The Limits of Influence Maximisation in Online Social Networks

Stefania Costantini*, Giovanna Costanzo†, Pasquale De Meo†, Rino Falcone‡, Fabio Persia* and Alessandro Provetti§
*DISIM - University of L'Aquila. L'Aquila, Italy
†DICAM - University of Messina. Messina, Italy
‡ISTC - National Research Council. Rome, Italy
§CMS - Birkbeck University of London. London, UK
SPS - University of Milan. Milan, Italy

***Abstract*—Influence maximisation over online social networks is a challenging ethical and technical problem. Among many applications, it is becoming relevant in public health and prevention: by choosing the right spokesperson for an advertising campaign, in fact, we can encourage people to adopt healthier lifestyles and increase organ donation rates. Traditional Influence Maximisation models assume that all members of a social network have the same propensity to spread a message, regardless of the topic, and that each user will forward a message to all of his or her contacts. These assumptions are clearly unrealistic: a user could be very sensitive to some messages (thus contributing to the information diffusion process) and completely insensitive to others (thus stopping the spread of information in a community). A user could also selectively decide to whom a message should be sent. To overcome the above limitations, we introduce a novel information diffusion model called IMBC (*Influence Maximization with Budget Constraints*), where the *budget* is an upper bound on the number of contacts to whom a message can be spread. We have experimentally verified our algorithm on six large real-world networks containing millions of nodes.**

***Index Terms*—Influence maximisation, susceptibility to persuasion in social networks, approximation algorithms, social networks in health care.**

I. Introduction

We examine the problem of understanding and modelling the spread of influence in a social network, a topic that has received much attention in many disciplinary fields. Social network analysis allows us to effectively model the mechanism by which a message spreads within a community, and provides the tools to understand what actions should be taken to maximise the success of a given information campaign [9], [13], [16].

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 2.2.2022 by the Italian Ministry of University and Research (MUR), funded by the European Union – NextGenerationEU– Project Title *TrustPACTX - Design of the Hybrid Society Humans-Autonomous Systems: Architecture, Trustworthiness, Trust, EthiCs, and EXplainability (the case of Patient Care)* – CUP J53D23007040001- Grant Assignment Decree No. 959 adopted on 30/0/2023 by the Italian Ministry of Ministry of University and Research (MUR).

Many organisations in the field of health (e.g., government agencies, NGOs, no-profit etc.) have made considerable efforts in recent years to promote prevention campaigns through mass media and Online Social Networks (OSNs) [5], [14], [18]. OSNs such as Facebook, X, YouTube and Instagram have become increasingly important, and celebrities from show business have been involved in promoting causes as diverse as organ donation, adopting a healthier lifestyle with a more balanced diet and daily physical activity, and regular clinical tests.

The success of these campaigns has a huge impact on the life of the community: for example, effective adoption of a healthy lifestyle prevents the onset of cardiovascular disease in old age, which is widespread in most industrialised countries. Targeted promotional campaigns have promoted a culture of donation, leading to rapid increases in organ, blood and tissue donation rates in many countries [18], [23].

In recent years, especially in the sociological field, there have been numerous studies on how to exploit online social networks (and their power to spread information) to promote the culture of organ and blood donations. The topic of organ donation certainly has a strong ethical impact and many scholars in the sociology sector consider parameters such as the rate of blood donors per million inhabitants as an index of the *social progress* of a community.

From different disciplinary angles, we have studied the recent phenomenon of organ donation in Italy, where in recent years donation rates have significantly improved[1]. The growth in the number of potential donors depends on many factors and, certainly, the use of effective and understandable media campaigns has increased the number of potential donors.

The most recent sociological studies reveal a surprising fact: if the 'yes' to organs (or blood) donation comes from a famous person then the promotion campaign is more successful than

[1]According to the latest report (2023) of the Italian National Transplant Center, Italy is currently the second European country after Spain in terms of number of donors per million inhabitants.

in the case in which the testimonials are very authoritative people (for example doctors or scientists) but less popular.

Consequently, the mathematical models that we use to date to describe the propagation of ideas in a social network and, in particular, to maximise the number of potential message receptors are ineffective if the target of users to be reached depends, to some extent, on the topic of the message or on who originated the message.

We present a new, extended model where OSN participants have a *varying degree of attention* to the message, in the sense that the propensity of that individual to listen and to propagate in turn the received message depends on the topic of the message and possibly on who generated the message. In other words, we make attention a finite quantity, called *budget.* These additional constraints lead to a more complex formulation of the classical models of influence that, in the limit case, degenerates to the standard model of literature, which is well know for being NP-hard and therefor prohibitive for OSN analysis.

