twitter, etc.). We use $\mathbb{OSN}^j$ to denote the set of users who belong to the OSN with identifier j. An advertiser selects one of these OSNs to be the target of its advertising campaign (i.e. the initial seed of users will be selected from this targeted network). We refer to this selected OSN as the *targeted OSN (TOSN)* (without loss of generality, we assume that the identifier of TOSN is 0). We use TOSN as both a term to refer to the selected OSN by the advertiser as well as an identifier of the selected OSN (the intended meaning should be clear from context). Moreover, we assume a universal ID for users. In other words, a user is identified by the same ID across all the OSNs (see [29] for an efficient method to assign universal IDs). For any user $i \in \mathbb{OSN}^{TOSN}$, we refer to the set of OSNs, other than TOSN, in which i is a member as *User Other Network (UON)* (i.e., UON of i is $UON(i) = \{j : j \in \mathbb{I} \setminus TOSN, i \in \mathbb{OSN}^j\}$). The set of followers (subscribers, friends) of user i in $\mathbb{OSN}^j$ is denoted by $F(i)^j$. We represent the social network as a directed graph $G = (V, E)$ where the vertices represent the users and the directed edges represent the direction of information flow (i.e. from user to its followers or from a user to its other social networks). Note here that a user may have more than one UON but for simplicity, we are assuming that each user may have up to one UON. Our analysis can be easily extended to accommodate the case of multiple UONs per user. We consider one-hop propagation within a given OSN. According to [3], there is little difference when the authors compare the cost-effectiveness of marketing campaign for 2, 3, and 4-hop propagation cases. Hence, restricting our analysis to one-hop should not affect the overall conclusion greatly. To clarify these terms, suppose that a company wants to advertise its product via YouTube (e.g. give a freebie to YouTuber to make a product review). A user **A** sees this video and shares a link to it via his Twitter account. In this example, YouTube is the TOSN while Twitter is the UON of user **A**. Linking a user profile in one network with the same user profile in another network is beyond the scope of this paper (see [33,34] for approaches to do the linking).

The UON of each user is characterized with a tuple (α_i, L_i) where α_i represents the size of that network and L_i is the identifier of UON (i.e. $\alpha_i = |F(i)^{L_i}|$ where $L_j = \{j : j \in \mathbb{I} \setminus TOSN, i \in \mathbb{OSN}^j\}$). With each user j, we associate two probabilities:

- $\rho^r_{j,i}$ the probability that the follower j reads the information (i.e. ad in our case) generated by another user i. Note that this probability is zero if j does not follow i.
- $\rho^s_{j,i}$ the probability that the follower j shares the information (i.e. ad in our case) generated by another user i given that the user had already read it. As $\rho^r_{j,i}$, $\rho^s_{j,i}$ is zero if j does not follow i.

We also define the following two sets for each user in TOSN:

- $F(i)$ the set of followers, i.e. set of users in TOSN who receive the content generated by user i (we also refer to this set as the children of i). This is basically $F(i)^{TOSN}$ but, for simplicity, we drop the superscript and use $F(i)$ instead.
- $P(i)$ the set of parents, i.e. set of users that user i is following in TOSN.

We note here that the information flow is controlled by the online social network logic. In our follower-based network model, whenever the user makes a post, it will be received by all of its followers.

Marketing reach: In marketing terminology, the *marketing reach* refers to the number of people who are exposed to the ad. This has nothing to do with how many people took action based on their exposure to the ad (note this is different from other models where they focus on *influencing* others). For example, if user A has 100 followers and he posted an ad of a certain product to his followers. Now, if 50 of his followers read (or saw) the ad, then we say that his marketing reach is 50 (the ad reached 50 followers). On the other hand, if 10 out of the 50 followers, who

read/saw the ad, were influenced by his product ad and bought the product, then we would quantify the influence spread as 10. It should be clear that quantifying how many followers read/saw the post is more practical compared to attempting to determine how many are influenced? The former depends on the actions taken by users on the OSNs, which is easy to track, while the later depends on what the user internalized about the ad. We are proposing to use the value of the marketing reach of a user as an indicator for the influence of that user. Hence, maximizing the influence will be equivalent to maximizing the marketing reach. Toward that goal, we would like to identify a set of users that will maximize the spread of the ad within the TOSN and to the UONs via sharing by users in the TOSN who received the ad. Each user will be given a weight to indicate its contribution in spreading the ad. We define the marketing reach $\sigma(i)$ of user i by the following equation:

$$\begin{aligned}\sigma(i) &= \alpha_i + \sum_{j\in F(i)} \left(\rho^r_{j,i} + \rho^r_{j,i}\cdot\rho^s_{j,i}\cdot\alpha_j\right)\\ &= \alpha_i + \sum_{j\in F(i)} \beta_{j,i}\end{aligned} \tag{1}$$

where $\beta_{j,i} = \rho^r_{j,i}\left(1+\rho^s_{j,i}\cdot\alpha_j\right)$. The reach of a node is quantified by two terms:

- The first term α_i characterizes the size of UON that user can share information with. We are assuming here that if a user is selected to be a "promoter", s/he will share the content with his or her other OSNs (i.e. $p^s_{i,i} = 1$)(maybe for extra incentive).
- The second term sums up the contribution of the followers to user i. Collectively, the summation term represents the expected number of followers who will read the ad disseminated by that user as well as the expected number of users in the followers' UONs who may receive it via the user's followers who decide to share it with their UONs. For a follower j with large number of subscribers in its UON, $\beta_{j,i} \approx \rho^r_{j,i}\cdot\rho^s_{j,i}\cdot\alpha_j$ which means that the main contribution of j to its parent is channeling the information to its UON (α_j, L_j).

Thus, the marketing reach of a user is not only related to the number of his followers who read the ad but it also incorporates the size of other social networks that he may leverage to spread the word via his followers who chose to share the ad with their UONs.

Collective marketing reach: We now turn our attention to defining the collective marketing reach of a set of users S. Obviously, the sizes of their UONs and the expected marketing reach via their followers will be factors in determining their collective influence. More specifically, if S represents the selected set of users (we call it *the marketing seed*) to spread the ad, which does not have any parent–child pair among them (the reason for this will be clarified shortly), and

$$U = \bigcup_{m\in S} F(m) \tag{2}$$

is the union of their followers, then we define the collective influence by:

$$\Phi'(S) = \sum_{j\in S}\alpha_j + \sum_{m\in U}\rho^r_m\left(1+\rho^s_m\cdot\alpha_m\right) \tag{3}$$

It has been assumed here that the sets of followers in other networks for users i and j in the seed set are disjoint. Here, ρ^r_m is the probability that a follower m will read the ad if it receives the ad from multiple parents who are in the marketing seed. Assuming that the probability of reading from each parent is independent from each other, we can compute ρ^r_m as follows:

$$\rho^r_m = 1 - \prod_{i\in S\cap P(m)}\left(1-\rho^r_{m,i}\right) = 1-\pi^r_m \tag{4}$$

where

$$\pi^r_m = \prod_{i\in S\cap P(m)}\left(1-\rho^r_{m,i}\right) = \prod_{i\in\mathbb{OSN}^{TOSN}}\left(1-\rho^r_{m,i}\right)^{x_i} \tag{5}$$

and x_i indicates whether user i belongs to the seed set S (1) or not (0). Note that, as mentioned previously, $\rho^r_{m,i}$ is zero if i is not a parent of m.

