UNIVERSITÀ DELLA
CALABRIA

UNIVERSITA' DELLA CALABRIA

Dipartimento di **Ingegneria Informatica, Modellistica, Elettronica e Sistemistica**

Dottorato di Ricerca in

Information and Communication Technologies

CICLO

XXXIII

TITOLO TESI

EMERGING PROBLEMS IN INFLUENCE PROPAGATION AND MAXIMIZATION

Settore Scientifico Disciplinare ING-INF/05

Coordinatore: Prof. Felice Crupi

Firma _____ Firma oscurata in base alle linee guida del Garante della privacy

Supervisore/Tutor: Prof. Andrea Tagarelli

Firma_____ Firma oscurata in base alle linee guida del Garante della privacy

Dottorando: Dott. Antonio Caliò

Firma _____ Firma oscurata in base alle linee guida del Garante della privacy

UNIVERSITÀ DELLA CALABRIA

Abstract

Dipartimento di Ingegneria Informatica, Modellistica, Elettronica
e Sistemistica (DIMES)

Doctor of Philosophy

Emerging Problems in Influence Propagation and Maximization

by Antonio Caliò

In the last two decades we witnessed the advent and the rapid growth of *online social networks* (OSNs). The impact of their pervasive diffusion on everyday life has been dramatic. In fact, social networks changed the way we interact with each other, the way we access information and the way companies engage with their audience or customers. A major consequence of the broad adoption and diffusion of social networks is the availability of an unprecedented amount of user data, which enables the opportunity for *social* and *network scientists* to investigate and observe many facets of human behaviors. Arguably, one of the most interesting facet is related to the notion of *social influence*.

Following this observation, this research project is mainly centered around the concept of social influence, specifically its propagation and maximization. Therefore, the goal of this thesis is twofold. To begin with, we investigate the complexity of influence propagation in real-world contexts. This leads to the definition of a novel class of diffusion models. Such models represent an attempt to unify, under a well-defined framework, all the aspects that contribute to the inherent complexity of any influence propagation phenomena. Afterwards, we devote our attention to the influence maximization problem. To this purpose, we first provide a detailed characterization of social influence from a topological perspective. Specifically, we want to understand if and to what extent being a good spreader depends on being located into strategic regions of a network.

Finally, we focus on the application of the influence maximization problem. In particular, we address a variant of the original problem, which is especially suitable for viral marketing scenarios. To this end, we propose two different *diversity-sensitive targeted influence maximization* problems. Both proposals share a common intent, which is assessing the benefit of embedding a notion of diversity into the process of the seeds identification. Nonetheless, diversity is considered from two different perspectives: (i) as a function of the topological properties of the nodes; (ii) as a function of some categorical data available on the node level.

"*That gulden I staked upon manque–and there is something in the feeling that, though one is alone, and in a foreign land, and far from one's own home and friends, and ignorant of whence one's next meal is to come, one is nevertheless staking one's very last coin! Well, I won the stake, and in twenty minutes had left the Casino with a hundred and seventy gulden in my pocket! That is a fact, and it shows what a last remaining gulden can do. But what if my heart had failed me, or I had shrunk from making up my mind?*

No: tomorrow all shall be ended!"

F. Dostoevsky

Acknowledgements

First and foremost, I would like to thank my advisor, Prof. Andrea Tagarelli, for believing in me and providing me with this amazing opportunity. Thanks for showing me that relentless work and ambition are the key to succeed in this fascinating and very competitive environment. Clearly, without his mentorship none of this would have been possible.

I would like to thank my two roommates and "fellow sufferers", Domenico and Nicola, who shared with me this complicated journey. They were able to make this three-years run much more fun.

Obviously, I cannot forget to thank all my friends, with the promise that they will not be charged with another big present this time.

Last but not least, I am infinitely thankful to my family. My father Oreste and my mother Rita, who encouraged me through these years. Once again, they made me realize how blessed am I to have such a caring and loving family behind me.

To my beloved partner Alessandra who, more than anyone else, has been a source of light and comfort during these years. With her optimism, or just her presence, she made me stay always focused and motivated, knowing that, no matter what, I would have her by my side.

Finally, as silly as it sounds, I would like to acknowledge my cat, Tata, who might have contributed to this manuscript by constantly jumping on my keyboard.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Starting from the early 2000s, we witnessed the booming of *online social networks*(OSNs). This led to a renovated attention to many research questions related to the field of *social network analysis*, or more in general *network science*.

Most of the merit of this renaissance can be ascribed to the proliferation of web platforms where people can interact with one another (e.g., Facebook, Linkedin, Instagram, etc.). Consequently, nowadays researchers have at their disposal an unprecedented amount of network data, which enabled the development of many challenging and exciting applications and studies. In particular, a rich body of these studies have been devoted to the analysis of *social influence* and *information diffusion*.

As it is emerged from many interesting research works, social influence plays an important role in shaping people's behavior. As an explanatory example, we can consider a famous study published in the *New England Journal of Medicine* [45]. For their experimental assessment, the authors extracted a (offline) social network starting from clinical records. The links between the individuals were discovered by taking into account relationships of different nature (e.g., wife/husband, acquaintances, neighborhood). The purpose of this analysis were to understand if, and to what extent, being obese implies having an obese neighborhood in the constructed social network. Remarkably, a person with an obese friend is 171% more likely to be obese than a person without such friend.

The authors further extended their experiment to other real social life contexts (e.g., musical tastes, wealth or beliefs). Their experiments provide compelling evidence on the profound impact that social influence has on many different aspects of a person's real life [44].

The above results are expression of a well known phenomenon in network science, which is referred as *network homophily*. According to this theory, which is implicitly connected to the model of *preferential attachment*, similarity drives the formation of new social ties in a social network. Intuitively, nodes are more inclined to connect with other nodes if, at some level, they are similar to each other. This tendency is a key-factor for determining social influence, which can be considered as the fuel of any diffusion process. In fact, we can arguably say that network homophily is a fundamental ingredient to trigger a pervasive information diffusion. There is however a major downside. In fact, as regards information diffusion, network homophily implies that users have the tendency to mostly access information from like-minded sources [102]. This is clearly a dangerous habit. In fact, the lack of diversity and pluralism inevitably favors the formation of information bubbles and the consequential *polarization* of a network [64].

To understand the importance of studying how information propagates through a network, we should consider the crucial role that platforms as Twitter had in both the 2016 presidential campaign and the entire presidency of Donald Trump, who actually used the social network as his main communication channel. Therefore, it is fair to say

that information diffusion can have a serious impact on many different scenarios (e.g., the adoption of political standpoints, technical innovations). Moreover, we recognize the significance and the urgency to develop a deeper knowledge on how information propagates through a network.

For the above reasons, in this work we address the inherent complexity underlying any propagation process. We propose in fact a novel class of diffusion models – informally, a diffusion model establishes the rules behind an information diffusion process – with the purpose of capturing the complexity of any real world propagation phenomena.

Among all the different applications of social influence analysis, *viral marketing* (also known as *word of mouth marketing*) can be regarded as the "poster" application. The goal of a viral marketing campaign is to detect, and then to activate, a small number of *influential* individuals in a social network, so to reach the largest possible fraction of users, leveraging on the virality of the propagation process. This vision led to the definition of one of the main algorithmic problems in the context of information diffusion, i.e., the *influence maximization* (IM) problem [97]. IM asks to find a set of k users in an online social network that has the maximum *influence spread*, i.e., it activates the largest number of users. Theoretically, influence maximization is a very challenging task, it is indeed an **NP**-hard problem. As a consequence, it is extremely difficult to design effective solutions that are also able to scale up to big social networks.

The importance and the significance of IM goes beyond viral marketing. In fact, IM is a cornerstone for a family of applications in seemingly different domains and settings. For instance, similarities can be found between influence maximization and network monitoring, i.e., the problem of determining the best spots to locate a set of expensive sensors, so that any malfunction can be detect as quick as possible [114]. There are also other algorithmic problems that support the relevance of IM, such as rumor control [24, 82], and social recommendation [198].

Over the years, researchers have also proposed a number of different variations on the IM problem. An interesting example is the *targeted influence maximization* problem. As compared with the classic formulation of the problem, the targeted version has a major difference, which makes him particularly suitable to address a marketing scenario. That is, instead of trying to activate the largest possible fraction of the entire network, a solution to the targeted influence maximization problem aims to maximize the engagement among a particular portion of the users base. Clearly, this represent a more realistic scenario, since the interest of a marketing campaign is typically aimed towards a particular segment of customers. This variant of the problem has been used in various fields. For instance, in [92] it is used in the context of users engagement, where the social capital of a node determines whether or not it has to be considered as target.

An interesting extension to this latter problem, which is addressed in this thesis, is proposed in [26], where the authors introduce a notion of *diversity*. Although there is evidence that diversity is able to enhance the performance in contexts as web searching, ranking and recommendation algorithms [52, 70, 170, 202], it appears to be surprisingly overlooked in the context of information diffusion. However, here we recognize the crucial role that diversity can play in combating the formation of echo chambers, namely those situations that favor the amplification or the reinforcement of beliefs. In fact, it is also known that one major reason behind the formation of an echo chamber is the indeed the lack of exposure to diverse sources of influence.

Contributions

This thesis is concerned with a variety of research topics centered around the concept of social influence, with emphasis on problems related to information diffusion and influence maximization. More specifically, the following research topics can be distinguished.

Modeling complex diffusion. Understanding the dynamics of information diffusion phenomena has emerged as one of the most challenging task in Web science and related fields of research. Since the first applications in contexts related to viral marketing, the design of information diffusion models has provided effective support to address a variety of influence propagation problems, first and foremost, the *influence maximization* problem.

However, one criticism that arises from existing diffusion models is the concern as to whether, and to what extent, they are sufficiently adequate to explain the real complexity of influence propagation in modern social networks. The adherence of a diffusion model with the mechanisms that drive an individual's information consumption becomes even more crucial if we consider the almost invisible boundary between real and virtual social life. Moreover, the process of acquiring and sharing information from reliable sources has often to cope with unlimited *misinformation* spots, which can alarmingly affect everyone's life.

Prompted by the above observations, in this research line which is *addressed in Chapter 3*, we define some of the key-ingredients that any diffusion model should have in order to face the inherent complexity of real world information diffusion. We then embed these factors into a novel class of diffusion models, named *Friend-Foe Dynamic Linear Threshold Models* (F^2DLT). The following aspects are essential constituents of our proposed models: (i) account for different kinds of social ties between users; (ii) incorporate time-dependent variable to represent the latency of any propagation process; (iii) account for users hesitation or inclination towards the adoption of an information item; (iv) enable the possibility for users to change their opinion towards alternative information items

Topological characterization of social influence. Understanding and measuring the spread of "contagion" of an individual is a problem that has attracted the attention of different research communities, e.g., physics, biology, epidemiology and network science. One of the most interesting and well studied problem in this area is related to the identification of the most effective spreaders, i.e., the most influential nodes, in a network. These studies have a very broad impact, as their application is significant in different domains, ranging from the diffusion of information/misinformation to the spread of viruses.

Several heuristics have been proposed to approximate the nodes' influence potential, mostly based on some notion of centrality (e.g., degree centrality, PageRank, betweenness centrality). In recent years, in contrast with the above approach, which considers social influence from a node-centric perspective, many have explored the effectiveness of meso-scale properties in predicting a node influence.