A. A formal model

We abstract away OSNs with the simplest of models: an undirected network (graph) $G = \langle N, E \rangle$, where the nodes in N are the actors involved in the information campaign (e.g., individual citizens, patient associations, hospitals, etc.) and the edges describe how a message flows from a source node to a target node. The problem of interest here is *Information Maximization* (in short, *IM*); it requires to determine a set $\mathcal{S} \subseteq N$ of nodes (called *seed nodes*) of size less than or equal to a predefined integer k that are responsible for disseminating a specific message (e.g. an organ donation campaign). At time $t = 0$, seed nodes generate a message and they (independently each other) disseminate their message to their neighbours; a node which receives a message can decide not to spread the message to its neighbours, or it may send the received message to its contacts, thus accelerating the information dissemination process. We want the fraction of nodes in G that have received and accepted the message to be as large as possible as $t \to +\infty$ and we point out that the constraint $|\mathcal{S}| \leq k$ models the fact that running an advertising campaign involves costs that are assumed to be proportional to the number of seed nodes.

The IM problem has been widely studied in the last decades [13]. However, some of the underlying assumptions of the IM problem are not verified in the healthcare domain: for some topics, in fact, a user may be willing to forward the message to his/her contacts, on the basis of ethical, religious or other beliefs; in contrast, for other topics, the user may refuse to forward the message.

A second criticism of the standard formulation of the IM problem is that users are always committed to receiving and forwarding messages. However, this assumption is false in real life: a typical social media user is overwhelmed by a myriad of messages of different types and often struggles to pay due attention to the messages received.

In this paper, we propose an extension of the standard IM problem in which each node is enriched with two additional pieces of information, namely the *interest* in a particular message and the *attention* a user can pay to the same message. We assume that the propensity of a node to propagate information to its neighbours is modelled by a random variable, and we have focused on the case where this random variable is either uniform or Gaussian. We also assume that a node can propagate a message to a subset of size B (called *budget*) of its contacts, where B is a random number between one and the degree of a node. Our assumptions enrich the standard mathematical model underlying the IM problem and allow us to respond to the criticisms highlighted above. The problem we consider (which we call *Influence Maximisation with Budget Constraints*, *IMBC* in short) generalises the problem of influence maximisation and we prove its NP-hardness.

We experimentally tested our system on six large real datasets. More specifically, we considered two well-known algorithms to solve the IM problem, known as *Reverse Influence Sampling* (RIS) [4] and Degree Discount (DD) [6]. We analysed how the performance of these algorithms (defined as the percentage of infected nodes or *spread*) depends on B and k. Our experimental analysis show that the propensity to propagate a message has a fundamental impact on the ability to spread the message: in fact, a low participation rate in the information dissemination processes is equivalent to the *sparsification* of the initial network and this strongly limits the ability to spread a message, independently of the algorithm we decide to apply to find out the seed set. The budget also plays a key role in network sparsification, and, in particular, our experiments show that the number of infected nodes grows approximately linearly as the budget increases.

II. Related Works

Domingos and Richardson [9] were the first to consider the problem of maximising the spread of a product within a community and coined the term Influence Maximization (IM). In IM, we consider a social network and assume that an external agent wants to sell a product to as many members of the social network as possible. The seller incurs costs associated with advertising the product, which lead to a specific constraint, that is the seller can recruit at most k members of the network (called *seed nodes*), who initially receive the product to sell and must convince their acquaintances to purchase it. The IM problem is thus to identify the set of seed nodes of size k that maximises the number of individuals (called *spread*) who decide to purchase the product. Kempe *et al.* [13] showed that the IM problem is NP-HARD. However, they showed also that if the spread is a submodular function, then there exists a greedy algorithm that produces an approximate solution that deviates from the optimal one by at most $1 - \frac{1}{e}$.

Kempe *et al.* [13] used Monte Carlo (MC) simulations to estimate the relative change in influence due to the inclusion of a node in the seed set, but the proposed approach is computationally expensive and therefore only applicable to modestly sized networks. Chen, Yuan and Zhang [7] have shown that computing influence in models is P-hard in networks of arbitrary size.

Recent literature has therefore focused on efficient influence estimation algorithms.

For example, the CELF [16] and CELF++ [11] algorithms have been developed to avoid unnecessary MC simulations with a significant improvement in computational efficiency. unnecessary using upper bound information.