In a similar way, ρ^s_m is the probability that the node m will share the ad with its other social networks if it receives it from multiple parents who are in the initial seed, i.e.,

$$\rho^s_m = 1-\pi^s_m \tag{6}$$

where

$$\pi^s_m = \prod_{i\in\mathbb{OSN}^{TOSN}}\left(1-\rho^s_{m,i}\right)^{x_i} \tag{7}$$

Fixing the parent–child dependence problem: There is one issue related to the definition of $\Phi'(S)$ that we need to fix to extend this definition to an arbitrary set of users S. The definition does not account for the dependence between parent and child that is manifested in the term $\beta_{j,i}$ that each child j contributes to its parent i. To illustrate the problem in a concrete way, assume that j is a follower of i and that the α_j is quite large. Hence, the term $\beta_{j,i}$ contributes considerably to the reach of the parent $\sigma(i)$. The problem will occur if both i and j are selected in the marketing set. Since j is selected, we are assured that we can reach the other social network of the user (α_j, L_j) directly via j. However, the value of this network reach (with some scaling) is also counted in the parent reach function (the term $\rho^r_{j,i}\cdot\rho^s_{j,i}\cdot\alpha_j$). Furthermore, if that term is removed from the parent reach (since the network (α_j, L_j) is reachable directly through the child), the parent reach may become too low and we are better off by replacing him with another user. Note, however, if the child j were not in the marketing set, then we needed to incorporate the term $\beta_{j,i}$ since the network (α_j, L_j) can be reached only via the parent (indirectly by passing the ad to its followers including j). To fix this parent/child dependence, we introduce the following corrective term:

$$\Phi''(S) = \sum_{i\in S}\sum_{j\in S}\beta_{j,i}\cdot x_i\cdot x_j \tag{8}$$

Observe that $\Phi''(S)$ will consider the contribution $\beta_{j,i}$ when both parent i and the child j are present $(x_i = x_j = 1)$ where we define $\beta_{i,i} = 0$. Note also that $\beta_{j,i} = 0$ if i and j do not have parent–child relationship since in that case $\rho^r_{j,i} = 0$. Based on this consideration, we define the marketing reach $\Phi(S)$ for an arbitrary set S as:

$$\Phi(S) = \Phi'(S) - \Phi''(S) \tag{9}$$

Note that Eqs. (3) through (9) define a non-linear set of equations that make finding the set S that maximizes $\Phi(S)$ impractical for large social networks. Moreover, the parent/child dependence makes most terms vary with the particular choice of S, which prevent us from pre-computing the values of the terms to speed-up the calculation.

An approximation for the collective marketing reach: To overcome the difficulties mentioned in the previous subsection, we will make an approximation to be able to compute the set S more efficiently. Since a user is likely to receive the information from several parents that have different likelihood to entice the

user to share with its other social network, we will link each user with its most influential parent in the sense that the probability that the user shares the received content with its social network is the highest among all parents of that user. More specifically, we define the influential parent $\overline{P}(j)$ of user j as follows:

$$\overline{P}(j) = \arg\max_{i \in P(j)} \rho^r_{j,i} \cdot \rho^s_{j,i} \tag{10}$$

Due to this linkage, each parent i will have a set of children that are highly influenced by that parent, which we denote by $\overline{F}(i)$, and can be defined formally as:

$$\overline{F}(i) = \{j \in F(i) : \overline{P}(j) = i\} \tag{11}$$

Based on the new linkage between the children and their most influential parents, we define an adjacency matrix $A = [a_{ji}]$ to indicate whether a child j is linked with parent i as follows:

$$a_{ji} = \begin{cases} 1 & \text{if } \overline{P}(j) = i \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Since each node j has at most only one influential parent, each row of A has only one 1 at most and the rest of entries are zero. Based on this new formulation, the value of the contribution of a follower in the union set U will be determined by its most influential parent only. In other words, both π^r_m and π^s_m will depend on the most influential parent $\overline{P}(m)$, i.e.

$$\rho^r_m = \rho^r_{m,\overline{P}(m)} \tag{13}$$

and ρ^s_m can be defined similarly. Consequently, $\Phi'(S)$ will become simply the sum of marketing reach $\sigma(i)$ for all the users in the set S (each child is influenced by one parent only). Despite this simplification in calculation of $\Phi'(S)$, the issue of parent/child dependence still exists in the new formulation. Nonetheless, the new formulation allows us to pre-compute the marketing reach of each user as well as the corrective term related to each parent–child pair; thereby, enabling computation speed-up. This can be cast in the form of a profit matrix $\Theta = [\vartheta_{ji}]$ with diagonal elements representing the gain from adding a particular user and the other elements act as corrective terms to adjust the contribution of the UON added to parent i from its child j. The variable ϑ_{ji} is defined as follows:

$$\vartheta_{ji} = \begin{cases} \sigma(i) & \text{if } i = j \\ -a_{ji} \cdot \beta_{j,i} & \text{if } i \neq j \end{cases} \tag{14}$$

Here, the terms $\sigma(i)$ and $\beta_{j,i}$ are calculated based on the follower set $\overline{F}(i)$. Hence, the collective marketing reach $\widetilde{\Phi}(S)$ of a set S becomes:

$$\widetilde{\Phi}(S) = \sum_{i \in S} \sum_{j \in S} \vartheta_{ij} \cdot x_i \cdot x_j \tag{15}$$

where we use $\widetilde{\Phi}(S)$ to indicate the marketing reach of the set S under the new formulation to avoid confusion with marketing reach $\Phi(S)$ under the original formulation.

Diversity factor: Another aspect that we want to take into account is the desire of a marketer to ensure that the ad reaches different UONs via sharing. For example, the marketer may target Facebook but may want to make sure that the ad reaches at least 5 different UONs via sharing from Facebook users to their other social networks. We refer to this requirement as the *diversity factor (DF)*, which indicates the number of different UONs that the ad is required to reach via sharing from the TOSN (i.e. $DF = |\bigcup_{j \in S} UON(j)|$). Moreover, the marketer may want to reach a specific UON (e.g. Twitter) via sharing from the TOSN (e.g. Facebook). We refer to this requirement as the *targeted diversity set (TDS)*, which is specified by a list of the types of UONs that marketer wants specifically to reach via sharing by the selected marketing seed and their followers (i.e. a list of L_m's values of the desired UONs). Obviously, such diversity requirement will influence which users will be selected in the seed set. For example, user i may have low number of friends in Facebook but large number of followers in Twitter. If the marketer is targeting Facebook but have Twitter in its *TDS*, user i may be more preferable over user j who has a larger number of friends in Facebook but no presence in Twitter.

4.2. Optimization problem

Let $\mathbb{N} = 1, 2, 3, \ldots, n$ be the set of all users in the targeted online social network and x_i indicate whether user i is selected (1) or not (0). We formulate the following *Diversity-Constrained Influence Maximization (DCIM)* problem:

$$\begin{aligned} &\text{maximize}_{x_i} \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{N}} \vartheta_{ij} \cdot x_i \cdot x_j \\ &\text{subject to} \\ &\qquad \sum_{i=1}^{n} x_i = k \\ &x_i \in \{0, 1\}, \quad i \in \mathbb{N} \\ &\qquad |\bigcup_{x_m=1} L_m| \geq DF \\ &\qquad TDS \subseteq \bigcup_{x_m=1} L_m \end{aligned} \tag{16}$$

In other words, we want to select a subset of k users $S \subset \mathbb{N}$ to maximize $\widetilde{\Phi}(S)$ with diversity requirement dictated by the last two constrains. The first diversity requirement ensures that the number of different UONs that the ad should reach via sharing from the target social network should be at least DF. The second diversity requirement specifies the set of UONs (TDS) that should be specifically targeted. Here, we are assuming without loss of generality that $|TDS| \leq DF$. If $|TDS| > DF$, the requirement regarding DF is redundant as it will be automatically satisfied once the TDS requirement is satisfied. If there is no diversity requirement imposed (i.e. $DF = 0$ and $TDS = \phi$), this problem reduces to Quadratic Knapsack Problem (QKP), which is NP-hard in the strong sense and hence our problem is an NP-hard as well since it is a generalization of QKP. Moreover, it is still an open research problem to determine whether or not a constant approximation ratio for QKP is possible [35].

5. Greedy algorithm

In this section, we present a greedy algorithm for the DCIM problem, which we refer to as *Greedy with Diversity Enforced at End* (GDEE). To highlight how the selection criteria that takes into account the users presence in other social networks result in better seed set, we use the popular degree centrality metric (i.e. number of followers in our case) that is typically used to determine the most influential nodes in social network [19] as a basis for comparison. A naive use of this metric would be to pick the top users in terms of their number of followers. However, this is not optimum in OSNs since there may be overlap between the followers of the top users. A better approach is to focus on the union set of the followers as a metric for selecting the users as described in Algorithm 1 below.