Remarkably, this latter approach turned out to be very promising. In fact, most of the times, properties such as the core-index, i.e., the index assigned by a *core-decomposition* method, have proven to provide better insights on the actual spreading potential of a node, as opposed to classic centrality measures.

However, we notice that these studies are mostly conducted on *undirected* graphs and under classic epidemic models. We recognize that this is an unusual setting in the

context of influence propagation and maximization, which is the main focus of this work. For this reason, in this line of research, which is *addressed in Chapter 4*, we aim at producing an extensive analysis to understand where state-of-the-art algorithms for influence maximization locate their best spreaders with respect to a variety of graph decomposition methods.

Our main goal is to understand if graph decomposition methods can consistently support the identification of subnetworks where nodes have a good influence-spreading potential.

Embedding Diversity into Targeted Influence Maximization problems. Online social networks are arguably the preferred communication tool for spreading information, or more in general, to reach out people. They can be considered as the privileged ground for any *viral marketing* campaign. The main goal of any marketing campaign is to engage the largest number of individuals, i.e., customers. A viral marketing campaign pursues this goal by exploiting the "word-of-mouth" phenomenon that takes place among the users of a social network.

This scenario inspired a classic optimization problem, namely the *influence maximization problem* (IM). Even though, in its classic formulation, the IM problem addresses the entire network, we believe that, as far a marketing campaign is concerned, a more natural choice would be to narrow the focus on an arbitrarily small portion of the network. Such portion comprises for the *target users* of the campaign. This is in fact the intuition underlying a well-studied variant of the IM problem, commonly referred as *targeted influence maximization*. In this context, we notice that much emphasis is given to the size of the set of early adopters, while the benefits brought by having diversified set of initial influencers is often, surprisingly, overlooked.

Intuitively, *diversity* means engaging people that are different from each others in terms of kind (e.g., age, gender), socio-cultural aspects, or other characteristics.

In this line of research, *addressed in Chapter 5 and Chapter 6*, we aim to include aspects related to diversity into a targeted influence maximization framework. We devise two different approaches when it comes to measure the users diversity. The first approach, proposed in Chapter 5, considers diversity from a topological standpoint. The second approach, discussed in Chapter 6, is based on the assumption that a set of categorical data is available at the node level, then diversity is defined as a function of these categorical values. It is worth noticing that, although Chapter 5 and Chapter 6 address a very similar optimization problem, their approach in measuring nodes diversity is substantially different from each other, as well as are the algorithmic solutions proposed for both problems. For this reason, they are addressed into two separate chapters.

Chapter 2

Background

This chapter offers an overview on preliminary concepts that will help the understanding of the techniques and models adopted in subsequent chapters. We first introduce the *influence maximization* (IM) problem in Section 2.1. In Section 2.2 we introduce the notion of stochastic diffusion model. Specifically, we focus on two major settings under which the influence maximization problem is commonly studied: *progressive diffusion* (Section 2.2.1) and *non-progressive diffusion* (Section 2.2.1). For each of the above two configurations we review some of the most used diffusion models under the IM framework. We conduct a theoretical analysis on the complexity of the IM problem in Section 2.3, highlighting the most fundamental theoretical results related to this problem. Such results are particularly important, since they enable the definition of effective algorithmic solutions. Finally, in Section 2.4 we provide an overview on some of the most well-known and significant algorithmic approaches to IM.

2.1 Essentials on influence maximization

The foundations of Influence Maximization (IM) as an optimization problem were initially posed by Kempe et al. in their seminal work [97]. The problem requires a social network graph $G = \langle V, E \rangle$, which consists of two sets V and E. $V \neq \emptyset$ is the set of nodes, i.e., users of the social network, while E is the set of *ordered* pairs of elements in V, i.e., the edges representing the social links between the users. The input graph is assumed to be *directed*. Therefore, for any node $u \in V$, $N^{in}(u) = \{v|(v,u) \in E\}$ denotes the set of *in-neighbors* of u, while $N^{out}(u) = \{v|(u,v) \in E)\}$. denotes the set of *out-neighbors* of u.

The IM problem asks to find a set of users, i.e., a *seed set*, with size at most k, that maximizes the total *influence* among all the nodes in G. A straightforward example of the IM problem is represented by a viral marketing campaign, where a company wishes to spread the adoption of a particular product, exploiting a "word-of-mouth" phenomenon. Therefore, behind any successful viral marketing campaign there is the ability to exploit the ability of users to interact with each other and consequently to influence each other decisions.

The amount of influence achieved by the detected *seed set* is measured by taking into account the *information diffusion* process, triggered by the seed set, that takes place between the users of the social graph. Typically, the outcome of this diffusion process is referred as *information cascade*. The volume of such cascade, i.e., the number of users involved in the propagation, determines the performance, thus the quality, of a seed set. To quantify the volume of the information diffusion process, we need to define an *influence spread* function. Moreover, we need to understand the dynamics of a diffusion process and to define the rules underlying propagation. These rules are defined in a *stochastic diffusion model*.

In a stochastic diffusion model the diffusion proceeds in discrete time steps (with time $t = 0, 1, 2, \dots$). At each time step, each node $v \in V$ is in a certain state. There are at least two possible states: *inactive* and *active*.

Intuitively, when a node goes from being inactive to being active means that it has adopted the new information, product or idea that is spreading along the network. Conversely, an inactive node is a node that has not adopted the new item that is propagating among the users.

In order to start, any diffusion model requires a seed set, i.e., the set or early adopters, which is active at the very beginning of the process ($t = 0$). At each time step t, the set of active users at that time is denoted by $S_t \subseteq V$.

Definition 1 provides a formal definition for a stochastic diffusion model.

Definition 1 (Stochastic Diffusion Model)**.** *Given a social network graph $G = \langle V, E \rangle$, an initial set of users $S \subseteq V$, a stochastic diffusion model $\mathcal{M}$ defines the randomized process of generating the set of active users S_t for all $t \geq 1$.*

The set of active users at the end of the propagation process, i.e., the final active set, is denoted by $\Phi(S)$, The influence spread function represents the size of the final active set, in *expectation*.

Definition 2 (Influence Spread)**.** *Given a social network graph $G = \langle V, E \rangle$, an initial set of users $S \subseteq V$, and a stochastic diffusion model $\mathcal{M}$, the influence spread of S, denoted by $\sigma(S)$, represents the expected number of users influenced by S.*

$$\sigma(S) = \mathbb{E}[|\Phi(S)|] \tag{2.1}$$

The influence function defined in Definition 2 is the objective function of the influence maximization problem, which is formally defined as follows.

Definition 3 (Influence Maximization Problem)**.** *Given a social network graph $G = \langle V, E \rangle$ a stochastic diffusion model $\mathcal{M}$, $S \subseteq V$ with $|S| \leq k$ of seed-nodes such that the influence spread of S, denoted by $\sigma(S)$, is maximized. That is, compute $S \subseteq V$ such that:*

$$S = \underset{S' \subseteq V \ s.t. \ |S'| \leq k}{\operatorname{argmax}} \sigma(S') \tag{2.2}$$

It should be noted that the IM problem is not defined with respect to a specific stochastic diffusion model. Nonetheless, some diffusion models are more suitable than others as they enable the possibility to design effective and efficient solutions to the problem.

2.2 Diffusion models

Diffusion models have captures the attention of a lot of researches from many different areas, e.g., data mining, databases networks and epidemiology. However, in this section we focus only on those model that are relevant to the IM problem.

We categorize diffusion models into two different classes: *progressive* and *non-progressive*. As regards the first class of models, once a node becomes active at time t it keeps the same state for every subsequent time step $t' \geq t$. On the contrary, in a non-progressive model nodes may switch back and forth between active and inactive states.

Progressive models are best suited for modeling the adoption of a new technology, a new product, etc., as the adoption is commonly regarded as a not reversible action

(e.g, the purchase of a new tablet). On the other hand, non-progressive models are best suited to model the diffusion of different opinions (e.g, the support for a particular political idea).

In this section we describe a selection of the most representative models for each of the two categories. Specifically, Section. 2.2.1 provides an overview on progressive diffusion models, while Section 2.2.2 addresses the class of non-progressive diffusion models.

2.2.1 Progressive diffusion

In any progressive diffusion model the active sets are monotonically non-decreasing, i.e., $S_0 \subseteq S_1 \subseteq S_2 \subseteq \ldots \subseteq \Phi(S_0) \subseteq V$, where S_t denotes the set of active users at time t. Therefore, if a node v belongs to S_t, than $v \in S_{t'}$ for any $t' \geq t$.

For each of the following models we provide the main definition along with an alternative one, which is based on the notion of *live-edge* and *possible world*. More specifically, given a graph $G = \langle V, E \rangle$, each edge is marked as either *live* or *blocked* accordingly to a certain randomized rule. This randomly generated subgraph is referred as *live-edge graph* and it represents a possible world for the propagation process. Under a live-edge graph the diffusion is *deterministic*. Therefore, a node u is able to activate another node v only if there exists a path connecting u to v in the live-edge.

Independent Cascade Model. The Independent Cascade (IC) model is introduced for the first time by Kempe et al. in [97], inspired by the studies on interactive particle systems in [55] and marketing [68, 69].

The IC model requires each edge $(u, v) \in E$ to be associated with an *influence probability* $p_{u,v}$, corresponding to the extent to which node u is able to influence v. Also, the activation attempts are independent from each other. The IC model is formally defined as follow.

Definition 4 (Independent Cascade Model)**.** *The Independent Cascade Model requires a social network graph $G = \langle V, E \rangle$, where each edge u, vE has a an influence probability denoted by $p_{u,v}$, and the initial seed set S. At every time step $t \geq 1$ the following rule applies: for any inactive node $v \in V \setminus (S_{t-1})$, for every node $u \in N^{in}(v) \cap (S_{t-1} \setminus S_{t-2})$ u executed an* activation attempt *by performing a Bernoulli trial with success probability $p_{u,v}$. If the attempt is successful v is added into S_t, therefore we say that u* activates *v at time t.*

Informally, when a node u is activated at time t, in the next time step it has a *single chance* of activating each of its inactive out-neighbors v. If u is not able to activate v, it will not be given with another opportunity to activate v. In case a node v is activated by more than one of its in-neighbors at time t, the outcome is the same, i.e., v will be added to S_{t+1}.

Example 1. *Figure 2.1 shows an example of a diffusion process. Initially at $t = 0$ a single seed, the node z, is activated. At step $t = 1$, z successfully activates u but it fails at activating v and x. At step $t = 2$, u successfully activates v, which at the following step, $t = 3$, fails at activating x. At this point, the diffusion stops since there are no more activation attempts to be performed.*

The independent cascade model is particularly suitable for modeling the diffusion in contexts where a single exposure may be sufficient to activate an individual (e.g., getting infected by a virus). This behavior is typically referred as *simple contagion*.

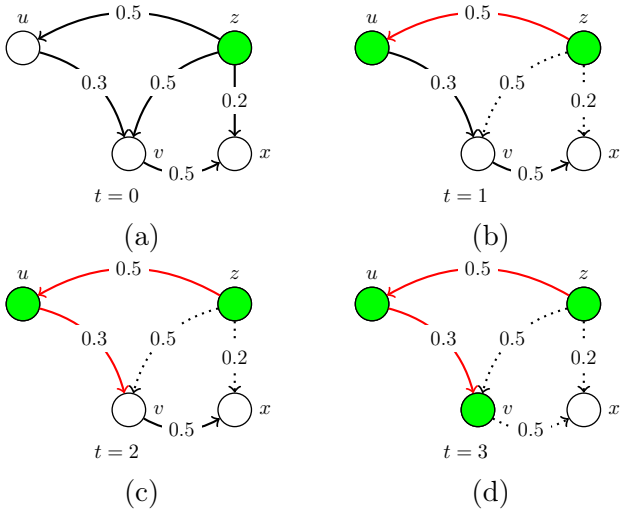


FIGURE 2.1: An example of diffusion process according to the Independent Cascade model. White denotes inactive nodes, and green denotes active nodes. A solid black arc represents an original edge of the graph. The value associated with each arc represents the influence probability. A dotted arc between node u and v denotes that u failed to activate v, while a solid red arc denotes that u has been able to activate v.