Borgs *et al.* [4] proposed a sampling-based approach to determine the optimal set of seed nodes. Degree-based heuristics have been widely applied to the IM problem and the Degree-Discount algorithm [6] (DD in short) is, perhaps, the most effective degree-based approach to maximise influence. In the DD approach, a modified version of the node degree is used so that the top-k ranked nodes are selected as seeds. Some authors have studied the problem of influence maximization (*IM*) in the presence of budget constraints [3], [8], [20].

To the best of our knowledge, [20] is the first paper to introduce the Budgeted Influence Maximization (*BIM*) problem, which is related to the objectives of our study. They assume that there is a cost to activate a seed node and that the overall objective is to maximise the expected spread, under the assumption that the sum of the costs to pay to activate all the seed nodes is less than a fixed budget. The BIM problem is NP-Hard and a greedy algorithm with a constant approximation ratio to so,lve it is proposed in [20].

A further approach to solving the BIM problem is presented in [3], [12], in which the authors suggest to use reverse sampling techniques to quickly obtain an accurate bound of the expected spread associated with the choice of a particular set of nodes as the seed set.

Both our approach and the ones reported above aim to study how the spread of influence can be maximised in case of cost constraints; however, the notion of cost in our approach completely differs from the notion of cost in the aforementioned approaches. Specifically, in [3], [8], [20], the cost is the amount of resources required to activate a node.

Our model assumes that the information propagation process leans only on some edges. However, edges that will actually be employed to convey information is *a-priori* unknown.

III. Standard Influence Maximisation

We begin our discussion with notation and problem statement that are largely taken from [13]. So, let again $G = \langle N, E\rangle$ be a network (graph) where N is a set of *nodes* and $E \subseteq N \times N$ is a set of *edges,* with $|N| = n$ and $|E| = m$. Each edge $\langle i, j\rangle \in E$ is associated with an *infection probability* p_{ij}. We define the *neighbourhood* $N(i)$ of a node i as the set of nodes adjacent to i, namely $N(i) = \{j \in N : \langle i, j\rangle \in E\}$. The degree d_i of a node i is the number of its neighbours, that is: $d_i = |N(i)|$. A *path* $p = \{i_0, i_1, \ldots, i_r\}$ iss a sequence of nodes in N such that pairs of consecutive nodes in p are connected by an edge.

Two of the most popular models of information diffusion in social networks are *linear threshold* (LT, for short) and *information cascades* (IC, for short). In this paper we focus on the IC model and plan to extend our study to the LT model as future work.

In the IC model, we consider a subset $S \subseteq N$ of nodes of G (called *seed set*). Nodes in S start producing and sending information through their neighbours, and, more in detail, node i transmits a message to $j \in N(i)$ with probability p_{ij}; thus, a message arrives at any other node $\ell \in N - S$ if and only if there exists a path from i to ℓ. The message propagation process creates a graph (called *sample graph*) $G_S = \langle N_S, E_S\rangle$ which contains all nodes and edges of G which participated to the information diffusion process.

Let $R(S) = |N_S|$ be the random variable that quantifies the number of nodes The *spread* $I(S)$ is the expected value of $R(S)$, that is $I(S) = \mathbb{E}[R(S)]$ and, the higher $I(S)$, the better S in diffusing information.

The *Influence Maximization Problem*, IM, of [13] is defined as follows.

> Given a network $G = \langle N, E\rangle$, a function $\mathbf{P} : E \to [0, 1]$ which associates each edge in E with its influence probability and a positive integer k, find the seed set S of size $k \leq n$ which maximises $I(S)$.

[13] proved that maximising $I(S)$ is NP-hard and moved to design a greedy algorithm that produces a solution that is $1-\frac{1}{e}$ factor close to the optimum [19]. Their Greedy algorithm is based on Monte Carlo simulations; unfortunately, the overall running time is still prohibitive even for networks of few thousand nodes. So, today heuristics solutions are in use, such as *reverse influence sampling* [4] and the *degree-based* method [6].

IV. Influence Maximisation with Budget Constraints

The IC model is an easy but powerful tool to describe diffusive processes in graphs but it has two main limitations that deserve of being considered in detail:

1) *Susceptibility to persuasion*. If an individual receives a message, then she/he *has to* propagate the message to her/his neighbors. Such an assumption is excessively conservative because, depending on the content of a message, a user can accept or refuse to forward the message.
2) *Limited Attention*. The IC model assumes an individual can propagate a message to *all* her/his neighbours, but such an assumption might be false because individuals could *selectively* decide to forward a message only to a fraction of their contacts.