Algorithm 1 Greedy based on Number of Followers (GNF): The goal of this algorithm is to select the set of users that results in the largest number of followers. To add a new user to the marketing set, the algorithm examines the size of the union set of followers of the marketing seed due to the addition of each user

Algorithm 1 Greedy based on Number of Followers (GNF)

Input: Target seed set size N
Output: Seed set S
1: Initialize the seed set $S = \phi$
2: **while** $|S| \neq N$ **do** ▷ loop while the required size of seed set S is not achieved
3: $u^* \leftarrow \arg\max_u |\bigcup_{m \in S \cup \{u\}} F(m)|$ ▷ find the user whose addition maximizes the total number of followers of the seed set
4: $S \leftarrow S \cup \{u^*\}$
5: **end while**

to the marketing seed and picks the user u^* that results in the maximum increase. This process is repeated until the required size of the seed set is achieved. Note here that the GNF is an improvement over the popular degree centric based selection algorithm. To see this, consider the case where users i and j are the top users in terms of their number of followers. A basic scheme that is based purely on the degree metric (i.e. the number of followers in this case) would pick these two users in its seed set. However, if i and j have the same followers sets, then one of them would be sufficient to reach this group of followers. Obviously in this particular example, it would be better to choose either i or j with another user (say m) who has a different group of followers. Having i (or j) with m in the seed set will ensure that the ad is spread to a larger audience.

Algorithm 2 Greedy based on Expected Number of Followers who read the content (GENF): This is a closely related variation of GNF. GENF selects the user that results in the maximum increase on the *expected* number of followers who read the content. We use the notation $EF^r(m)$ to refer to the expected number of followers who read the content posted by user m (see the discussion related to Eq. (1) for how to calculate $EF^r(m)$).

Algorithm 3 Greedy with Diversity Enforced at End (GDEE): This algorithm computes the increase in the marketing reach that is achieved by adding a user i to the existing seed set (i.e. $\widetilde{\Phi}(S \cup \{i\}) - \widetilde{\Phi}(S)$) and picks the user u^* that results in the largest increase. The algorithm delays the enforcement of the diversity requirement to the very end in the hope that some of these requirement will be met in the early iterations. This gives the algorithm the freedom in the early iterations to pick the users that result in the largest increase in $\widetilde{\Phi}$. However, when the number of users remaining to meet the required seed set size is just enough to satisfy the diversity requirement (line 6), the algorithm will ignore all the users whose addition will not result in meeting the diversity requirement (lines 7–9). For example, if the required diversity factor is 6 and we had achieved a diversity factor of 5 (i.e. marketing seed set contain users that have five different UONs types), then when there is one user remaining to be selected, the algorithm will not consider any user that will not result in increasing the diversity factor to 6 (i.e. it should have a different UON type than the five already covered by the marketing seed).

Algorithm 4 Enhanced GDEE (EGDEE): We will now exploit the properties of $\widetilde{\Phi}(S)$ to speed up the GDEE algorithm. First, observe that in step 11 of GDEE, we compute the function $\widetilde{\Phi}(S)$ twice to determine the increment due to adding the user i. From (15), we observe that the change in $\widetilde{\Phi}(S)$ due to the addition of a new user i will be due to the terms that relate this new user with existing ones. Hence, the change in $\widetilde{\Phi}(S)$ can be expressed as:

$$\widetilde{\Phi}(S \cup \{i\}) - \widetilde{\Phi}(S) = \vartheta_{ii} + \sum_{j \in S} (\vartheta_{ij} + \vartheta_{ji}) \cdot x_i \cdot x_j \tag{17}$$

Algorithm 2 Greedy based on Expected Number of Followers (GENF) who read the content

Input: Target seed set size N
Output: Seed set S
1: Initialize the seed set $S = \phi$
2: **while** $|S| \neq N$ **do** ▷ loop while the required size of seed set S is not achieved
3: $u^* \leftarrow \arg\max_u |\bigcup_{m \in S \cup \{u\}} EF^r(m)|$ ▷ find the user whose addition maximizes the total expected number of followers who read the content
4: $S \leftarrow S \cup \{u^*\}$
5: **end while**

Algorithm 3 Greedy with Diversity Enforced at End (GDEE)

Input: Target seed set size N and target diversity factor DF
Output: Seed set S
1: Initialize the seed set $S = \phi$
2: **while** $|S| \neq N$ **do** ▷ loop while the required size of seed set S is not achieved
3: $inc^* \leftarrow 0$
4: $R \leftarrow N - |S|$ ▷ number of remaining users to achieve the target seed set size
5: **for** $i \in \mathbb{N} \setminus S$ **do** ▷ loop through all users who do not belong to the current seed set
6: **if** $R = DF$ **then** ▷ check if the number of remaining users to be selected is just enough to satisfy the diversity requirement
7: **if** $|\bigcup_{j \in S \cup \{i\}} UON(j)| = |\bigcup_{j \in S} UON(j)|$ **then** ▷ check if user i will not improve the diversity requirement
8: **continue** ▷ skip this user
9: **end if**
10: **end if**
11: $inc \leftarrow \widetilde{\Phi}(S \cup \{i\}) - \widetilde{\Phi}(S)$ ▷ compute the increase in marketing reach due to the addition of user i
12: **if** $inc > inc^*$ **then**
13: $inc^* \leftarrow inc$
14: $u^* \leftarrow i$
15: **end if**
16: **end for**
17: $S \leftarrow S \cup \{u^*\}$
18: **end while**

Observe that maximum increase in $\widetilde{\Phi}(S)$ due to the addition of user i is $\vartheta_{ii} = \sigma(i)$. This can be exploited to speed up the GDEE algorithm in the following way. We first sort the users in descending order in terms of their $\sigma(i)$. This ordering will be used in the **for** loop of the GDEE algorithm to evaluate the increment ***inc*** in $\widetilde{\Phi}(S)$ due to different users starting with the user that has the highest $\sigma(i)$. Moreover, in each iteration of the **for** loop, we test whether the maximum increment found so far inc^* is greater than $\sigma(i)$ of the user being considered in that loop iteration. If that is the case, then we can break the loop. The logic behind aborting the rest of the loop is that we will not be able to find a larger increment since the users are sorted in descending order based on their σ's values and we have already found a user i that satisfies $inc^* \geq \sigma(i) \geq \sigma(j)$ for any user j that comes after user i in the sorted list. We refer to the GDEE algorithm with these speed-up improvements as *Enhanced GDEE* (EGDEE) (Algorithm 4).

A similar trick can be used with GNF and GENF by sorting the users in terms of their number of followers for GNF and expected number of followers who will read the ad for GENF.

Algorithm 4 Enhanced GDEE (EGDEE)

Input: Target seed set size N and target diversity factor DF
Output: Seed set S

Initialize the seed set $S = \phi$
$\mathbb{N}^* \xleftarrow[\text{order in term of } \sigma(i)]{\text{Sort in descending}} \mathbb{N}$
while $|S| \neq N$ **do** ▷ loop while the required size of seed set S is not achieved
  $inc^* \leftarrow 0$
  $R \leftarrow N - |S|$ ▷ number of remaining users to achieve the target seed set size
  for $i \in \mathbb{N}^* \setminus S$ **do** ▷ loop through sorted set (in terms of $\sigma(i)$) of users who do not belong to the current seed set
    if $\sigma(i) < inc^*$ **then**
      break ▷ the current user under consideration has $\sigma(i)$ lower than the maximum increase found so far and hence there is no need to check the remaining users as their $\sigma(i)$'s will be lower (due to the use of the sorted set)
    end if
    if $R = DF$ **then** ▷ check if the number of remaining users to be selected is just enough to satisfy the diversity requirement
      if $|\bigcup_{j \in S \cup \{i\}} UON(j)| = |\bigcup_{j \in S} UON(j)|$ **then** ▷ check if user i will not improve the diversity requirement
        continue ▷ skip this user
      end if
    end if
    $inc \leftarrow \vartheta_{ii} + \sum_{j \in S} (\vartheta_{ij} + \vartheta_{ji}) \cdot x_i \cdot x_j$ ▷ compute the increase in marketing reach due to the addition of user i
    if $inc > inc^*$ **then**
      $inc^* \leftarrow inc$
      $u^* \leftarrow i$
    end if
  end for
  $S \leftarrow S \cup \{u^*\}$
end while

We briefly describe the trick for GNF (similar reasoning can be applied for GENF). The users should be considered in this sorted order and whenever the maximum increment found is greater than the number of followers for the next user being considered for addition to the seed set for GNF, we can terminate the search as it would not be possible to find a greater increment. The increment is at most equal to the number of the followers for the user being added (equality is achieved when the followers of the user do not overlap with those in the union set of the followers of the users who have been already selected).