The IC model has an equivalent live-edge based formulation. The randomized rule to construct the live-edge graph is described in the following definition.

Definition 5 (Live-edge Graph Model with Independent Cascade Edge Selection.)**.** *Given a social network graph $G = \langle V, E \rangle$, the corresponding live-edge graph model is obtained by selecting each $(u, v) \in E$ with probability p_{uv}.*

Intuitively, to generate a possible world according to the above definition we need to carry out a Bernoulli trial for each edge in the graph. If the trial is successful the corresponding edge is marked as live, otherwise it is marked as blocked, thus it is removed from the resulting graph.

Linear Threshold Model. The Linear Threshold (LT) model is first proposed by Kempe et al. in [97]. As opposed to the IC model, which focuses on simple contagions, the LT model is inspired upon studies in the area of social science related to *threshold behaviors* [29, 75]. A key feature of this model is represented by an aggregate function (e.g., count, sum) that takes into account all the positive inputs received by a target node, which becomes active as the aggregate signal exceeds a certain threshold.

The LT model requires each edge $(u, v) \in E$ to be associated with an *influence weight* $w_{uv} \in [0, 1]$, which reflects the importance of u in influencing v. The weights are normalized such that for each node v the cumulative sum of its incoming edges is at most one, i.e., $\sum_{N^{in}_{+}(v)} w_{uv} \leq 1$. The LT model is formally defined as follow.

Definition 6 (Linear Threshold Model)**.** *The Linear Threshold Model requires a social network graph $G = \langle V, E \rangle$, where each edge $(u, v) \in E$ has a an influence weight value denoted by w_{uv}, and the initial seed set S. Initially, each node $v \in V$ independently selects a threshold $\theta_v \in [0, 1]$, uniformly at random.*

At every time step $t \geq 1$ the following rule applies: for any inactive node $v \in V \setminus (S_{t-1})$, if the total weight of the edges from its active in-neighbors is at least θ_v,

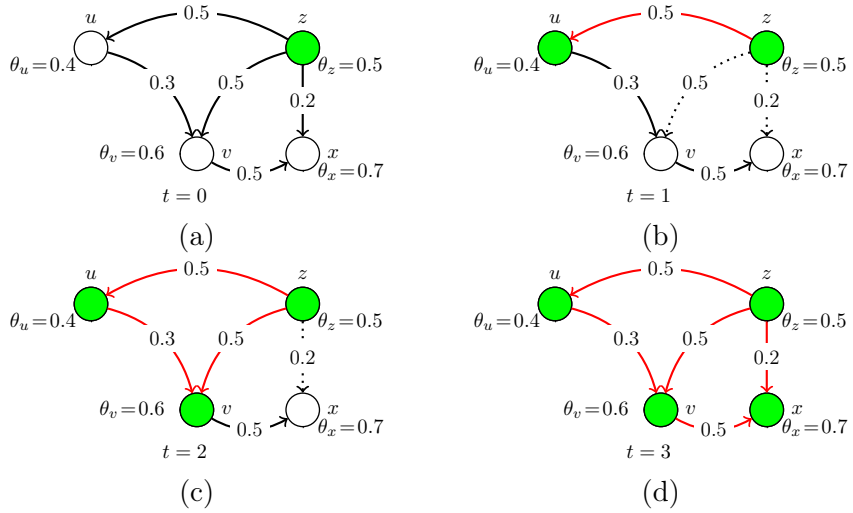


FIGURE 2.2: An example of diffusion process according to the Linear Threshold model. White denotes inactive nodes, and green denotes active nodes. A Solid black arc represents an original edge of the graph. The value associated with each arc represents the influence weight. A dotted arc between node u and v denotes that u is not able to activate v, while a solid red arc denotes that u contributes to v activation. Each node is also associated with the corresponding threshold.

i.e., $\sum_{N_+^{in}(v)\cap S'_{t-1}} w_{uv} \geq \theta_v$*, then* v *is added to* S_t*, thus we say that* v *is activated at time* t*.*

Intuitively, the threshold θ_v represents the effort required to activate v. The larger the value of θ_v the harder will be to activate v. It should be noted that once all the thresholds are determined, the diffusion unfolds in a deterministic fashion. The stochastic nature of the model is therefore represented by the random initialization of the users' thresholds. The random initialization reflects our lack of knowledge about a user specific threshold.

Example 2. *Fig. 2.2 shows an example of a diffusion process. At the beginning, each node selects its activation threshold accordingly to Def. 6. Initially at* $t = 0$ *a single seed, the node* z*, is activated. At step* $t = 1$*,* z *successfully activates* u *but it fails at activating* v *and* x*. At step* $t = 2$*, thanks to the combined effort of* u *and* z*,* v *is activated. At step* $t = 4$ *the combined influence of* v *and* z *triggers the activation of* x*. At this point, the diffusion stops are there as there are no more activation attempts to be performed.*

The Linear Threshold model is particularly suitable for modeling the diffusion in contexts where a target has to be exposed to multiple and independent sources of influence to change its behavior. This dynamics is typical for the adoption of a new and unproven technology, a controversial idea or a costly new product.

More generally, this scenario takes place when people may need positive reinforcement from their friends before taking a particular action. This mechanism is referred as *complex contagion*.

The LT model has an equivalent live-edge based formulation. The randomized rule to construct the live-edge graph is described in the following definition.

Definition 7 (Live-edge Graph Model with Linear Threshold Model Edge Selection.)**.** *Given a social network graph* $G = \langle V, E \rangle$*, the corresponding live-edge graph model is*

obtained by selecting for each node $v \in V$ at most one incoming edge with probability proportional to the weight of the edge. Therefore, given a node v, among all incoming edges $(u, v) \in E$ only an edge is selected with probability w_{uv}, and with probability $1 - \sum_{u \in N^{in}(v)} w_{uv}$ no edge is marked as live. Each node selects its only live incoming edge independently from the other nodes.

Triggering Model. The Triggering (TR) model is introduced by Kempe et al. in [97], as a generalization the aforementioned IC and LT models. It is based on the notion of *triggering set*. For any node $u \in V$, the triggering set, denote by T_u, is a subset of u in-neighbors. A u node selects a subset among all the $2^{|N^{in}(u)|}$ configurations, according to some probability distribution. The triggering set of a node is responsible for its activation.

The TR model is formally defined as follow.

Definition 8 (Triggering Model)**.** *The Triggering Model requires a social network graph $G = \langle V, E \rangle$ and the initial seed set S.*

At $t = 0$, each node $v \in V$ independently selects a triggering set T_v according to some probability distribution.

At every time step $t \geq 1$ the following rule applies: for any inactive node $v \in (V \setminus S_{t-1})$, if it has an active neighbors in its selected triggering set, i.e., $T_v \cap N^{in}(v) \cap S_{t-1} \neq \emptyset$, then v is added to S_t, therefore we say that v is activated at at time t.

Since it is a generalization of the IC and LT model, the TR model is suitable for modeling diffusion in contexts that have either a simple or a complex contagion dynamics.

The TR model has an equivalent live-edge based formulation. The randomized rule to construct the live-edge graph is described in the following definition.

Definition 9 (Live-edge Graph Model with Triggering Model Edge Selection.)**.** *Given a social network graph $G = \langle V, E \rangle$, the corresponding live-edge graph model is obtained by selecting each edge $(u, v) \in E$ if $u \in T_v$.*

Informally, in the live-edge model corresponding to the Triggering model, an edge $(u, v) \in E$ is marked as *live* only if u belongs to the triggering set of v, i.e., $u \in T_v$.

2.2.2 Non-progressive diffusion

This section offers an introduction on a number of non-progressive models. A common denominator for the following models is that they are original conceived for contexts that are apparently distant from the context of influence propagation. Nonetheless, at some point, they have been used to mode information diffusion, and influence maximization.

It should be noted that, the purpose of this section is not to provide a comprehensive introduction on non-progressive diffusion model. In fact, here we wish to provide the reader with all necessary information to understand the main difference between these models and the ones discussed in Section 2.2.1.

First of all, what makes a non-progressive diffusion model is the ability of any active node to turn back to the inactive state. It means the active sets are non-monotonically increasing, i.e., if a node is active at time step t, it will not necessarily be active at any future time step $t' \geq 1$.

In this section we describe a class of well studied non-progressive models, i.e., *epidemic models*. More specifically, we introduce three of the most used models: *SI*, *SIR* and *SIS*.

Finally, we explore the strong connection that there exists between epidemic models and the IC model described in Section 2.2.1

Epidemic Models. Epidemic models were originally conceived to study the spread of diseases among biological populations. However, in recent years researchers have started to adopt these models to represent the diffusion of information in social networks. Unlike every model discussed above, epidemic models are formulated as *fully mixed* models. It means they assume every individual can make contact with any other person. In other words, they consider a network of contacts as a complete graph. Often, epidemic models are treated as continuous time models and their dynamics is represented via a system of differential. The solution to such system of equations provides numerical results on the dynamics of the diffusion among the population.

Analogously to stochastic diffusion models each individual transition between several possible states. The taxonomy used to denote the different states considers the transmission of a disease. Therefore, a node can assume one of the following three states: (i) S which stands for *susceptible*, it means that a node can get the disease if it gets in touch with an infected individual: (ii) I which stands for *infected*, a node in this state is able to transmit the disease to other individuals; R which stands for *recovered*, a node in this state is a node healed from the disease, thus it is not contagious anymore.

Each node passes through the same sequence of transitions and the model is named after this sequence of possible transitions. For instance, the most basic model is the *SI* model. It stands for susceptible-infected model. In allows a node to only transition from susceptible to infected. The rate at which these transitions happen is called *infection rate*, denoted by β.

The *SI* model is formally defined as follows.

Definition 10 (SI Model)**.** *Let s, i and n denote the fraction of susceptible, the fraction of infected nodes and the total number of nodes in the system, respectively. In a small unit of time dt, each infected node makes contact with all other nodes, among which only $s \cdot n$ are susceptible. An infected node infects any susceptible node with probability β. Since there are $i \cdot n$ infected nodes in total, after dt time unit, the reduction to the fraction of susceptible node is $i \cdot n \cdot s \cdot n \cdot \beta / n = \beta \cdot s \cdot i \, dotn$. This leads to the following differential equation:*

$$\frac{ds}{dt} = -\beta \cdot s \cdot i \cdot n$$

Based on the observation that $s + i = 1$, the differential equation in Definition 10 is solvable by the following closed-form:

$$i(t) = \frac{i(0)e^{\beta nt}}{1 - i(0) + i(0)e^{\beta nt}}$$

Where $i(t)$ denotes the fraction of infected nodes ad time t. The solution corresponds to the classic logistic growth curve (or S-shaped) illustrated in Figure. 2.3.