Some previous studies focused on susceptibility to persuasion [1], [25] and we observe that individuals highly susceptible to persuasion favour the process of information dissemination while individuals displaying low susceptibility hinder the information flow in a social network. To model the susceptibility to persuasion we associate each node $i \in N$ with a random variable $f_i : N \to [0, 1]$: the higher f_i, the higher the probability the node i will decide to continue propagating a message. We considered two configurations to model the function f, namely *(i) Uniform*, that is f is a uniform random variable in the range $[0, 1]$, or *(ii) Normal*,

that is f is a truncated Gaussian variable with mean 0.5 and standard deviation 1 and it is constrained to lie in the [0,1] interval.

As for limited attention, we point out that previous studies highlight the existence of an upper bound, known as *Dunbar's number*, on the number of stable social relationships of an individual [10]. The Dunbar's number depends on both biological constraints [10] and limitations imposed by the real world in which we live. In fact, because of the time of each individual is finite, each person will decide with whom she/he can interact on the basis of personal interests and needs, for instance. Thus, we modified the IC model to consider the limited attention of an individual and, to this end, we assume that a node i can only propagate a message to a finite number B, called *individual budget*, of its neighbors. We assume that B_i is drawn, uniformly at random, from the interval $[1, d_i]$: in this way, the node i constructs, uniformly at random, a subset $A_B(i) \subseteq N(i)$ of its neighbors with $|A_B(i)| = B_i \leq d_i$ to which it can forward its message. We leave as future study more sophisticated procedures to construct $A_B(i)$ (e.g., procedures which take individual features of the neighbors of i into account).

We call our model *Influence Maximization with Individual Budget Constraint* (in short, *IMBC*) and, in the next section, we discuss its main features.

V. Computing the IMBC of networks

A. Mathematical Formulation

The input of the IMBC problem are *(i)* a network $G = \langle N, E \rangle$, *(ii)* a random vector $\mathbf{f} : N \to [0, 1]$ that maps each node $i \in N$ with its susceptibility to persuasions and *(iii)* a random function $\mathbf{q}_B : E \to [0, 1]$ that captures limited user attention and, more concretely, it associates each edge $\langle i, j \rangle \in E$ with its probability of being part of the set $A_B(i)$.

Observe that if $\mathbf{f}$ is equal to a constant value $\overline{f}$ for any node in N and $\mathbf{q}_B$ is equal to a constant value $\overline{q}$ for each edge in E, then the IMBC problem reduces to the standard IM problem defined in Section III.

In the most general case, we note that neither $\mathbf{f}$ nor $\mathbf{q}_B$ is constant; moreover, we can include the contribution of $\mathbf{q}_B$ in the infection probabilities and thus consider a new infection probability vector $\hat{\mathbf{p}} = \mathbf{p} \odot \mathbf{q}_B$, where $\odot$ is the Hadamard product of the two vectors.

Let $I(S, \mathbf{f}, \hat{\mathbf{p}})$ be the expected spread in the IMBC problem as function of $\mathbf{f}$ and $\hat{\mathbf{p}}$ provided that we have chosen $S \subseteq N$ as the seed node. With the constraint $|S| = k$ we have that $I(S, \mathbf{f}, \hat{\mathbf{p}})$ is monotone and submodular and the problem of maximizing it is NP-Hard:

Theorem 1. *The expected spread $I(S, \mathbf{f}, \hat{\mathbf{p}})$ associated with the IMBC problem is monotone and submodular and the problem of optimizing it is NP-Hard.*

Proof. We rewrite $I(S, \mathbf{f}, \hat{\mathbf{p}})$ in a more manageable fashion. Specifically, if we fix $\mathbf{f}$ and $\hat{\mathbf{p}}$, we obtain a graph G_l; in this case, let $I(S|G_l)$ be the optimal expected spread in G_l.

For any possible choice of $\mathbf{f}$ and $\hat{\mathbf{p}}$ we generate an (infinite) collection $\mathcal{G}$ of graphs and, by the law of iterated expectations [2], we can write:

$$I(S, \mathbf{f}, \hat{\mathbf{p}}) = \sum_{G_l \in \mathcal{G}} I(S|G_l) \times \Pr(G_l) \tag{1}$$

The function $I(S|G_l)$ monotone and submodular because it is the objective function of the IM problem [13]; the function $I(S, \mathbf{f}, \hat{\mathbf{p}})$ is also monotone and submodular because it is the convex combination of monotone and submodular functions.