Upper Bound for Marketing Reach (UPMR): An upper bound for the marketing reach can be obtained by observing that $\Phi(S) \leqslant \Phi'(S)$ (Eq. (9)). For $|S| = k$, the maximum value for $\Phi'(S)$ will be attained if S contains the top k users (in terms of $\sigma(i)$) and these users do not have any overlap between their followers (i.e. $F(i) \cap F(j) = \phi$ for any $i, j \in S$). Hence, ρ_m^r and ρ_s^s will represent the probability of reading and sharing that relates a single user in S and one of its followers. As a result, $\Phi'(S)$ will simply be the sum of the marketing reach $\sigma(i)$ of the users in S. In general, if $\sigma_s(i)$ is sorted in descending order, then the upper bound based on k users can be expressed as:

$$\text{UPMR} = \sum_{i=1}^{k} \sigma_s(i) \tag{18}$$

Complexity Analysis: To assess the computational complexity of the proposed algorithm, let n_p and n_f be the maximum number of parents and followers that a user can have, respectively. We first analyze the computational complexity for functions $\sigma(i)$, $\Phi(S)$, and $\widetilde{\Phi}(S)$:

- $\sigma(i)$: since the calculation is repeated for every follower of the user, the computational complexity will be $\mathcal{O}(n_f)$.
- $\Phi(S)$: For $\Phi'(S)$, the first summation term has a complexity of $\mathcal{O}(|S|)$. To determine the complexity of the second summation term of $\Phi'(S)$, observe that we have $|S|n_f$ followers in the union U at most, which results in complexity $\mathcal{O}(|S|n_f n_p)$ (cost of computing ρ_m^r and ρ_m^s is $\mathcal{O}(n_p)$). For $\Phi''(S)$, the cost is $\mathcal{O}(|S|^2)$. Typically, the seed set size $|S|$ is much smaller than the maximum number of parents n_p and followers n_f. Hence, the complexity of calculating $\Phi(S)$ is $\mathcal{O}(|S|n_f n_p)$.
- $\widetilde{\Phi}(S)$: The determination of the most effective parent will involve evaluating all parents for each user, which has at most n_p parents and hence the cost will be $\mathcal{O}(nn_p)$. The determination of the adjacency matrix coefficients a_{ij} has a cost of $\mathcal{O}(n)$. The computation of the diagonal elements of Φ matrix has a cost of $\mathcal{O}(n_f)$ and for off-diagonal elements, the cost is $\mathcal{O}(1)$. Hence, the total cost for computing Φ is $\mathcal{O}(nn_f + n^2 - n)$. The previous computations will be done once with a total cost of $\mathcal{O}(n(n + n_p + n_f))$. With pre-computed values for Φ at hand, $\widetilde{\Phi}(S)$ will cost $\mathcal{O}(|S|^2)$.

Now, let us analyze the complexity of the algorithms presented at the beginning of this section:

- **GNF:** For each iteration, the algorithm examines each user in the TOSN to determine the one that results in the largest increase in the number of followers. To do this, the algorithm computes the union of user's followers and the union set of all the followers of previously selected users. A linear approach to the set union will mean that in each iteration, the cost of computing the union between user with n_f followers at most and union set that may have at most $(i-1)n_f$ members if we consider iteration i is $\mathcal{O}(in_f)$ with a worst case value when we have selected the last user (i.e. $\mathcal{O}(|S|n_f)$). This needs to be repeated for n users to determine which one will result in the highest increase. Hence, the complexity of GNF is $\mathcal{O}(|S|nn_f)$.
- **GENF:** The only difference from GNF is that we need to compute ρ_m^r, which costs $\mathcal{O}(n_p)$. Hence, the overall cost of GENF is $\mathcal{O}(|S|nn_f n_p)$.
- **GDEE:** For each of the $|S|$ iterations of the **while** loop, the algorithm computes $\widetilde{\Phi}(S)$ twice to determine the increment inc per candidate user, which costs $\mathcal{O}(|S|^2)$. Hence, the cost per iteration is $\mathcal{O}(|S|^2 n)$ and the overall complexity of GDEE is $\mathcal{O}(|S|^3 n)$. It is worth noting here that the complexity of the algorithm would be $\mathcal{O}(|S|^2 nn_f n_p)$ if we were to use $\Phi(S)$ instead of $\widetilde{\Phi}(S)$.
- **EGDEE:** This is similar to GDEE but involves sorting the users in terms of their σ's value (cost $\mathcal{O}(n \log n)$). In addition, the cost of evaluating the increment in $\widetilde{\Phi}(S)$ is $\mathcal{O}(|S|)$. Hence, the loop will cost $\mathcal{O}(|S|n)$ per user. Therefore, the overall cost is $\mathcal{O}(n(\log n + |S|^2))$. Note that, as we have explained previously, the algorithm will run faster than the worst case cost obtained here due to premature termination of the loop since in a real social network, small percentage of users have large follower base which in turn means that the larger number of users may not need to be examined as they do not have large marketing reach.

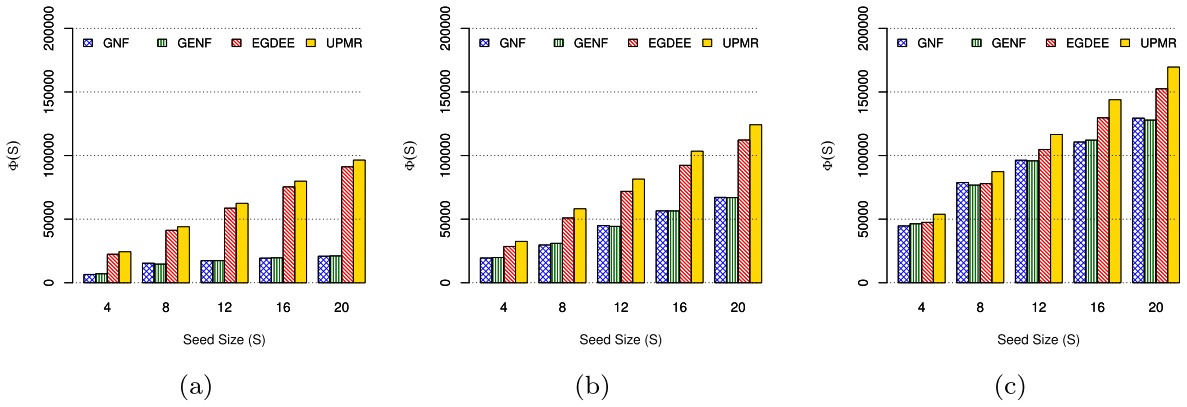

Fig. 2. Marketing Reach for the different algorithms when the percentage of users who have UON with sizes uniformly distributed between 1 and 5k are: (a) 1% (b) 5% and (c) 10%.

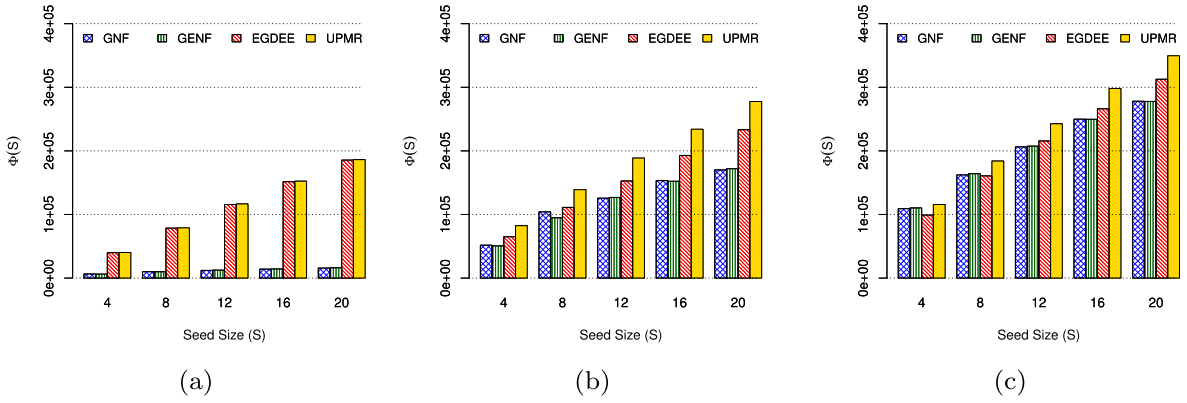

Fig. 3. Marketing Reach for the different algorithms when the percentage of users who have UON with sizes uniformly distributed between 1 and 10k are: (a) 1% (b) 5% and (c) 10%.