Another interesting model is the *SIR* model. It has the same behavior of the *SI* model, but it accounts for an additional transition. Specifically, once a node is infected, it always transitions to the recovered state. This model is particularly suitable to represent the diffusion of a disease that generates life-time immunity once an individual recover from it.

The rate at which the transitions from I to R happen is referred as *recovery rate*, denoted by γ. Intuitively, this parameter represents the likelihood that an infected

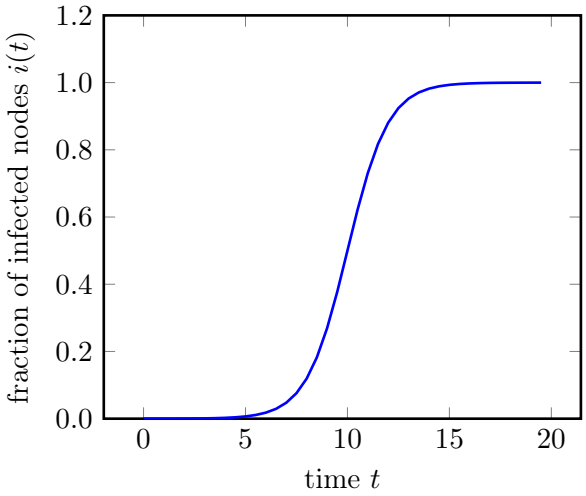


FIGURE 2.3: Logistic growth curve (S-shaped curve) for the SI model.

individual recovers from the disease obtaining an immunity that will prevent him/her to further develop the disease.

The *SIR* model is formally defined as follows.

Definition 11 (SIR Model)**.** *Let s, i, r and n denote the fraction of susceptible, the fraction of infected nodes, the fraction of recovered nodes and the total number of nodes in the system, respectively. In a small unit of time dt, the reduction to the fraction of infected nodes is $\gamma \cdot i$. This leads to the following system of differential equations:*

$$\begin{cases} \frac{ds}{dt} = -\beta \cdot s \cdot i \cdot n \\ \frac{di}{dt} = \beta \cdot s \cdot i \cdot n - \gamma \cdot i \\ \frac{dr}{dt} = \gamma \cdot i \\ s + i + r = 1 \end{cases}.$$

The solution to the above system of equation returns the fraction of nodes for each of the three categories, i.e., susceptible, infected and recovered.

An important metric of the *SIR* model is the *basic reproduction number* $R_0 = \beta n/\gamma$. This parameter represents the average number of new infections caused by an infected node, during the time span of its infection, which on average lasts $1/\gamma$ time units.

Interestingly, when $R_0 > 1$ a significant portion of nodes are infected, on the contrary when the reproduction number is below 1 only a small fraction of nodes are infected, thus the diffusion tends to rapidly stop.

Intuitively, R_0 quantifies the virality of the diffusion process. It depends on both the infection and the recovery rate. Also, when this value exceeds the so called *epidemic threshold* a large fraction of the population will be infected by the disease.

Finally, another commonly used model is the *SIS* model, where an infected node, after it recovers from the disease, always returns to the *susceptible* state. The *SIS* model is particularly suitable to represent the diffusion of a disease for which an individual cannot develop a life-time immunity (e.g., the flu). The *SIS* model is governed by differential equations that are similar to the ones derived for the *SIR* model. The steady state in the *SIS* model is reached as time approaches to infinity. Analogously to the *SIR* model, the epidemic threshold is 1. If the basic reproduction number R_0 is above that threshold then a non-zero fraction of nodes is infected at

the steady state. On the contrary, if R_0 is below the threshold the diffusion stops exponentially, therefore no node is infected at the steady state.

Comparison with the Independent Cascade Model. Epidemic models can also be used with on *contact networks*. As opposed to the classic model formulation, where a node can potentially infect any other nodes in the network, under this setting an individual can only infects the nodes within its neighborhood.

The *SIR* model on a contact network is extremely close to the IC model. In fact, as described in Definition 12 and Definition 13, given an instance of the *SIR* model we can derive an equivalent instance of IC model, and vice versa.

Definition 12 (Conversion from SIR to IC)**.** *Given an* SIR *model defined on a social graph* $G = \langle V, E \rangle$ *with parameters* β *and* γ*, an equivalent instance of an IC model can be derived as follows.*

For each edge $(u, v) \in E$*, the influence probability is defined as:* $p_{uv} = 1 - (1 - \beta)^{1/\gamma} \simeq \beta/\gamma$*. In fact, each infected node is given with* $1/\gamma$ *attempts to activate* v*, each of which with success probability* β*.*

Definition 13 (Conversion from IC to SIR)**.** *Given an IC model defined on a social graph* $G = \langle V, E \rangle$*, where each edge* (u, v) *is associated with an influence probability* p_{uv}*, an equivalent instance of the* SIR *model can be derived as follows. First we set* $\gamma = 1$*, so that each node remains infected only for a single a unit of time, thus it has a single chance of infecting its neighbors. Then we define a pairwise specific infection rate for each edge* (u, v)*, so that* $p_{uv} = \beta_{uv}$*, where* β_{uv} *denotes the probability that* u *infects* v*.*

It should be noted that epidemic models are mostly addressed by studies not necessarily related to the context of information diffusion or influence maximization. For example, in [30, 63, 147] is investigated how the network topology affects the epidemic threshold. In [23, 46, 144] the authors are interested in investigating the efficacy of immunization program under different models.

2.3 Complexity of influence maximization

Here we assess the computational complexity of the influence maximization problem under the main progressive diffusion models presented in Section 2.2.1, namely the IC and the LT models.

Solving an instance of an IM problem under any stochastic diffusion models involves dealing with two different tasks: (i) the *influence computation*, i.e., evaluating the influence spread $\sigma(\cdot)$; (ii) solving the combinatorial problem, that is finding the optimal seed set. In the following, we argue that both problems are extremely hard.

Complexity of the influence computation. Under both the IC and the LT model computing the influence spread of a seed set of users is a #**P**-hard problem. The class #**P** contains counting problems whose counterpart decision version is in **NP**.

A **NP** problem asks if a particular instance has a solution (e.g., if a conjuctive-normal-form (CNF) formula has a satisfying assignment), while a #**P** problem asks to evaluate the number of possible solutions for the given instance (e.g., how many satisfying assignments to a CNF formula). A problem $\mathcal{P}$ is said to belong to #**P**-complete if it belongs to #**P** and every problem in this class can be reduced to $\mathcal{P}$ with a polynomial time reduction. $\mathcal{P}$ is also said to belong to #**P**-hard if there exist a problem in #**P**-complete that can be reduced to $\mathcal{P}$ in polynomial time.

It should be noted that a #**P**-hard problem is also at least a **NP**-hard problem, since counting the number of solutions implicitly answers to whether or not a solution exists.

Theorem 1 (Theorem 1 of [37] and Theorem 1 of [197])**.** *Computing the influence spread $\sigma(S)$ in a social graph $G = \langle V, E\rangle$, with seed set S is #**P**-hard under both the IC and LT models, even when $|S|= 1$*

Proof (sketch). The #**P**-hardness under the IC model can be proven with a reduction from the #**P**-complete problem *s-t connectedness counting problem* in a directed graph. The #**P**-hardness under the LT model can be proven with a reduction from the #**P**-complete *simple path counting problem* [192]. See [37, 197] for a complete reduction.

□

The above theorem implies the #**P**-hardness of the influence maximization problem.

Corollary 1. *The influence maximization problem is #**P**-hard under both the IC and LT models, even for $k = 1$.*

Complexity of the combinatorial problem. Besides the hardness of the influence computations task, since the IM problem has a combinatorial nature, it contains some **NP**-complete problems as special cases. Based on this observation, the following theorem proves the **NP**-hardness of the problem.

Theorem 2 (Theorem 2.4 and 2.7 in [97])**.** *Influence maximization problems under both IC and LT models contain #**P**-complete problems as special cases, and thus are **NP**-hard. Moreover, influence maximization under the IC model is **NP**-hard even if the influence computation can be done in polynomial time.*

Proof (sketch). Influence maximization under the IC model contains the **NP**-complete set cover problem as a special case, while influence maximization under the LT model contains the **NP**-complete vertex cover problem as a special case. Consequently, IM is **NP**-hard under both the IC and LT models (see the proofs of Theorem 2.4 and 2.7 of [97]).

Moreover, for influence maximization under the IC model, its set cover special case is a bipartite graph with arcs from one partition to the other partition, with influence probabilities as 1 on all arcs. In such special cases, influence computation given any seed set is trivial by computing the size of the reachable set of nodes from the seed set, so influence computation can be done in polynomial time. □

2.3.1 Greedy approach to influence maximization

In the previous section we show that IM is hard. More specifically, there are two sources of hardness: the combinatorial nature of the problem and the influence computations task. A possible approach to address the combinatorial nature of the problem is represented by a greedy selection strategy, described in Algorithm 1. The adoption of a greedy strategy is motivated by the following theorem.

Theorem 3 ([152])**.** *Let $S^* = \text{argmax}_{|S|\leq k} f(S)$ be the set maximizing $f(S)$ among all sets with size at most k. If f is a* monotone *and* submodular *set function and $f(\emptyset) = 0$, then if S^g is the solution provided by a greedy selection we have:*

$$f(S^g) \geq (1 - \frac{1}{e}) f(S^*)$$

Algorithm 1 *Greedy*$(\mathcal{G}, k, f)$: a general greedy algorithm

Input: The diffusion graph $\mathcal{G} = \langle V, E \rangle$; the size of the returned seed set k; a monotone submodular function f.
Output: Seed set S of size k.

1: $S \leftarrow \emptyset$
2: **for** $i \leftarrow 1$ **to** k **do**
3: $\quad u \leftarrow \text{argmax}_{w \in V \setminus S}(f(S \cup \{w\}) - f(S))$
4: $\quad S \leftarrow S \cup \{u\}$
5: **end for**
6: **return** S

Proof. Let $S^* = \{s_1^*, s_2^*, \ldots, s_k^*\}$. Let the greedy algorithm in Algorithm 1 select elements $s1, s2, \ldots, s_k$ in this order, such that $S^g = \{s_1, s_2, \ldots, s_k\}$. Let $S_i^* = \{s_1^*, \ldots, s_i^*\}$ and $S_i^g = \{s_1, \ldots, s_i\}$ for $i = 1, dots, k$ and $S_0^* = S_9^g = \emptyset$. Then, for every $i = 0, 1, \ldots, k-1$, we have:

$$\begin{array}{rcll} f(S^*) & \leq & f(S_i^g \cup S^*) & \text{monotonicity of } f \\ & = & f(S_i^g \cup S_{k-1}^* \cup \{k_k^*\}) & \\ & \leq & f(S_i^g \cup \{s_k^*\}) - f(S_i^g) + f(S_i^g \cup S_{k-1}^*) & \text{submodularity of } f \\ & \leq & f(S_{i+1}^g) - f(S_i^g) + f(S_i^g \cup S_{k-1}^*) & \text{line 3 of Algorithm 1} \\ & \leq & k(f(S_{i+1}^g) - f(S_i^g)) + f(S_i^g) & \text{repeating the above steps } k \text{ times} \end{array}$$

Arranging the inequality we have:

$$f(S_{i=1}^g) \geq (1 - \frac{1}{k}) f(S_i^g) + \frac{f(S^*)}{k} \tag{2.3}$$

Multiplying by $(1 - 1/k)^{k-i-1}$ on both sides of the above inequality and then adding up all inequalities for $i = 0, 1, \ldots, k-1$, we obtain the final result:

$$\begin{array}{rcl} f(S^g) & = & f(S_k^g) \\ & \geq & \sum_{i=0}^{k-1} (1 - \frac{1}{k})^{k-i-1} \frac{f(S^*)}{k} \\ & = & (1 - (1 - \frac{1}{k})^k) f(S^*) \\ & \geq & (1 - \frac{1}{e}) f(S^*) \end{array}$$

□

The greedy approach referred by Theorem 3 is described in Algorithm 1. Intuitively, a greedy selection iteratively adds the node with the largest marginal gain with respect to the current solution.