The optimization of each function $I(S|G_l)$ is NP-Hard, and, thus, the optimization of $I(S, \mathbf{f}, \mathbf{q})$ is NP-Hard too. □

To compute $I(S, \mathbf{f}, \hat{\mathbf{p}})$ we generate M independent and identical distributed (i.i.d) samples from $\mathcal{G}$, say $\mathcal{G}^\star = \{G_1, G_2, \ldots, G_M\}$ and we apply the standard sample mean estimator:

$$\hat{I}_M = \frac{1}{M} \sum_{G_l \in \mathcal{G}^\star} I(S|G_l) \times \Pr(G_l)$$

Furthermore, for each $G_l \in \mathcal{G}^\star$ let S_l be the corresponding optimal seed set. We define $\mathcal{S}^\star = \{S_1, S_2, \ldots, S_M\}$.

Observe that graphs belonging to $\mathcal{G}^\star$ could display substantial differences at the structural level and, thus, we would expect significant differences among the expected spread associated with each graph in $\mathcal{G}^\star$. As a result, we expect that there will be significant differences in the composition of each of the elements in $\mathcal{S}^\star$.

Our goal is to construct a set of seed nodes $\hat{S} \subseteq N$ that is *as similar as possible* to all of the sets in $\mathcal{S}^\star$ for a given similarity metric. We will call the set $\hat{S}$ the *median* associated with $\mathcal{S}^\star$.

We provide a procedure to construct the median and, to this end, let $N(k)$ be the collections of all subsets of size k drawn from G; we also introduce a *similarity function* $\phi : N(k) \times N(k) \to [0, 1]$ which takes as input a pair $\langle S_a, S_b \rangle$ of subsets of size k drawn from N and returns a number from 0 and 1 as output. The higher the output of $\phi(S_a, S_b)$, the more similar S_a and S_b.

For each set $S_l \in \mathcal{S}^\star$, we fix a threshold $\theta_l \in [0, 1]$ and we require that $\phi(\hat{S}, S_l)$ is equal to θ_l. The term $\varepsilon_l = |\phi(\hat{S}, S_l) - \theta_l|$ quantifies to what extent the similarity between the median $\hat{S}$ and the set S_l differs from the target value θ_l. Other approaches to measuring the discrepancy between $\phi(\hat{S}, S_l)$ and θ_l are possible; however, in this paper, we use the absolute deviation because our notion of median of a collections of sets resembles the notion of median for a collection of numbers.

Our goal is to make ε_l as small as possible for each $l = 1 \ldots M$. In the light of such a reasoning, we define the following *loss function*:

$$\mathcal{L} = \sum_{S_l \in \mathcal{S}^\star} |\phi(\hat{S}, S_l) - \theta_l|$$

and we are in charge of solving the following optimization problem:

$$\begin{aligned} \min \quad & \mathcal{L} = \sum_{S_i \in \mathcal{S}^\star} |\phi(\hat{S}, S_i) - \theta_i| \\ \text{s.t.} \quad & |\hat{S}| = k. \end{aligned} \tag{2}$$

A numerically-efficient procedure for the above optimisation problem has been implemented by choosing the similarity function $\phi(\cdot, \cdot)$: to be the *Dice-Sorensen coefficient* [24], a popular metric introduced in the field of Information Retrieval.

VI. Experiments

We have designed experiments to compare the performance of the RIS and DD algorithms under the assumption that the conditions given by the IMBC problem hold. Experiments aims to answer the following question: *How does the expected spread of the RIS and the DD algorithms vary as B varies*?

Such a question has to be repeated for the Uniform and Normal configurations, respectively, for a total of four tests. We experimented with $B \geq 5$ and $k \geq 5$, which seemed a reasonable setup for all six datasets involved. Finally, all experiments included a baseline strategy, called *Rand,* where seed nodes are selected uniformly at random.

A. Dataset Description

Our experiments were run over six publicly-available social network datasets. The main features of each are summarised in Table I, let's see them in detail.

Amazon. This is the co-purchase network of Amazon based on the customers *who bought this also bought* feature [26]. Nodes are products; an edge between two nodes indicates that the corresponding products have been frequently bought together.

Deezer. This dataset depicts the social network of Deezer users, a music streaming service [22]: nodes identify Deezer users from European countries and edges are mutual follower relationships between them.