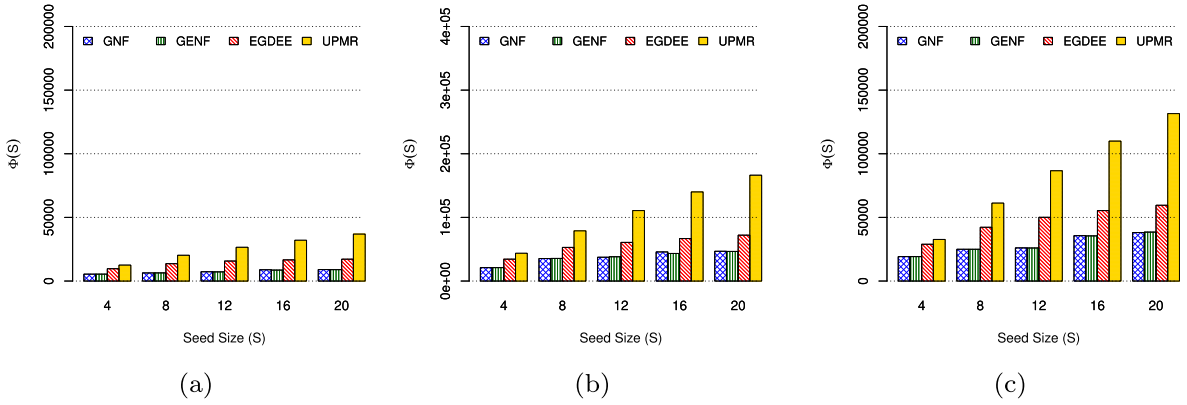

Fig. 4. Marketing Reach for the different algorithms for Facebook–Twitter dataset when the percentage of Facebook users who have UON are: (a) 1% (b) 5% and (c) 10%.

6. Simulation results

6.1. Simulation setup

To evaluate the proposed algorithms, we use the following datasets:

- **Dataset0:** This is the base dataset and used for all simulation unless otherwise stated. we generated target networks using the Barabási–Albert model [36]. The generated network size consists of 10k users. In previous work that analyzed diffusion models (e.g. [17,19,30]), unknown parameters (e.g. influence probability, weight of graph edge) are assigned values from a uniform random distribution between 0 and 1. Following this, the values of read and share probabilities are assigned values from a uniform distribution between 0 and 1. Certain percentage of the total number of users have UONs (we denote this percentage by ρ). The sizes of these UONs (i.e. α_i's) are generated using a uniform random distribution between 1 and β. This is used to simulate scenarios where diversity is of no concern (and hence UON type is the same for all users who have UON).

Table 1
Running time in milliseconds of the various algorithms for different seed set size. T_{parent} is the time to compute the most influential parent for each user and T_{profit} is the time to compute the profit matrix Θ ($\beta = 10k$, $\rho = 1\%$).

Seed size	GNF	GENF	T_{parent}	T_{profit}	GDEE	EGDEE	UPMR	Original
4	2489	2484	186	2890	6	10	2492	22 769
8	2509	2504	185	2895	14	11	2513	78 895
12	2510	2507	182	2894	26	11	2509	166 906

Table 2
Marketing reach of the various algorithms for different seed set size ($\beta = 10k$, $\rho = 1\%$).

Seed size	GNF	GENF	EGDEE	UPMR	Original
4	6388	6171	39 890	40 187	40 170
8	9923	9853	78 477	78 967	78 934
12	12 060	12 442	115 641	116 547	116 487

- **Dataset1:** Similar to Dataset0 with $\alpha = 1\%$ but UON types are different to investigate diversity impact. Here, the types of UONs (i.e. L_i's) are chosen between 1 and 12 (inclusive) according to the following distribution: 45% of these UONs are set to type 1 while the remaining ones are assigned types 2 to 12 randomly and in equal proportion (i.e. for types 2 to 12, each represents 5% of total number of UONs). The size of each UONs is uniformly chosen between 1 and 10k. This dataset allows us to examine the impact of diversity constraint when there is an UON that is more popular than the others (i.e. UONs of type 1 in this case).
- **Dataset2:** Similar to Dataset1 but here there is no popular UON and all UONs are equally likely. Users with UONs are equally assigned UONs of types(i.e. L_i's) between 1 and 10 (inclusive) (i.e. the number of users who are assigned type i UON are 0.1% of the total number of users where $i = 1, 2, \ldots, 10$). However, the sizes of these UONs are different: the sizes of UON type i are distributed uniformly between 1 and $10k - 1k * (i - 1)$. Hence, the average size of UON type 1 is the highest while average size of UON type 10 is the smallest. In this dataset, we do not have a popular UON but users are equally presented in all the available UONs. However, the impact of users within different UONs is different (users who have UON type 1 have more followers on average compared to users who have other UON type).
- **Facebook–Twitter dataset:** This dataset is constructed using SNAP datasets of Facebook (4092 nodes, 88 234 edges) and Twitter (81 306 nodes, 1 768 149 edges) [37]. The network is constructed as follows: Facebook data is used to represent the target network. A certain percentage ρ of the Facebook users, who are randomly selected, are assigned a UON. The size of a UON is the number of followers for a randomly selected Twitter user.

To compare the different algorithms fairly, the marketing reach is computed based on the value $\Phi(S)$ for the seed set S obtained by the algorithm even if the algorithm does not explicitly take UONs into account (e.g. GNF and GENF). It may be argued here that if a user is going to share with his UON, it would not matter what criteria is used to select this user; he is going to share any way. Sharing is a user behavior while selecting which users to be in the seed set is dependent of the selection algorithm. Note that the results shown here are the average of 10 000 iterations.

6.2. Numerical results

Marketing reach: Fig. 2 shows the marketing for different algorithms when the percentage of users who have UONs ρ is 1%, 5% and 10%, respectively, when $\beta = 5k$ (i.e. UONs sizes are distributed uniformly between 1 and 5k users). The figure shows that when the percentage of users who have UONs ρ is small, the EGDEE is performing much better than the basic algorithms GNF and GENF since it incorporates the weights of these UONs in its search for the optimal seed set. However, as this percentage increases, the advantage of EDGEE decreases. This can be explained as follows. With large number of users having, on average, the same size of UONs, the chances that GNF and GENF gain from these UONs value increases as they pick users that maximize the number of followers. These large sets of followers are likely to contain members who have large UONs, which contributes to the marketing reach (even though that GNF and GENF do not explicitly take these into consideration). Fig. 3 shows the marketing reach when $\beta = 10k$. It is clear that the general pattern is the same as the case when $\beta = 5k$. Moreover, the figure shows also that the performance of GNF and GENF is even worse here as the size of UONs is larger and taking them into account becomes more vital to achieving higher marketing reach especially when the number of users who have UONs is small. Missing this small percentage of users and simply picking users with large followers who may not have UONs result in high penalty in terms of marketing reach. For example, when $\rho = 1\%$, GNF and GENF marketing reach achieves values that vary between 21% and 35% of the upper bound of marketing reach (UPMR) when $\beta = 5k$. However, these percentages drop to values that vary between 8% and 16% of UPMR when $\beta = 10k$. Note here that the marketing reaches for GDEE and EGDEE are identical as both use the same criteria for selecting the users; the difference is that EGDEE runs faster than GDEE. For this reason, we do not show the marketing reach for GDEE separately to improve the figure clarity. Generally speaking, EGDEE is performing quite well when compared with the upper bound of marketing reach. It achieves a value that is over 88% of the upper bound value in most cases. More specifically, it achieves this performance in 19 cases out of the 30 possible combinations of $|S|$, ρ, and β shown in Figs. 2 and 3. For the other cases, its value is between 75% to 79% of UPMR in 4 cases and between 80% to 87% of UPMR in 7 cases. For the Facebook–Twitter dataset (shown in Fig. 4), we observe the same general pattern about the relative performance of EGDEE compared to GNF/GENF. However, the gap between the UPMR and EGDEE is bigger in this case. This may be due to the higher density of links that exist in the Facebook set compared to the synthesis networks used in Figs. 2 and 3. Based on the analysis of the datasets, we found that, in Facebook, there are 88 234 links between the 4039 users while there are 39 992 links between the 10k users in the synthesis sets. This means that there will be more users with large number of friends in the Facebook set and as a result, these users will be picked by UPMR as the optimal set. However, the selected set of users may have large overlap between them which is not taken into consideration by UPMR and hence gives a larger value for the marketing reach. On the other hand, the EGDEE takes into consideration this factor and hence, it produces a lower value for the marketing reach.