As stated in Theorem 3, Algorithm 1 provides a solution with approximation guarantee only if the objective function, i.e., the influence spread function $\sigma(\cdot)$, has both the properties of *monotonicity* and *submodularity*. These two properties are formally defined as follow.

Definition 14 (Monotonicity)**.** *Let f be a set function defined over a domain denoted by V, so that $f : 2^V \mapsto \mathbb{R}$. We say f is (non-decreasing) monotone if and only if $f(S) \leq f(T)$ for any $S \subseteq T \subseteq V$.*

Definition 15 (Submodularity)**.** *Let f be a set function defined over a domain denoted by V, so that $f : 2^V \mapsto \mathbb{R}$. We say f is submodular if and only if $f(S \cup \{v\}) - f(S) \geq f(T \cap \{v\}) - f(T)$, for any $S \subseteq T \subseteq V$ and for any $v \in V$.*

Informally, monotonicity means that adding more nodes to a seed set does not reduce its influence spread. Submodularity can be understood as diminishing marginal gains of the influence spread.

Therefore, Whenever the *influence spread function* of a diffusion model, i.e., $\sigma(\cdot)$, satisfies the above properties,we can successfully approximate the optimal solution accordingly to Theorem 3.

Fortunately, as stated in the following theorem, under the two main progressive diffusion models, i.e., the IC and LT models, the influence spread is both monotone and submodular.

Theorem 4 (Theorem 2.13 [33]). *The influence spread function $\sigma(\cdot)$ in both the independent cascade (IC) and the linear threshold (LT) models is monotone and submodular.*

Proof. In order to prove the above theorem we use the equivalent live-edge graph obtained accordingly to the randomized rules defined in the Definition 5 and the Definition 7 for the IC and the LT model, respectively. Given a social network graph $G = \langle V, E \rangle$, we denote with $\mathcal{G}$ the set of all possible live-edge graphs of G. Each of the possible live-edge graph instances G_L has a probability denoted by $Pr(G_L)$. Regardless of the diffusion model, we can define the influence spread function with a possible world semantics as in the following equation.

$$\sigma(S) = \sum_{G_L \in G} Pr(G_L) |R_{G_L}(S)|$$

IN the above equation $R_{G_L}(S)$ denotes the set of nodes reachable by any node in S under the live-edge graph instance R_{G_L}.

Since the linear combination of monotone (resp. submodular) functions with non-negative coefficients – the probabilities associated with each possible live-edge graph instance – is also monotone (resp. submodular), to prove the monotonicity (resp. submodularity) of the influence spread function $\sigma(\dot{})$ it is sufficient to prove that for any live-edge graph G_L, $|R_{G_L}(\dot{})|$ is monotone (resp. submodular).

The monotonicity of $|R_{G_L}(\dot{})|$ is straightforward, in fact the number of nodes reached by the seed set S cannot decrease as we add more nodes in S.

To prove the submodularity of the function we nee to show that for any two subsets $S \subseteq T \subseteq V$ and a node $v \in V$, $R_{G_L}(T \cup \{v\}) \subseteq R_{G_L}(S \cap \{v\}) \setminus R_{G_L}(S)$.

For any node $u \in R_{G_L}(T \cup \{v\}) \setminus R_{G_L}(T)$, u is reachable from $T \cup \{v\}$ but it is not reachable from T. Consequently, u must be reachable from v. Therefore, there is no node in T th at is able to reach u. Since $S \subseteq T$, u cannot be reached by any node in S neither, as it would otherwise belongs to $R_{G_L}(T)$, leading to a contradiction. We can say $|R_{G_L}(\cdot)|$ is submodular under every possible G_L, thus $\sigma(\cdot)$ is submodular. □

There is a major drawback with Algorithm 1: it needs to evaluate the influence spread function $n \cdot k$ times. Since computing these values is a very expensive task (#**P**-hard), we first need to overcome this problem before we can apply the greedy approach in reasonable time.

The solution proposed by Kempe et al in [97] is to estimate the influence spread with Monte Carlo (MC) simulations. Thus, the value of the influence spread associated with S is approximated by computing the average number of activated nodes over a large number of simulations. Algorithm 2 shows the Monte Carlo based greedy algorithm for influence maximization. The structure is the same as Algorithm 1, the

Algorithm 2 *Greedy*$(\mathcal{G}, k)$: Monte Carlo greedy algorithm for influence maximization

Input: The diffusion graph $\mathcal{G} = \langle V, E \rangle$; the size of the returned seed set k; a monotone submodular function f.
Output: Seed set S of size k.

$S \leftarrow \emptyset$
for $i \leftarrow 1$ **to** k **do**
  $u \leftarrow \text{argmax}_{w \in V \setminus S} MC - Estimate(S \cup \{w\}, G)$
  $S \leftarrow S \cup \{u\}$
end for
return S
procedure MC-Estimate(S, G)
$cnt \leftarrow 0$
for $i \leftarrow 1$ **to** N **do**
  Simulate the diffusion process on graph G with seed set S
  $cnt \leftarrow cnt + |\Phi(S)|$
end for
return cnt/N

only exception is represented by the use of the procedure MC-Estimate for computing the influence spread.

The accuracy of the estimate provided by MC-Estimate depends on the number of different diffusion processes N. In general, larger values of N correspond to a higher accuracy.

However, regardless of the level of accuracy offered by the Monte Carlo estimation, we cannot rely on the results in Theorem 3, as it requires the influence spread to be computed exactly. In fact, the MC procedure always returns a multiplicative γ-error estimate of the original set function f. It means that for any subset $S \subseteq V$ and for any $\gamma > 0$ we have $|f(S) - \hat{f}(S)| \leq \gamma v$, where $\hat{f}$ denotes the estimated of the set function.

Nevertheless, in [33] the authors show that with an adjustment to Theorem 3 it is possible to prove that the Algorithm 2 achieves a $(1 - 1/e - \epsilon)$-approximation of the optimal solution.

Theorem 5 (Theorem 3.6 of [33]). *Let* $S^* = \text{argmax}_{|S| \leq k} f(S)$ *be the set maximizing* $f(S)$ *among all sets with size at most* k*, where* f *is a* monotone *and* submodular *set function and* $f(\emptyset) = 0$*. For any* $\epsilon > 0$*, for any* γ *with* $0 < \gamma \leq \frac{\epsilon/k}{2+\epsilon/k}$*, for any function estimate* $\hat{f}$ *that is a multiplicative* γ*-error estimate of the function* f*, the output* S^g *of the Monte Carlo based greedy algorithm guarantees:*

$$f(S^g) \geq (1 - \frac{1}{e} - \epsilon) f(S^*)$$

Proof. The proof follows the same structure as the proof of Theorem 3. Unfortunately, we can no longer claim that $f(S_i^g \cup \{s_k^*\}) \leq f(S_{i+1}^g)$ since the element s_{i+1} found by the greedy algorithm with respect to the approximated function $\hat{f}$ may not be the optimal element with respect to the correct function f.

Suppose that $\bar{s}_{i+1} = \text{argmax}_{w \in V \setminus S_i^g}(f(S_i^g \cup \{w\}) - f(S_i^g)) = \text{argmax}_{w \in V \setminus S_i^g} f(S_i^g \cup \{w\})$, then we have:

$$\begin{aligned}
f(S_i^g \cup \{^*_k\}) &\leq f(S_i^g \cup \{\bar{s}_{i+1}\}) && \text{by } \gamma\text{-error estimate of } \hat{f} \\
&\leq \tfrac{1}{1-\gamma} \hat{f}(S_i^g \cup \{\bar{s}_{i+1}\}) && \text{by algorithm } \mathit{Greedy}(k, \hat{f}) \\
&\leq \tfrac{1}{1-\gamma} \hat{f}(S_i^g \cup \{s_{i+1}\}) && \text{by algorithm } \mathit{Greedy}(k, \hat{f}) \\
&\leq \tfrac{1+\gamma}{1-\gamma} f(S_i^g \cup \{S_i^g \cup \{s_{i+1}\}\} && \text{by } \gamma\text{-error estimate of } \hat{f}
\end{aligned}$$

Plugging the above inequality into Inequality 2.3 we have:

$$f(S_{i+1}^g) \geq \frac{1-\gamma}{1-\gamma}((1 - \frac{1}{k}) f(S_i^g) + \frac{f(S^*)}{k})$$

Multiplying both sides by $((1-1/k)(1-\gamma)/(1+\gamma))^{k-i-1}$ and then adding up all inequalities for $i = 0, 1, \ldots, k-1$, we have:

$$\begin{aligned}
f(S) &= f(S_k^g) \\
&\geq \textstyle\sum_{i=0}^{k-1} (\frac{(1-1/k)(1-\gamma)}{1+\gamma})^{k-i-1} \cdot \frac{1-\gamma}{(1+\gamma)k} \cdot f(S^*) \\
&= \frac{1-(\frac{1-\gamma}{1+\gamma})^k (1-\frac{1}{k})^k}{(1+\gamma)k/(1-\gamma)-k+1} f(S^*) \\
&\geq \frac{1-(\frac{1-\gamma}{1+\gamma})^k \cdot \frac{1}{e}}{(1+\gamma)k/(1-\gamma)-k+1} f(S^*) \\
&\geq \frac{1-\frac{1}{e}}{(1+\gamma)k/(1-\gamma)-k+1} f(S^*) \\
&\geq (1-\tfrac{1}{e})(1 - \tfrac{(1+\gamma)k}{1-\gamma} + k) f(S^*) && \text{since } \frac{1}{1+x} \geq 1-x, \forall x 0 \\
&\geq (1 - \tfrac{1}{e} - (\tfrac{(1+\gamma)k}{1-\gamma} - k)) f(S^*) \\
&\geq (1-\tfrac{1}{e}) f(S^*) && \text{since } \gamma \leq \frac{\epsilon/k}{2 + /k}
\end{aligned}$$

□

he time complexity of Algorithm 2 is $O(k|V| \cdot N \cdot |E|)$. In fact, it runs for k iterations, at every iteration, for each node in $V \setminus S$ an MC simulation is carried out, which means starting N diffusion processes, each of which can take at most $|E|$, i.e., the traversal of the entire graph.

Theoretically, we can provide a bound on N, i.e., the number of simulations, so that Algorithm 2 achieves the $(1-1/e-\epsilon)$ approximation ratio with high probability, as stated in the following theorem.