Epinions. This is a *who-trust-whom* online social network drawn from the Website site Epinions.com [21]. Nodes in the graph correspond to Epinions members; an edge from a source node to a target one indicates that the source trusts the target.

EU Mail First introduced by [15], this dataset represents the network of e-mail messages exchanged within a large (undisclosed) European institution. Each node is associated with an individual whereas edges indicate that at least one e-mail has been sent from the source to the target.

Gnutella This is a snapshot of the Gnutella peer-to-peer file-sharing network collected in August 2002 [15]. Nodes are associated with hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts.

Twitter This dataset consists of *lists* from Twitter [17]. Twitter allows its members to create up to 1 000 Lists, each with up to 5 000 users to help people businesses to follow tweets from particular people.

TABLE I: Main features of the graphs used in our experimental evaluation

Dataset	Nodes	Edges	Density
Amazon	334 863	925 872	0.000017
Deezer	28 281	92 752	0.000232
EU Mail	265 214	365 570	0.000010
Gnutella	62 586	147 892	0.000076
Epinions	75 879	405 740	0.000141
Twitter	23 370	32 831	0.000120

B. Sensitivity to budget values

This section reports the expected spread obtained by the RIS and DD algorithms as B varies from 5 to 50, whereas the seed set size has been fixed to $k = 10$.

We observe that the expected spread in the Normal configuration, in Figure 1, is generally lower than that observed in the Uniform case, in Figure 2.

We submit that this result is due to the fact that in the Normal configuration some nodes could have a small susceptibility and, therefore, their edges will not contribute to the spread of information. The graph actually used to convey messages in the Normal configuration is, thus, expected to be sparser than in the Uniform one and this justifies smaller values of the expected spread.

It is worth noticing that network *density* may help increase the expected spread in both the Uniform and Normal configurations. For instance, the density of **Epinions** is 8.29 times bigger than that of **Amazon** and, in some cases, the expected spread in **Epinions** is more than double than that observed in **Amazon**.

In terms of expected spread, interesting behaviours emerge. For the DD algorithm, the spread grows almost linearly with B and the same behaviour has been encountered with RIS.

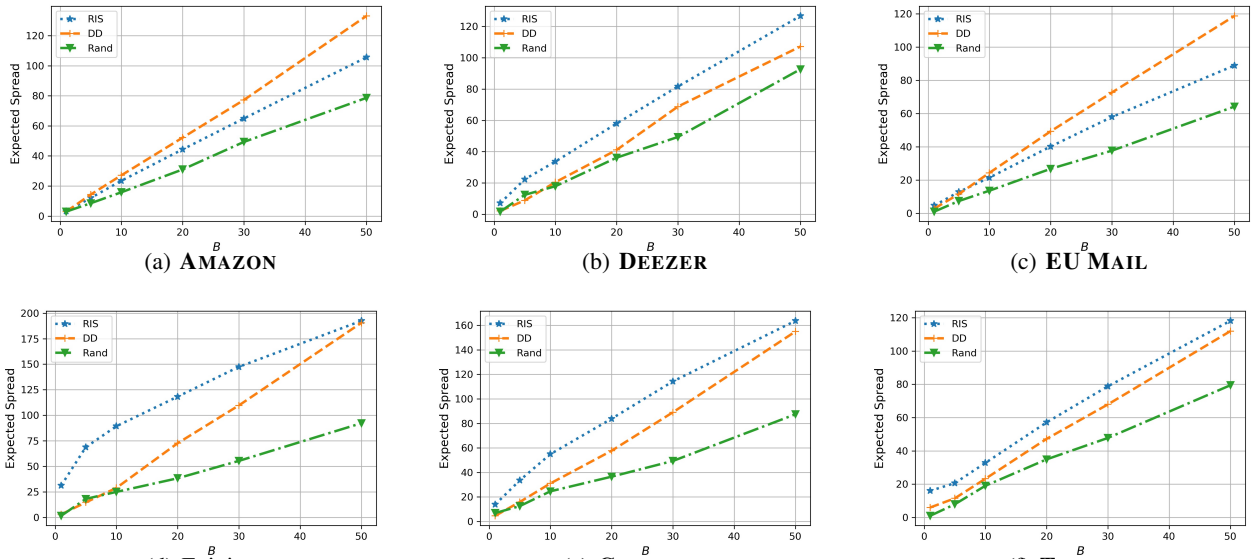
From our experiments, there is no clear "winner" in both the Uniform and Normal configurations; for example, RIS achieves a better spread than DD on **Deezer** and **Epinions** whereas the spread obtained by RIS closely matches the spread of DD on **Gnutella/Twitter**. Finally, DD achieves a larger expected spread than RIS in **EU Mail/Amazon** datasets.