Running time: The running time of various algorithms is shown in Table 1 and the corresponding marketing reach is shown in Table 2. For GDEE/EGDEE, there is a one time computation cost to calculate the most influential parent (T_{parent}) and the profit matrix (T_{profit}). These two times are shown separately in Table 1. Once the profit matrix is computed (which is based on influential parent information), the actual search time for seed set is quite small for GDEE/EGDEE (shown under the heading GDEE and EGDEE). It is clear from the table, that the GDEE time increases with seed size while the EGDEE time changes slightly due to the speed-up improvements described previously. Although the difference in this particular case is not that big, the saving will

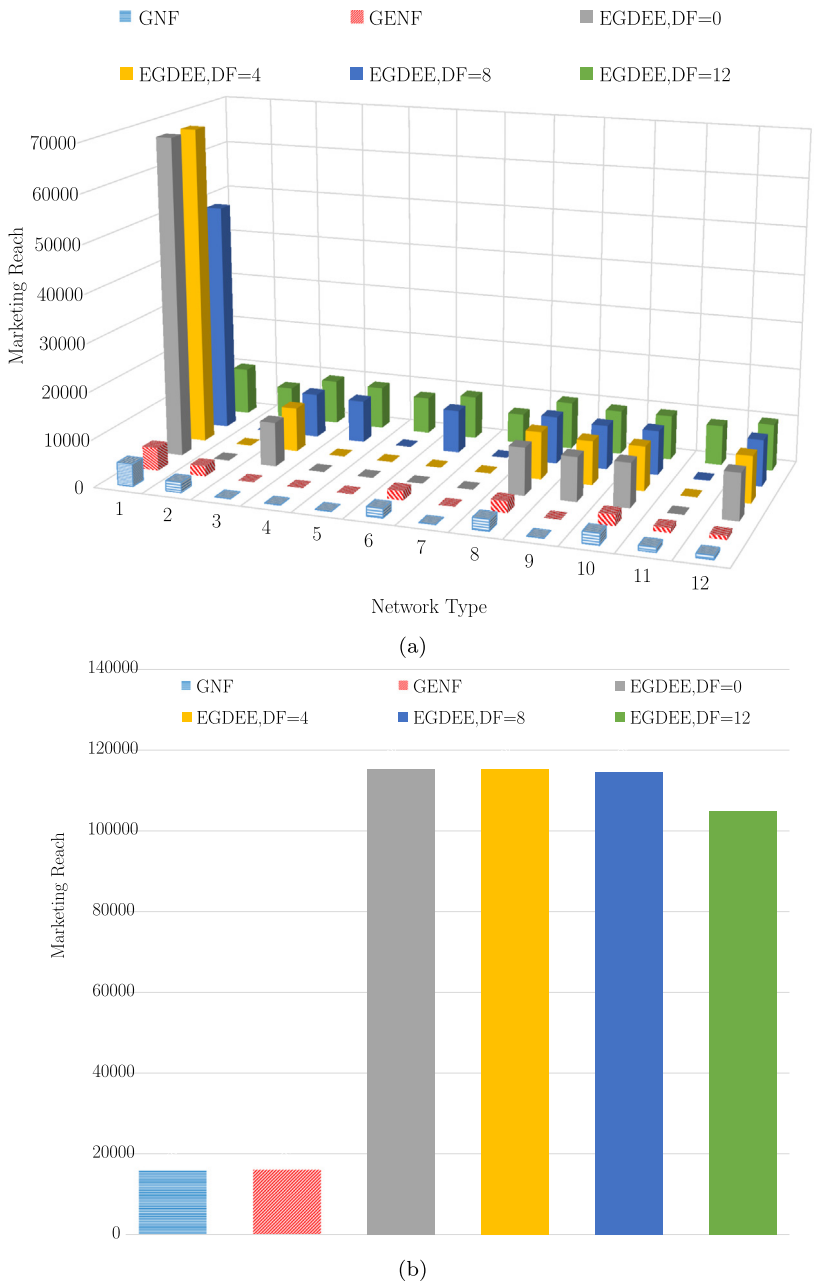

Fig. 5. (a) Marketing reach in various other social networks for GNF, GENF and EGDEE with different diversity factor. (b) Total marketing reach. [$\rho = 1\%$, $\beta = 10k$; dataset1].

be much bigger if we need to compute a bigger set size. Another point worth noting is that when we compare the total time to run EGDEE to the time of GNF and GENF, we observe that the increase in EGDEE time is not drastic but its pay off in terms of marketing reach is quite high. For example, for a seed set size of 12, the total running time of EGDEE is $182+2894+11 = 3087$ ms which is 22% higher than the time of GNF (2510 ms). On the other hand, the marketing reach of EGDEE is almost 10 times that of GNF (Table 2). The tables also include the results for running the EDGEE using the original marketing reach function (i.e. running Algorithm 4 with $\Phi(S)$ instead of $\tilde{\Phi}(S)$) which we refer to as "original"). It is clear from the tables that the original formulation will result in marginal improvement in terms of marketing reach compared to EGDEE but high price in terms of the running time. For example, for a seed size of 12, the marketing reach for the original formulation is 116 487 which is 0.7% better than the value of EGDEE, but EGDEE is 54 times faster than the original formulation in this case. Thus, EGDEE provides excellent speed improvement without compromising the performance.

Diversity impact: to investigate the impact of the diversity, we use dataset1 and dataset2 since they have users with different UONs (otherwise diversity constraint will not be of use). The marketing reach in different UONs for different diversity factor values is shown in Fig. 5(a). It is clear from the figure that the number of users in UONs, who received the ad that is initiated in the target network, is quite low when using GNF and GENF compared to EGDEE since GNF/GENF simply picks users based on the number of followers in the TOSN without considering their presence in UONs. Moreover, for EGDEE with low diversity factor (DF), most of the ads are received by UON type 1 since it is the most popular (45% of UONs are type 1) and EGDEE will simply attempt to maximize the marketing reach and hence users who have type 1 UON will be picked more frequently due to their ability to channel the ad to a very popular network. On the other

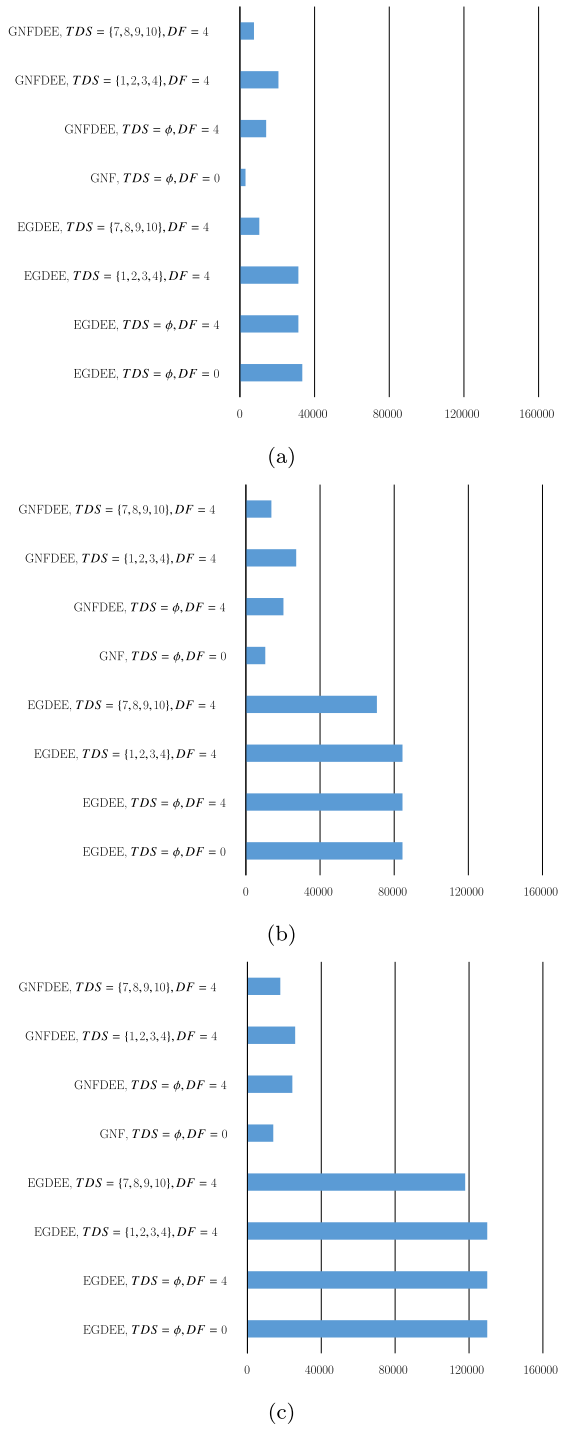

Fig. 6. Total marketing reach for a seed set size of: (a) 4, (b) 12, and (c) 20. [$\rho = 1\%, \beta = 10k$; dataset2].