Theorem 6 (Theorem 3.7 of [33]). *With probability $1 - 1/|V|$ Algorithm 2 achieves $(1-1/e-\epsilon)$ approximation ratio in time $O(k|V| \cdot N \cdot |E|)$, for both IC and LT models*

Proof (sketch). By Theorem 5, to achieve the $(1-1/e-\epsilon)$ approximation ratio, we need to guarantee the influence spread estimate error $\gamma \leq \frac{\epsilon/k}{2+\epsilon/k}$. For any valid ϵ it is sufficient to use $\gamma \leq \frac{\epsilon}{3k}$. We then use the Hoeffding bound for the number of required simulation N. Let $X_1, X_2, \ldots, X_N$ be a set of N independent random variables. Each X_i varies in a range $[a_i, b_i]$. Let $\bar{X} = \sum_{i=1}^N X_i/N$. According to the Hoeffding bound we have:

$$Pr(|\bar{X} - \mathbb{E}(\bar{X})) \leq 2exp(-\frac{2N^2t^2}{\sum_{i=1}^N (b_i - a_i)^2})$$

TABLE 2.1: Summary of influence maximization algorithms under classical diffusion models

Category	Method	Approximation
Simulation	MC-Greedy[97]	$1-1/e-\epsilon$
	CELF[114]	$1-1/e-\epsilon$
	CELF++[74]	$1-1/e-\epsilon$
	GCA[197]	$1-1/e-\epsilon$
Proxy	DegDist [40]	N.A.
	GroupPR [134]	N.A.
	IRIE [96]	N.A.
	SPM[100]	$1-1/e$[1]
	MIA [38]	$1-1/e$[1]
	LDAG [35]	N.A.
	Simpath [72]	N.A.
Sketch	NewGreIC [40]	N.A.
	StaticGreedy [43]	$1-1/e-\epsilon$
	StaticGreedyDU [43]	$1-1/e-\epsilon$
	PrunedMC [156]	$1-1/e-\epsilon$
	RIS[20]	$1-1/e-\epsilon$
	TIM/TIM+ [189]	$1-1/e-\epsilon$
	IMM [188]	$1-1/e-\epsilon$
	SSA [154]	$1-1/e-\epsilon$

For the case of Monte Carlo simulations, X_i is the number of active nodes in the i-th simulation for some seed set S, and $a_i = 1, b_i = |V|$.$\bar{X}$ is the Monte Carlo estimate. $\mathbb{E}(\bar{X})$ is the true influence spread $\sigma(S)$, and $t = \frac{\epsilon}{3k}\sigma(S)$. When $t = \frac{\epsilon}{3k}\sigma(S)$, the event $\{|\bar{X} - \sigma S| \leq t\}$ implies that $\bar{X}$ is a $\frac{\epsilon}{3k}$-error estimate of $\mathbb{E}(\bar{X}) = \sigma S$. Thus, it the event $\{|\bar{X} - \sigma S| \leq t\}$ is true for any seed set S, then we have a $\frac{\epsilon}{3k}$-error estimate of $\sigma(\cdot)$, as a consequence the greedy algorithm achieves the $(1-1/e-\epsilon)$ approximation ratio. Finally, in the greedy algorithm, we estimate influence spread for totally nk seed sets. Thus, we want the probability of exceeding the multiplicative $\frac{\epsilon}{3k}$ error for each seed set be at most $\frac{1}{n^2k}$, so that by union bound we can have probability of at most 1 that some estimate exceeds the $\frac{\epsilon}{3k}$ error bound. Putting all these into the above inequality, we have that the number of runs required by each MC simulation is $\Theta(\epsilon^{-2}k^{-2} \cdot |V|^{-2}\log(|V|^2k))$. It means that the running time of the algorithm is $O(\epsilon^{-2}k^3 \cdot |V|^3|E| \cdot \log|V|)$ □

2.4 Algorithmic solutions

Despite its polynomial time complexity and the strong approximation ratio, Algorithm 2 is far from being practical efficient. Its biggest problem is with the Monte Carlo simulations required to estimate the influence spread. For this reason, since the seminal work of Kempe et al. [97], many studies have tried to overcome the efficiency issue with the classic Monte Carlo based greedy algorithm.

This section provides an overview on the main solutions for IM. Algorithms are categorized as in the excellent survey [125]. More specifically, we have three different classes: (i) *simulation based*; (ii) *proxy-based*; (iii) *sketch-based*.

For each class we highlight the main strengths and weaknesses as regards the following aspects:

[1]The approximation bound is with respect to the reduced model

- *Model Generality*: the ability to embrace different diffusion models with minimum effort
- *Practical Efficiency*: the ability to efficiently solve the IM problem on large networks
- *Theoretical Efficiency*: the ability to provide solutions with approximation guarantee

Table 2.1 provides a guide to the following discussion.

Simulation Based. All the algorithms in this category adopt Monte Carlo simulations to estimate the influence function. However, unlike Algorithm 2, they also leverage on meta-heuristic search strategies in order to speed up the combinatorial optimization task.
Pros: The main advantage is their *model generality*. They can be adapted to work with any diffusion model – It is sufficient to incorporate the specific Monte Carlo estimation procedure. Another advantage is their *theoretical efficiency*. As long as the diffusion model enable a monotone and submodular influence function, we can rely on the approximation ratio in Theorem 5.
Cons: Simulation based algorithms suffer of a poor practical efficiency. The main reason is ascribed to the expensive the Monte Carlo simulations.

Proxy based. The key idea underlying this approach is to devise an efficient strategy to *approximate* the influence function. Therefore, these algorithms are concerned with the definition of a good proxy for the influence maximization function, instead of having an accurate estimate.
Pros: The main advantage is their practical efficiency.
Cons: Typically, solutions provided by proxy-based algorithms have not any theoretical guarantee. Consequently, the solution provided by any of these algorithm might be arbitrarily bad.

Sketch-based. The key idea underlying this approach is to combine together practical efficiency and theoretical efficiency. Solutions provided by sketch-based algorithms couple a theoretical approximation guarantee with a reasonable execution time. At the center of their definition there is the concept of *sketch*. A sketch is a realization of the influence network graph under a specific diffusion model, i.e., a possible world.
Pros: The main advantage is their ability to ensure both theoretically and practically efficiency.
Cons: There is lack in terms of model generality. In fact, it is not easy to incorporate different diffusion models, since they must enable the definition of an equivalent *sketch* based formulation to compute the influence spread (e.g., the reachability set on a live-edge graph model as for the IC and LT models).

2.4.1 Simulation based

The greedy procedure described in Algorithm 2 is a clear expression of a simulation based approach. As shown in Theorem 6, the time complexity of is prohibitively expensive to deal with large graphs. Therefore, many researchers have tried to optimize the simulation based greedy framework. In their attempts, it is possible to identify two orthogonal strategies, which are discussed below.

Reducing the number of MC simulations. In each iteration of the greedy algorithm, there are exactly $|V \setminus S|$ MC simulations to estimate the marginal gain of each node. However, most of these computations could be avoided as they involve nodes

with insignificant influence. Therefore, the intuition underlying this approach is to prune all the nodes that have irrelevant marginal influence.

This is the main idea behind the CELF algorithm [114], which exploits the submodularity of the influence function to reduce the number of influence computation. Specifically, let $\Delta(u|S_i) = \sigma(S_i \cup \{u\}) - \sigma(S_i)$ represent the marginal gain of node u with respect to the set S_i, that is the set constructed by the greedy algorithm up to the i-th iteration. According to the definition of submdoularity, $\Delta(u|S_i)$ is also an upper bound for the marginal gain of u with respect to any S_j so that $S_i \subseteq S_j$. Based on this observation, the CELF algorithm first computes the marginal gain $\Delta(u|\emptyset)$ for any $u \in V$ and it adds the best node into S_1. At each subsequent iteration $i = 2, \ldots, k$, CELF visits each node in $u \in V \setminus S_{i-1}$ in a descending order of their upper bounds of $\Delta(\cdot|S_i)$ and it computes $\Delta(u|S_{j-1})$ by the means of an MC simulation.

The main trick is the early termination on the marginal gains update, which prevents the algorithm to visit the entire node set. In fact, if the maximum upper bound of an unvisited node is smaller than the maximum marginal gain among all the visited nodes, then we can stop updating the upper bound of the remaining unvisited nodes. In each iteration, CELF the node with the largest marginal gain.

Even though CELF does not improve the theoretical time complexity, the early termination technique is able to provide a better practical efficiency, as it be up to 700 times faster than the greedy algorithm.

Following the above approach, in [74] is proposed the CELF++ algorithm, which further reduces the number of MC simulations required by CELF. The main difference between the two algorithms is that CELF++ computes both $\Delta(u|S_i)$ and $\Delta(u|S_i \cup \{v_j^u\})$ for each user u, where v_j^u is the node having the largest marginal gain among all the nodes visited before u. In this way, CELF++ prevents the computation of $\Delta(\cdot|S_{i+i})$ if $S_{i+1} = S_i \cup \{v_j^u\}$ at the $i+1$ iteration.

Despite this slight change in the way the two algorithms update the nodes' marginal gain, the speedup brought by CELF++ over CELF is often negligible [5, 141].

Reducing MC complexity. This approach aims to speed up each individual MC simulation, by rearranging the influence computation task. The algorithm proposed in [197], i.e., the Community-based Greedy Algorithm (CGA), is clear example of this approach. The key idea is to use a divide-and-conquer paradigm. At the beginning, the graph is partitioned into a number of different communities. In the divide step GCA computes the influence of each node within its community, while in the conquer step the community-wise influence of each node is combined to decide the final seed set. The main benefit derived by this approach is that the influence computations are usually faster, since they are carried out with respect to only a portion of the entire graph, i.e., the subgraph induced by a given community.

It should be noted that both the approaches discussed above merely mitigate the scalability problem with any Monte Carlo based algorithm, without actually solve the problem. For this reason, more recent studies have explored different paths, such as the proxy and the sketch based algorithms introduced in the following sections.

2.4.2 Proxy based

The proxy-based approach trades theoretical efficiency for practical efficiency. The heavy MC simulations to estimate the influence are replaced by the definition of some efficient, yet effective, measure to approximate the influence spread. Also, most of the time, a proxy-based algorithm is specifically designed to work with a specific

diffusion model, therefore, as compared with simulation based algorithms, proxy based algorithms have a disadvantage in terms of model generality.

Nonetheless, despite the theoretical disadvantage, empirical evaluations showed that the quality of the solutions provided by proxy-based algorithms are often good enough to compete with those provided by simulation-based approach. This result is even more remarkable if we consider the significant gain in terms of efficiency brought by proxy-based methods.

In the following we discuss two main strategies for the proxy-based algorithms: (i) influence ranking; (ii) diffusion model reduction.

Influence Ranking Proxy. The key idea is to define a ranking function based on a specifically designed *metric* that must be able to capture the nodes' influence. This ranking function is then used to drive the seed set selection process, so that the top k nodes are included into the resulting seed set.

Two of the most basic approaches to influence ranking are the PageRank [159] and the DistanceCentrality [59] algorithms. The former method uses the page rank score of a node to approximate its influence while the latter uses the distance centrality.

Unfortunately, these simple approaches share a common weakness, they tend to overestimate the influence spread of a set because of their inability to account for *i*nfluence overlaps (e.g., two seed nodes that have two overlapping sets of influenced nodes).

Several other ranking methods have been proposed to address the above issue. For instance, DegDis [40], based on the degree proxy – the influence of a node is approximated by its degree – implements a discount mechanism to mitigate the effect of the influence overlaps. More specifically, once a node u is selected as seed, the influence score of any u's neighbor v is *discounted* by a certain factor, e.g., v' score could be subtracted by 1 to account for the influence overlap with the selected seed, i.e, u. A major flaw of the above discount mechanism is that it ignores indirect influence paths.