VII. Conclusion and Future Works

We introduced a novel model of influence diffusion, called IMBC (Influence Maximization with Budget Constraints) where the spreading of a message from a node to its neighbours has a cost, and, in addition, the neighbours can show a varying amount of *resistance* to the message itself.

Unlike previous models about influence diffusion in social networks, we have assumed that each node can only propagate a message to a subset of size B of its neighbors. Also, unlike previous works, we have considered the resilience of a node to the receipt of a message. We finally defined the Median Problem in the IMBF Problem and we provided an efficient algorithm to construct it.

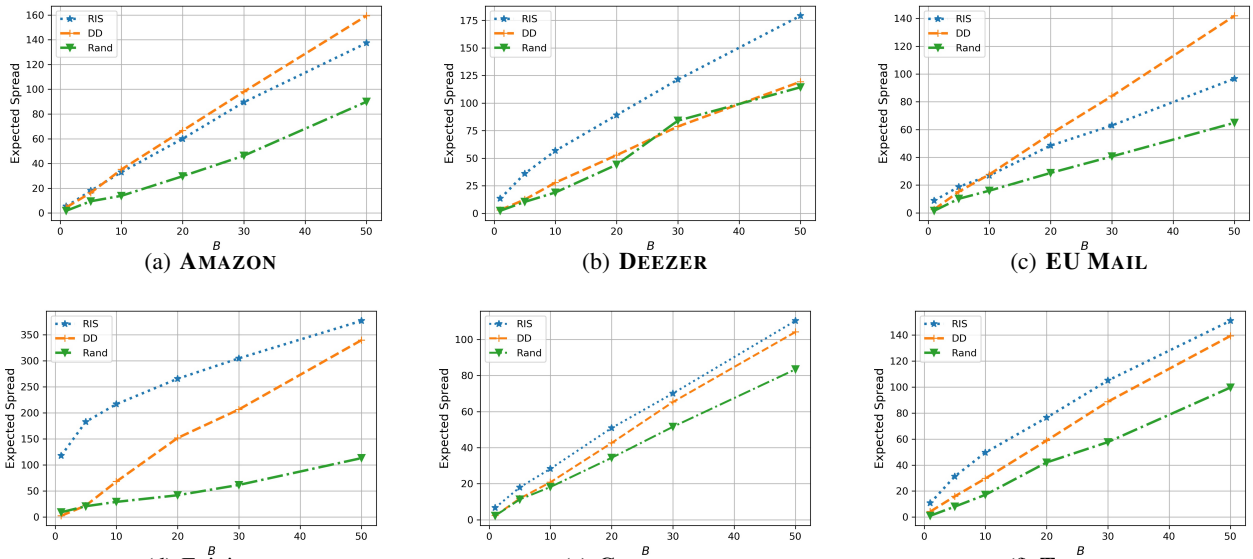
We believe that this new framing of information diffusion, and the algorithms we presented herein will open the way to



(a) AMAZON (b) DEEZER (c) EU MAIL

(d) Epinions (e) GNUTELLA (f) TWITTER

Fig. 1: Expected spread of *RIS*, *DD* and *Rand* strategies as function of the budget B in the *Normal* configuration.



(a) AMAZON (b) DEEZER (c) EU MAIL

(d) Epinions (e) GNUTELLA (f) TWITTER

Fig. 2: Expected spread of *RIS*, *DD* and *Rand* strategies as function of the budget B in the *Uniform* configuration.

further modeling the online activity of social networks. For instance, an immediate research goal would be to formalise competitive/adversarial diffusion processes on networks, namely the evolution of influence processes that were simultaneously generated by groups of people displaying divergent opinions. This new scenario could play out as follows. Groups of nodes in $\mathcal{G}$ start generating messages with the aim of persuading public opinion to support (or not) a certain social cause. For example, with regard to the COVID-19 pandemic, we might consider a group of individuals in $\mathcal{G}$ who support mass vaccination and the use of individual protective devices such as masks; at the same time, other groups of individuals in $\mathcal{G}$ might be skeptical about vaccination. The two groups of nodes are, of course, driven by opposite goals, and each group tries to convince the largest segment of the population.