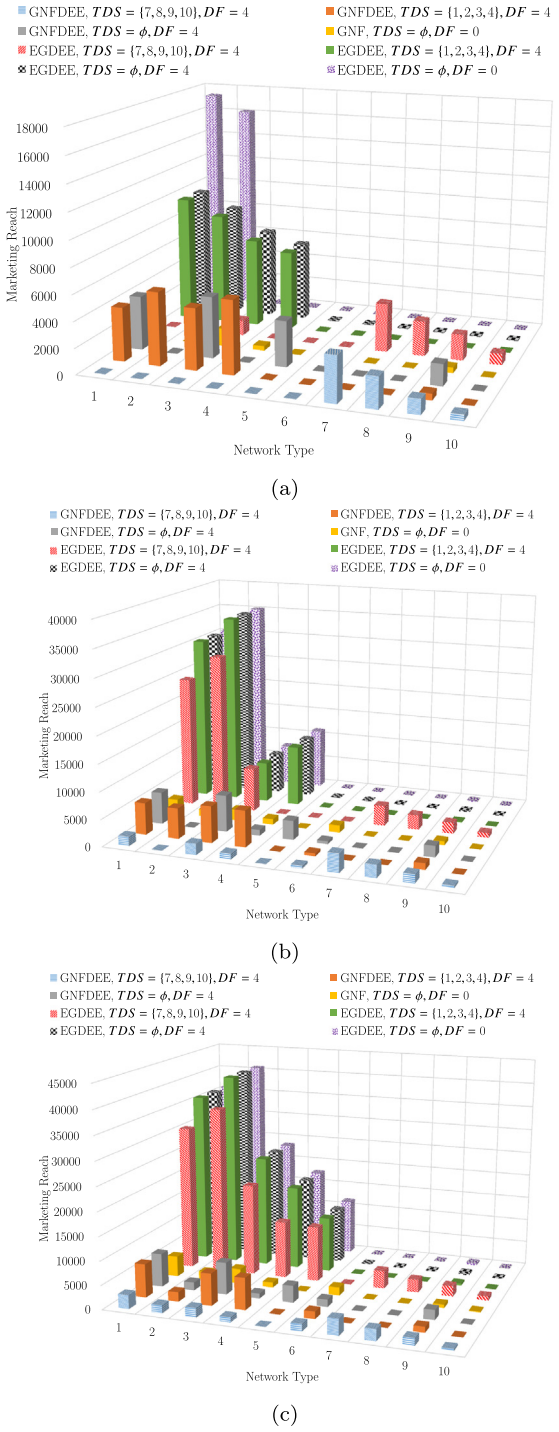

Fig. 7. Marketing reach in the UONs under different diversity constraint conditions for a seed set size of: (a) 4, (b) 12, and (c) 20. [$\rho = 1\%, \beta = 10k$; dataset2].

hand, when we go for high diversity (e.g. DF = 12), the ads reach all UONs in almost equal proportion. The total marketing reach is shown in Fig. 5(b). From this figure, we see that there is a drop in the total marketing reach of EGDEE when we are forced to diversify over more UONs since it will have to pick users who have presence in UONs that are needed to satisfy the DF requirement but may not be necessarily the most optimal in terms of maximizing the total marketing reach.

Now, let us turn our attention to analyzing the impact of the TDS diversity constraint on the marketing reach. Recall that TDS specifies a set of UON types that must be covered by the seed set (i.e. some users of the seed set must have presence on the UON types specified by TDS to ensure that the ad is propagated to these desired TDS). Obviously, if a particular type is more popular than the other types and if TDS specifies the popular type, then it is unlikely that the marketing reach will be impacted heavily as the seed set is likely to cover that popular type anyway due to its wide spread among users. To investigate this further, we use dataset2 described previously to mitigate the popularity issue and

focus more on the impact exercised by users in the different UONs which is manifested in the dataset by the different distribution of followers for the different UONs. The total marketing reach for different seed sizes is shown in Fig. 6. The figure shows that, when we restrict TDS to types that have high weight (e.g. {1, 2, 3, 4} in this particular simulation), there is little impact on the total marketing reach of EGDEE compared to the case where there is no diversity restriction (i.e. $TDS = \phi, DF = 0$). On the other hand, the marketing reach becomes lower when we restrict TDS to UON types that have lower weight (e.g. {7, 8, 9, 10}). Note also when there is no restriction on terms of *TDS* but the value of *DF* is set to 4, the value of the marketing reach is comparable to the case when *TDS* is restricted to {1, 2, 3, 4} since it is likely that EGDEE will attempt to satisfy the diversity factor requirement by picking users with higher weight UONs to maximize the marketing reach and hence the comparable results. The figure also shows that GNF has the worst performance of all since it is not sensitive to the presence of users in other social network. We also tested a variation of GNF that considers the diversity constraint which we refer to as *GNF with diversity enforced at the end (GNFDEE)*. Basically, GNFDEE works in a similar way to EGDEE in the sense that when the number of users remaining to be chosen is just enough to satisfy the diversity constraint, GNFDEE will add the user that results in the maximum increase in the number of followers of the seed set *and* satisfy the diversity constraint. Interestingly, GNFDEE has better performance than GNF (although way below EGDEE). This once more underscores our previous observations that blindly basing the selection of users on the number of followers in TOSN leads to missing the real weight of users since we do not consider their impact on other social networks. By forcing GNF to diversify, it ended up picking users that have higher weight due to the presence in other social networks and hence improved the overall marketing reach. Observe that the relative performance of the different algorithms remains the same when the seed size increases. However, for EGDEE, the gap between the case where *TDS* is restricted to {7, 8, 9, 10} (lower weight UONs) and the case where *TDS* is restricted to {1, 2, 3, 4} (higher weight UONs) decreases. The value of this gap is 23.6%, 16.4%, and 9.2% of marketing reach value of the unrestricted EGDEE (i.e. $TDS = \phi, DF = 0$) when the seed size is 4, 12, and 20, respectively. With larger seed size, EGDEE is able to compensate the decrease that is caused by the constraint $TDS = \{7, 8, 9, 10\}$ through the selection of larger number of users who have higher weight.

Fig. 7 shows the marketing reach in the UONs for the previous scenario. Note that for GNF and GNFDEE, there is no clear concentration of marketing reach on the high weight networks (i.e. low numbered network types) as they pick users with high number of followers even if these users do not have high weight in terms of their presence in other social networks. On the other hand, this concentration is evident for EGDEE for all cases except when it is restricted by the constraint $TDS = \{7, 8, 9, 10\}$ to select users so that the ads reaches these chosen network types.

7. Conclusion

In this paper, we presented a model to better capture the interactions that occur in modern online social networks. We also introduced a new optimization problem of maximizing the combined influence of a set of users that incorporates a diversity constraint of the users' other social networks. In addition, we proposed a greedy algorithm to tackle this diversity-constrained influence maximization problem. Numerical results show that our proposed algorithm performs much better than the algorithms that simply base their users selection on the targeted network metrics as they fail to take into account user influence that he exercised in other social networks that he is a member of.

For future directions, it will be interesting to study the problem of estimating the parameters of the presented model, which may not be an easy task given that most social networks companies will not share the users actions and interactions recorded over their platforms due to privacy laws. Hence, the problem need to be tackled by observing the public interactions between users to estimate different parameters such as the probabilities of reading posts and sharing them. Another interesting direction for investigation is to explore how to dynamically update the seed set of the viral marketing campaign based on the sentiment analysis of the users' opinions about the product being advertised. For example, once a negative sentiment starts to emerge in a given social network, the seed set in that network may need to be changed or increased to combat the negative views before they become widespread in that particular social network.

CRediT authorship contribution statement

Dawood Al Abri: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Shahrokh Valaee:** Conceptualization, Methodology, Formal analysis, Writing - review & editing.