The issue on the influence overlaps has been also addressed in [134] where the authors define the Group-PageRank GPR. The GPR of a set of nodes S is straightforward, as it is simply the cumulative sum of the page rank score of every node in S, discounted by a certain factor. As in DegDis, the discount factor penalizes the selection of nodes with influence overlaps. Based on the above definition, the authors define the GroupPR algorithm. It adheres to the classic greedy paradigm. In fact, at each iteration, the node with the largest marginal gain is selected. The marginal gain of the any node $v \in V \setminus S$ is computed accordingly to either one of the following methods: (i) $Linear(S, v)$ - it recomputes the GPR for $\{S \cup \{v\}\}$ in $O(|E\ vert)$; (ii) $Bound(S, v)$ - it uses the GPR of S along with the page rank of each node in $j \in \{S \cup \{v\}\}$ to derive the GPR of $\{S \cup \{v\}\}$ in $O(k)$.

Another interesting method, that generalize the Page-Rank proxy, is the Influence Ranking Influence Estimation (IRIE) algorithm proposed in [96]. The key idea of IRIE is to define a system of n linear equations with n variables, where $n = |V|$. Each linear equation recursively defines the influence of a node $u \in V$ as $r(u) = (1 - AP_S(u))(1 + \alpha \sum_{v \in N_u^{out}} p_{uv} r(v)))$. Intuitively, the influence of a node u comprises the influence to itself (i.e., 1) plus the influence of u's neighbors v scaled by the probability of u to activate v. Moreover, the equations account for a damping factor $\alpha \in (0, 1)$ and the probability of a node to be activated by the current seed set selection, denoted by $AP_S(\cdot)$, which can computed with Monte Carlo simulations.

The final seed set returned by IRIE is obtained by updating and solving the above system of linear equations k consecutive times.

Diffusion Model Reduction Proxy. The key idea is to design a reduction for the diffusion model provided as input, so that we can speed up the computation of $\sigma(\cdot)$. There are two main strategies to pursue the above goal: (i) transforming the stochastic diffusion model into a deterministic model; (ii) restricting the influence computation to small portion of the graph.

Regardless of the strategy adopted to design the reduction, the seeds selection stage always follows a greedy approach.

It should be noted that reductions are model-specific. For this reason, we separately discuss the main algorithms for both the IC and LT models.

IC Model Reduction. The key idea is to reduce the number of paths involved in the influence computation. This intuition is based on the observation that for any pair of nodes u and v, among all the possible influence paths that might lead u to activate v, it is possible to individuate a set of *insignificant* paths, namely those paths with a small probability of success. Clearly, these paths can be ignored (almost) without affecting the quality of the influence computation.

The first attempt is represented by the *Shortest-Path Model* (SPM) and SP1M model proposed in [100]. The ideas is to regard as *significant* influence path only those having the shortest path distance between any pair of nodes. Therefore, let $d(u, v)$ denote the minimum distance between u and v and $d(S, v) = \min_{u \in S} d(u, v)$ denote the minimum distance between any node in S and v. In the SPM model, a seed set S is given with single chance to activate v after $d(v, S)$ steps. The SP1M model slightly generalizes the SPM model, as the seed set S is provided with an additional attempt to activate v, at step $d(S, v)+1$. The intuition behind the SPM (resp. SP1M) model is to prune all the paths with length larger than $d(S, v)$ (resp. $d(S, v) = 1$).

The influence function under both models is monotone and submodular, thus a greedy selection strategies provides a solution with $(1-1/e)$ approximation guarantee. However, the approximation guarantee is related to influence function under the SPM model, not the original IC model. Actually, it is most likely that the solution derived under the SPM model performs poorly under the IC model. This is particularly true if the influence probabilities are neither small or constant.

Another reduction of the IC model, probably the most well-known, is the *maximum influence arborescence* (MIA) model [38]. The main idea underlying this model is to construct an arborescence of the input graph, so that the influence of each node u is locally restricted to a small region of the graph with a tree structure rooted in u. This strategy takes advantage of the fact that influence in trees can be computed efficiently and exactly.

The MIA model consists of two main steps: (i) for any pair of nodes the *maximum influence paths* (MIP) are computed; (ii) all the influence paths with probability below a given threshold θ are ignored. The probability of an influence path is given by the product of the influence probabilities of all the edges included in the path.

For any node $u \in V$ the MIA model computes two arborescence:

- *maximum influence in-arborescence*, $MIIA(v, \theta)$, it contains all the maximum influence paths ending at v, with probability at least θ
- *maximum influence out-arborescence*, $MIOA(v, \theta)$, it contains all the maximum influence paths starting from v, with probability at least θ.

Once the arborescence are derived, the marginal gain of a node can be computed extremely fast. The influence function under the MIA model is also monotone and submodular, thus a greedy selection strategy provides a $(1 - 1/e)$-approximation of the optimal solution, with respect to the MIA influence function.

However, the MIA model has a major flaw. In fact, its performance decays either when the graph is too dense or the influence probabilities are relatively high. Specifically, with denser graphs a substantial fraction of the influence paths is pruned by the threshold θ, this leads to a poor approximation of the influence spread under the IC model. Moreover, in graphs with high influence probabilities, especially if θ is small, few influence paths are pruned, thus the influence computation becomes more expensive.

LT Model Reduction. The first reduction for the LT model is the LDAG model proposed in [35]. The approach is very similar the one adopted for the IC model. In fact, it is based on the observation that influence, under the LT model, can be computed exactly and efficiently in *directed acyclic graphs* (DAGs). Therefore, the influence of any node $v \in V$ is restricted to a particular subgraph, i.e., $LDAG(v, \theta)$, which is constructed by computing the shortest paths containing all the nodes exerting an influence on v at least θ.

Unfortunately, the LDAG construction procedure is both computation and memory intensive and it does not scale well with the size of the input graph.

Another interesting LT reduction method is the Simpath algorithm proposed in [72]. It is based on a simple result: the influence of a set of node, under the LT model, can be computed by enumerating all the simple paths starting from every node in the set. Enumerating the simple paths is a #**P**-hard problem, for this reason the algorithm introduces a pruning threshold θ so that the enumeration is limited only those paths having influence at least θ. The influence of a node $u \in V$, i.e., $\sigma(u)$,is then computed by summing up all the simple paths starting from u having influence at least θ. Analogously, the influence of a set of nodes, i.e., $\sigma(S)$ is given by the sum of the influence of each node $u \in S$, with respect to the subgraph induced by $V \setminus S \cup \{u\}$.

The algorithm adopts two optimization techniques to speed up the greedy selection process. More specifically, in the first iteration, the Simpath algorithm finds a vertex cover of the graph. Then, only the influence of the nodes included in the vertex cover is computed directly, while the influence of the remaining nodes is derived starting from the influence of nodes in the vertex cover. The second optimization consists in finding, at each iteration after the first one, the top-l most promising seed candidates, for which the marginal gain has to be computed exactly.

The main advantage of the Simpath algorithm over LDAG is that it does not require to enumerate every simple path in advance. For this reason, Simpath often provides a higher time and space efficiency.

The proxy-based approaches discussed in this section have an advantage in terms of efficiency over any other simulation based approach. However, this advantage often translates to having to deal with a problem that may not be directly related to IM – it is the case of the influence ranking methods – or having to design a specific approach for any diffusion model – it is the case of the model reduction approach.

2.4.3 Sketch based

A sketch-based approach aims at improving the simulation-based theoretical efficiency, while preserving the same approximation guarantee. The focus is put on the main bottleneck of any simulation-based approach, that is the need for heavy MC simulations to evaluate the influence marginal gain of nodes. In order to avoid rerunning

the MC simulations, all the methods discussed in this section pre-compute a number of *sketches* upon which they evaluate influence.

Based on how these sketches are computed, we can devise two different categories: (i) Forward-Influence Sketch (FI-Sketch); (ii) Reverse-Reachable Sketches (RR-Sketch).

Forward Influence Sketch. The idea of this approach is to construct an instance of the influence propagation graph according to the given diffusion model. For instance, under the IC model, a *sketch* of the input graph $G = \langle V, E \rangle$, denoted by G_i, is obtained by removing each edge $(u, v) \in E$ with probability $1 - p_{uv}$.

Given a sketch G_i and a set S, the influence of S is the number of nodes reachable from S. More generally, if we have as set of θ different sketches $\{G_1, G_2, \ldots, G_\theta\}$ the influence of a set of users, i.e., $\sigma(S)$, is given by the average number of users reached by S on the different θ sketches. That is, $\sigma(S) = \frac{1}{\theta} \sum_{i=1}^{\theta} |R_{G_i}(S)|$, where $R_{G_i}(S)$ denotes the set of nodes reachable by S on the sketch G_i.

Theoretical results show that this approach provides the same approximation bound of the simulation based greedy algorithm, with high probability. This probability depends on the accuracy of the influence estimation, which is affected by the number of sketches. In general, the larger the number of sketches the higher the accuracy of the influence spread estimation will be.

Several algorithms are designed on top of the forward influence framework. NewGreIC proposed in [40] applies the FI-Sketch procedure at the beginning of each iteration of the greedy selection stage to evaluate the nodes' marginal gain.

It should be noted that, even though the asymptotic complexity of generating a sketch under the IC model is the same as running an MC simulation, the main advantage of NewGreIC is that sketches are shared among the $O(|V|)$ influence function evaluation.

An issue with the above approach is the need to generate a new set of sketches at the beginning of every iteration. However, it would be preferable to generate all the required sketches in advance and then to use them for every influence estimation required by the greedy selection process. The StaticGreedy algorithm [43] is based this idea. It first generates θ different sketches, then it greedily selects the seed nodes estimating their marginal gain considering the previously generated influence sketches. StaticGreedy ensures a $(1 - 1/e-)$ approximation of the optimal solution high probability, as it is established by the following lemma.

Lemma 1 (Lemma 1 of [125]). *With probability* $1 - n^{-1}$, *StaticGreedy requires* $\theta = (8 + 2\epsilon)n \cdot \frac{\log n + \log \binom{|V|}{k} + \log 2}{\epsilon^2}$ *sketches to achieve the* $(1 - 1/e - \epsilon)$ *approximation ratio.*

The complexity of the static greedy is $O(k \cdot |V| \cdot \theta \cdot |E|)$. Although it represents an improvement over MC-Greedy it is still prohibitive for large graphs. The main problem is with the influence evaluation, which takes $O(|E|)$ time. For this reasons many studies have tried to overcome this problem. For instance, StaticGreedyDU [43] empirically improves the running time of StaticGreedy with the introduction of a pruning mechanism. Specifically, after a node u is selected as seed in the i-th iteration, all the nodes reachable by u are pruned from every pre-computed sketches. As a consequence, any subsequent iteration has the benefit of estimating the influence on sketches with a smaller size.

To further improve the performance of StaticGreedy, Ohsaka et al. propose PrunedMC [156]. This algorithm combines the pruning technique of StaticGreedyDU with the construction of an index structure on the influence sketches. More specifically, the algorithm

Algorithm 3 $RR-Sketch(\mathcal{G}, k, \theta)$: [189]

Input: The diffusion graph $\mathcal{G} = \langle V, E \rangle$; the size of the returned seed set k; Number of RR Sets θ to be generated.

Output: Seed set S of size k.