In the case of competing diffusion processes, the concept of node fragility which we introduced herein could be generalised to define an individual's propensity to accept/reject a given message M. At the same time, the budget could, instead, quantify the effort incurred by a group to spread their message, M or by the opponents to spread the opposite message $\overline{M}$. Numerical simulations could help reveal to what extent parameters such as the topology of the network (*e.g.*, the degree of a node, the fraction of neighbours in favour of M vs.

the fraction against and so on). In the end, the susceptibility of a node to accept/reject a message M and the budget will determine the success (prevalence) of M over $\overline{M}$ or *vice versa*.

REFERENCES

[1] Rediet Abebe, Jon M. Kleinberg, David C. Parkes, and Charalampos E. Tsourakakis. Opinion dynamics with varying susceptibility to persuasion. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD*, pages 1089–1098, London, UK, 2018. ACM.

[2] Dimitri P Bertsekas and John N Tsitsiklis. *Introduction to probability*. Athena Scientinis, New York, 2000.

[3] Song Bian, Qintian Guo, Sibo Wang, and Jeffrey Xu Yu. Efficient algorithms for budgeted influence maximization on massive social networks. *Proc. of the VLDB Endowment*, 13(9):1498–1510, 2020.

[4] Christian Borgs, Michael Brautbar, Jennifer T. Chayes, and Brendan Lucier. Maximizing social influence in nearly optimal time. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014*, pages 946–957, Portland, Oregon, USA, 2014. SIAM.

[5] Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334(6060):1269–1272, 2011.

[6] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 199–208, Paris, France, 2009. ACM.

[7] Wei Chen, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. In *Proc. of the IEEE International Conference on Data Mining*, pages 88–97. IEEE, 2010.

[8] Wei Chen, Weizhong Zhang, and Haoyu Zhao. Gradient method for continuous influence maximization with budget-saving considerations. In *Proc. of the AAAI Conference on Artificial Intelligence*, pages 43–50, New York, USA, 2020. AAAI.

[9] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proc. of the ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, pages 57–66, 2001.

[10] Robin IM Dunbar. The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, 6(5):178–190, 1998.

[11] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++ optimizing the greedy algorithm for influence maximization in social networks. In *Proc. of the International Conference on World Wide Web*, pages 47–48, 2011.

[12] Qintian Guo, Chen Feng, Fangyuan Zhang, and Sibo Wang. Efficient algorithm for budgeted adaptive influence maximization: An incremental rr-set update approach. *Proceedings of the ACM on Management of Data*, 1(3):1–26, 2023.

[13] David Kempe, Jon M. Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, Washington, DC, USA, 2003. ACM.

[14] David A Kim, Alison R Hwong, Derek Stafford, D Alex Hughes, A James O'Malley, James H Fowler, and Nicholas A Christakis. Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet*, 386(9989):145–153, 2015.

[15] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2, 2007.

[16] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proc. of the ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pages 420–429, 2007.

[17] Julian J. McAuley and Jure Leskovec. Learning to discover social circles in ego networks. In *Proc. of the International Conference on Neural Information Processing Systems (NIPS 2012)*, pages 548–556, Lake Tahoe, NV, 2012. NIPS.

[18] Michael Douglas Murphy, Diego Pinheiro, Rahul Iyengar, Gene Lim, Ronaldo Menezes, and Martin Cadeiras. A data-driven social network intervention for improving organ donation awareness among minorities: analysis and optimization of a cross-sectional study. *Journal of Medical Internet Research*, 22(1):e14605, 2020.

[19] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

[20] Huy Nguyen and Rong Zheng. On budgeted influence maximization in social networks. *IEEE Journal on Selected Areas of Communications*, 31(6):1084–1094, 2013.

[21] Matthew Richardson, Rakesh Agrawal, and Pedro M. Domingos. Trust management for the semantic web. In *Proc. of the International Semantic Web Conference (ISWC 2003)*, pages 351–368, Sanibel Island, FL, USA, 2003. ISWC.

[22] Benedek Rozemberczki and Rik Sarkar. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, CIKM '20, page 1325–1334, New York, NY, USA, 2020. Association for Computing Machinery.

[23] Baruch Shomron. The solicitation of altruistic kidney donations on facebook. *Social Science & Medicine*, 344:116615, 2024.

[24] T. Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skar.*, 5:1–34, 1948.

[25] Wendy Wood. Retrieval of attitude-relevant information from memory: Effects on susceptibility to persuasion and on intrinsic motivation. *Journal of personality and social psychology*, 42(5):798, 1982.

[26] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, MDS '12, New York, NY, USA, 2012. Association for Computing Machinery.