References

[1] P. Hui, J. Crowcroft, E. Yoneki, BUBBLE rap: Social-based forwarding in delay-tolerant networks, IEEE Trans. Mobile Comput. 10 (11) (2011) 1576–1589, http://dx.doi.org/10.1109/TMC.2010.246.

[2] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, U.C. Kozat, Minimizing peak load from information Cascades: Social networks meet cellular networks, IEEE Trans. Mobile Comput. 15 (4) (2016) 895–908, http://dx.doi.org/10.1109/TMC.2015.2436381.

[3] T. Dinh, H. Zhang, D. Nguyen, M. Thai, Cost-effective viral marketing for time-critical Campaigns in large-scale social networks, IEEE/ACM Trans. Netw. 22 (6) (2014) 2001–2011, http://dx.doi.org/10.1109/TNET.2013.2290714.

[4] T. Schelling, Dynamic models of segregation, J. Math. Sociol. 1 (1971).

[5] M. Granovetter, Threshold models of collective behavior, Am. J. Sociol. 83 (6) (1978) 1420–1443.

[6] J.A. Morente-Molinera, G. Kou, K. Samuylov, R. Ureña, E. Herrera-Viedma, Carrying out consensual Group Decision Making processes under social networks using sentiment analysis over comparative expressions, Knowl.-Based Syst. 165 (2019) 335–345.

[7] E. Muller, R. Peres, The effect of social networks structure on innovation performance: A review and directions for research, Int. J. Res. Mark. 36 (1) (2019) 3–19.

[8] J.A. Morente-Molinera, G. Kou, C. Pang, F.J. Cabrerizo, E. Herrera-Viedma, An automatic procedure to create fuzzy ontologies from users' opinions using sentiment analysis procedures and multi-granular fuzzy linguistic modelling methods, Inform. Sci. 476 (2019) 222–238.

[9] J. Ren, D. Zhu, H. Wang, Spreading-vanishing dichotomy in information diffusion in online social networks with intervention, Discrete Contin. Dyn. Syst. Ser. B 24 (4) (2019).

[10] A. Al Kindi, D. Al Abri, A. Al Maashri, F. Bait-Shiginah, Analysis of malware propagation behavior in Social Internet of Things, Int. J. Commun. Syst. 32 (15) (2019) e4102, URL https://doi.org/10.1002/dac.4102.

[11] J.I.R. Molano, J.M.C. Lovelle, C.E. Montenegro, J.J.R. Granados, R.G. Crespo, Metamodel for integration of internet of things, social networks, the cloud and industry 4.0, J. Ambient Intell. Human. Comput. 9 (3) (2018) 709–723.

[12] J. Lies, Marketing intelligence and big data: Digital marketing techniques on their way to becoming social engineering techniques in marketing, Int. J. Interact. Multimedia Artif. Intell. 5 (5) (2019).

[13] C. Moreno, R.A.C. González, E.H. Viedma, Data and artificial intelligence strategy: A conceptual enterprise big data cloud architecture to enable market-oriented organisations, IJIMAI 5 (6) (2019) 7–14.

[14] Y. Feng, H. Li, Z. Chen, B. Qiang, Improving recommendation accuracy and diversity via multiple social factors and social circles, in: Innovative Solutions and Applications of Web Services Technology, IGI Global, 2019, pp. 132–154.

[15] P. Domingos, M. Richardson, Mining the network value of customers, in: Proc. 7th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, 2001, pp. 57–66, URL http://dl.acm.org/citation.cfm?id=502525.

[16] M. Richardson, P. Domingos, Mining knowledge-sharing sites for viral marketing, in: Proc. 8th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, 2002, pp. 61–70, URL http://dl.acm.org/citation.cfm?id=775057.

[17] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, in: Proc. 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, 2003, pp. 137–146, URL http://dl.acm.org/citation.cfm?id=956769.

[18] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, Cost-effective outbreak detection in networks, in: Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, ACM, 2007, pp. 420–429, URL http://dl.acm.org/citation.cfm?id=1281239.

[19] W. Chen, Y. Wang, S. Yang, Efficient influence maximization in social networks, in: Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, in: KDD '09, ACM, New York, NY, USA, 2009, pp. 199–208, http://dx.doi.org/10.1145/1557019.1557047, URL http://doi.acm.org/10.1145/1557019.1557047.

[20] C. Borgs, M. Brautbar, J. Chayes, B. Lucier, Maximizing social influence in nearly optimal time, in: Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SIAM, 2014, pp. 946–957.

[21] R. Narayanam, Y. Narahari, A shapley value-based approach to discover influential nodes in social networks, IEEE Trans. Autom. Sci. Eng. 8 (1) (2011) 130–147, http://dx.doi.org/10.1109/TASE.2010.2052042, URL http://dx.doi.org/10.1109/TASE.2010.2052042.

[22] J.-R. Lee, C.-W. Chung, A query approach for influence maximization on specific users in social networks, IEEE Trans. Knowl. Data Eng. 27 (2) (2015) 340–353, http://dx.doi.org/10.1109/TKDE.2014.2330833.

[23] S. Bhattacharya, K. Gaurav, S. Ghosh, Viral marketing on social networks: An epidemiological perspective, Physica A 525 (2019) 478–490.

[24] Y. Li, B. Zhao, J. Lui, On modeling product advertisement in large-scale online social networks, IEEE/ACM Trans. Netw. 20 (5) (2012) 1412–1425, http://dx.doi.org/10.1109/TNET.2011.2178078.

[25] B. Liu, G. Cong, Y. Zeng, D. Xu, Y.M. Chee, Influence spreading path and its application to the time constrained social influence maximization problem and beyond, IEEE Trans. Knowl. Data Eng. 26 (8) (2014) 1904–1917, http://dx.doi.org/10.1109/TKDE.2013.106.

[26] O. Yagan, D. Qian, J. Zhang, D. Cochran, Information diffusion in overlaying social-physical networks, in: Info. Sciences and Systems (CISS), 2012 46th Annual Conf., 2012, pp. 1–6, http://dx.doi.org/10.1109/CISS.2012.6310749.

[27] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, J. Han, Event-based social networks: Linking the online and offline social worlds, in: Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, KDD '12, ACM, New York, NY, USA, 2012, pp. 1032–1040, http://dx.doi.org/10.1145/2339530.2339693, URL http://doi.acm.org/10.1145/2339530.2339693.

[28] Y. Shen, T.N. Dinh, H. Zhang, M.T. Thai, Interest-matching information propagation in multiple online social networks, in: Proc. 21st ACM Int. Conf. on Information and Knowledge Management, CIKM '12, ACM, New York, NY, USA, 2012, pp. 1824–1828, http://dx.doi.org/10.1145/2396761.2398525, URL http://doi.acm.org/10.1145/2396761.2398525.

[29] D.T. Nguyen, S. Das, M.T. Thai, Influence maximization in multiple online social networks, in: 2013 IEEE Global Communications Conference, GLOBECOM, 2013, pp. 3060–3065, http://dx.doi.org/10.1109/GLOCOM.2013.6831541.

[30] H. Zhang, D.T. Nguyen, H. Zhang, M.T. Thai, Least cost influence maximization across multiple social networks, IEEE/ACM Trans. Netw. 24 (2) (2016) 929–939, http://dx.doi.org/10.1109/TNET.2015.2394793.

[31] A. Bosworth, Bringing people better ads, 2016, URL http://newsroom.fb.com/news/2016/05/bringing-people-better-ads/.

[32] S. Englehardt, A. Narayanan, Online tracking: A 1-million-site measurement and analysis, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016, pp. 1388–1401.

[33] T. Iofciu, P. Fankhauser, F. Abel, K. Bischoff, Identifying users across social tagging systems, in: Proc. 5th Int. Conf. on Weblogs and Social Media, Nejdl Publication, 2011, pp. 522–525.

[34] F. Buccafurri, G. Lax, A. Nocera, D. Ursino, Discovering links among social networks, in: P.A. Flach, T. De Bie, N. Cristianini (Eds.), Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 467–482, http://dx.doi.org/10.1007/978-3-642-33486-3_30.

[35] R. Taylor, Approximation of the quadratic Knapsack problem, Oper. Res. Lett. 44 (4) (2016) 495–497.

[36] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, science 286 (5439) (1999) 509–512.

[37] J. Leskovec, A. Krevl, SNAP Datasets: Stanford large network dataset collection, 2014, http://snap.stanford.edu/data.