1: $S \leftarrow \emptyset$
2: $\mathcal{R} \leftarrow \emptyset$
3: Generate θ random RR sets and insert them into $\mathcal{R}$
4: **for** $i \leftarrow 1$ **to** k **do**
5: Pick the node v_i that covers the most RR sets in $\mathcal{R}$
6: Add v_i into S
7: Remove from $\mathcal{R}$ all the RR sets that are covered by v_i
8: **end for**
9: **return** S

builds a directed acyclic graph (DAG) for each sketch G_i. For each DAG, the algorithm creates an index storing the ancestors and the descendants of the hub node, namely the node with the largest degree. The trick to speed up the influence evaluation is the following: given a node v, if it is the ancestor of a hub for a particular sketch, there is no need to visit the descendants of the hub to get the set of users reached by v. With this expedient the time required by any influence estimation is significantly reduced.

Reverse Reachable Sketch. Any algorithm discussed in the previous section is able to overcome the main problem with the FI-Sketch based approach, which is the excessive time required for the sketches generation.

The Reverse Reachable Sketch (RR-Sketch) approach is able to overcome this bottleneck. It is based on the intuition of Borgs et. al in [20], which are the first to understand that it is not necessary to involve the entire graph during the generation of a sketch.

Following this intuition the authors develop the RR-Sketch framework. The key idea is that the influence of a set S can be computed by first the selecting a number random nodes, and then seeing the fraction of these randomly selected nodes that can be reached by S, under some influence sketch. Central to the entire approach are the concept of *R*everse Reachable Set (RR-Set) and *R*andom Reverse Reachable Set (Random RR-Set), which are formally defined as follow.

Definition 16 (RR-Set and Random RR-Set)**.** *Let G' denote an FI-Sketch constructed on G. The Reverse Reachable set of a node v, denoted by $\mathcal{R}_v$, is the set of nodes that can reach v on the considered sketch G'. A random RR set, $\mathcal{RR}(v)$ is generated on an instance G' sampled from the original graph G, where v is randomly selected from the nodes in V.*

Intuitively, a random RR set generated from v contains all the nodes that are potentially able to influence v. Clarly, highly influential nodes a appear in many random RR sets. Analogously, given a collection of Random RR Sets, if a set S covers the largest fraction of those sets, it is most likely to be the optimal set.

Algorithm 3 describes, at a high level, the behavior of the RR-Sketch approach. First, the algorithm generates θ random RR-Sets, then it uses the standard greedy algorithm for the maximum coverage problem. In fact, the seed set S is essentially the set covering the largest fraction of Random RR Sets.

Crucial for Algorithm 3 is the definition of the number of RR sets to be generated, i.e., θ, as it strikes the balance between accuracy and performance.

Borgs et al., in [20] design the RIS framework to bound the number of RR-Sets to a threshold τ. Such threshold is set on the number of visited edges. Also, if τ is set to $O(\epsilon^{-3} \cdot k \cdot (|V|+|E|) \cdot \log^2 n)$, the RIS framework provides a $(1-1/e-\epsilon)$-approximate solution with probability $1-1/n^{-1}$.

Unfortunately, the proposed framework does not achieve the practical efficiency, as its running time complexity is still prohibitive to be used on large graphs.

The first to achieve the goal of practical efficiency is the TIM algorithm proposed by Tang et al. in [189]. It is based on the RIS framework of Borgs et al., but it has the merit of finding a better bound on the number of RR-Sets required to obtain a solution with the same approximation guarantee as the RIS framework.

In fact, TIM requires $O(^{-2} \cdot n \cdot (\log n + \log \binom{n}{k}))/OPT)$ RR-Sets. However, this bound relies on the value OPT, i.e., the optimal value of the influence function, which is unknown. The algorithm is therefore organized into two different stages: the first stage is concerned with the parameter estimation, i.e., it estimates the value of OPT, thus it established the number of required RR-Sets, while the second stage is concerned with seeds selection. While the second stage follows the same schema described in Algorithm 3, the first stage is based on two different bootstrap techniques, designed to estimate OPT The expected time complexity of TIM is $O(\epsilon^{-2} \cdot (n+m) \cdot \log n)$.

Tang et al. further improve the performance of their TIM algorithm in [188], where they propose the IMM algorithm. It follows the same schema as TIM, but it uses a martingale analysis to provide a better bootstrap estimation of OPT.

One common issue with both IMM and TIM, is their large memory consumption. This is due to the need of storing in the main memory a large number of RR-Sets, especially for small values of ϵ, for the entire duration of the algorithm.

For this reason, many followup studies have tried to optimize the memory demand of the above methods. Among these studies, an interesting approach is the one adopted in the SSA algorithm proposed in in [154]. SSA starts by generating an initial number of RR-Sketches. Then, at each iteration, it doubles the number of sketches and extracts the seeds based on the current generated sketches. It stops when the estimated influence of the seed set extracted in the i-th iteration is close enough to the estimated influence of the seed set extracted in the previous iteration. The authors of SSA claim their algorithm can achieve the $(1-1/e-\epsilon)$ approximation ratio, even though in [87] the authors discover an error in the original analysis of [154], which can be easily fixed without a significant impact on the efficiency of the algorithm.

It should be noted that RR-Sketch based approaches have a significant advantage over the FI-Sketch based approaches in terms of efficiency. In fact, the RR-Sketch methods discussed in this section can be regarded as the state of the art algorithms for IM.

2.5 Chapter notes

The influence maximization problem, proposed for the first time in [97], is extremely challenging. It is an example of **NP**-hard problem, thus it is computationally intractable. The source of its complexity lies with the two computational task that are implicitly embedded into the definition of the problem, i.e., the influence computation and the combinatorial problem of finding the best seed nodes for the propagation.

Nonetheless, under the two most classic diffusion models, i.e., the Independent Cascade and the Linear Threshold model, the monotonicity and submodularity of the influence spread function can be exploited to design an effective greedy approach. In fact, a greedy seeds' selection strategy provides a $(1 - 1/e)$-approximation of the optimal solution [152]. However, the bound is not directly applicable, unless we are able to overcome the burden derived from the influence computation.

A first approach to mitigate the hardness of the influence computation, suggested in [97], consists in relying on Monte Carlo simulations to estimate the influence function, as in Algorithm 2. Unfortunately, relying on MC simulations prevent us to claim the original approximation bound of [152], since we need to account for the multiplicative error introduced on the influence spread computation. As consequence, the above bound needs to be modified to take into account the fact that the influence function is *estimated*, rather than computed exactly. Therefore, Algorithm 2 provides an $(1 - 1/e - \epsilon)$-approximation of the optimal solution, where ϵ depends on the accuracy of the MC estimate.

Even though Algorithm 2 has a polynomial time complexity, it is far from being practical efficient. For this reason, for many years researchers have tried to improve over the classic Monte Carlo based greedy algorithm. In pursuing this goal, a number of different paths have been explored (cf. Section 2.4). Nowadays, the sketch-based algorithms, especially the ones based on the RIS framework [20], are arguably the state-of-the-art for influence maximization.

In the following chapters, we first investigate if and to what extent the most commonly used stochastic diffusion models – based on the ideas discussed in this chapter – are able to capture the surprising complexity of real-world propagation phenomena. As a result of this investigation, we formulate a novel class of diffusion models, based on the Linear Threshold Model (cf. Chapter 3). We then devote our attention to the classic influence maximization problem as introduced in [97]. More specifically, we assess the opportunity of exploiting some graph-mining techniques to effectively detect interesting regions of the diffusion graph, i.e., regions populated with optimal spreaders (cf. Chapter 4). Finally, in order to capture emerging scenarios in viral marketing applications, we revisit the classic IM problem. In particular, we design a novel framework which introduces a combination of targeted aspects and diversity-awareness (cf. Chapter5-6).

Chapter 3

Complex Influence Propagation

To properly capture the complexity of influence propagation phenomena in real-world contexts, such as those related to viral marketing and misinformation spread, information diffusion models should fulfill a number of requirements. These include accounting for several dynamic aspects in the propagation (e.g., latency, time horizon), dealing with multiple cascades of information that might occur competitively, accounting for the contingencies that lead a user to change her/his adoption of one or alternative information items, and leveraging trust/distrust in the users' relationships and its effect of influence on the users' decisions. To the best of our knowledge, no diffusion model unifying all of the above requirements has been developed so far. In this work, we address such a challenge and propose a novel class of diffusion models, inspired by the classic linear threshold model, which are designed to deal with trust-aware, non-competitive as well as competitive time-varying propagation scenarios. Our theoretical inspection of the proposed models unveils important findings on the relations with existing linear threshold models for which properties are known about whether monotonicity and submodularity hold for the corresponding activation function. We also propose strategies for the selection of the initial spreaders of the propagation process, for both non-competitive and competitive influence propagation tasks, whose goal is to mimic contexts of misinformation spread. Our extensive experimental evaluation, which was conducted on publicly available networks and included comparison with competing methods, provides evidence on the meaningfulness and uniqueness of our models.

3.1 Introduction

Since the early applications in viral marketing, the development of information diffusion models and their embedding in optimization methods has provided effective support to address a variety of influence propagation problems.

However, due to the shrinking boundary between real and online/virtual social life [14] along with the unlimited *misinformation* spots over the Web, e.g., *fake news* [99, 105], deciding whether a source of information is reliable or not has become a delicate task. For these reasons, understanding the complex dynamics of information diffusion phenomena has emerged as a task of paramount importance, since the way people act on the Web reflects how people behave in reality, which eventually depends to some extent on the way everyone consumes and acquires information.

A few studies on the spreading of fake news and hoaxes [149, 151] argued that, the likelihood of people to be deceived by a spreading information item is increased because assessing the reliability and trustworthiness of the source generating and/or sharing such item becomes harder. Within this view, one side effect is the tendency of users to access information from like-minded sources [102] and at the same time, to

be trapped inside information bubbles, thus favoring network polarization phenomena [64].

Remarkably, the research community is still divided on this subject. In fact, few studies argue that, due to the inherent diversity of the entire media environment (e.g., newspapers, television, social networks), eco chambers are often overstated, especially for political matters [54]. Nonetheless, the analysis carried out in [54] does not necessarily neglect the presence of information bubbles, and it also recognizes the need of avoiding their formation.

To this purpose, debunking and fact-checking are two important tools. One can devise two main strategies when it comes to debunk information: *real-time detection and correction*, or *delayed correction* [104]. However, in both cases, the response time plays a crucial role into the effectiveness of the correction attempt, because users tend to reinforce their own belief — a cognitive phenomenon known as *confirmation bias*. Moreover, there is no guarantee about the effectiveness of such correction, on the contrary, highlighting a fake news may even produce a *backfire* effect, i.e., driving users' attention towards the misleading piece of information.

In this scenario, it appears that one recipe to deal with the interleaving of information and dis/misinformation should be to educate people to be mindful of the informative source. Unfortunately, it is often difficult to understand where an information item originated from. Therefore, it turns out to be essential to capture the effects that different types of social ties, particularly *trust/distrust relationships*, can have on both the user behavior and propagation dynamics. Two related questions hence arise:

Q1 *What are the key-features that make a diffusion model able to explain the inherent dynamic, and often competitive, nature of real-world propagation phenomena?*

Q2 *Do the currently used models of diffusion already incorporate such features?*

To address question **Q1**, we recognize a number of aspects as essential constituents of a "realistic" information diffusion model, namely: (1) leveraging trust/distrust information in the user relationships to capture different effects of influence on decisions taken by a user; (2) accounting for a user's change in adopting one or alternative information items (i.e., relaxation of the diffusion progressivity assumption); (3) accounting for a user's hesitation or inclination towards the adoption of an information over time; (4) accounting for time-dependent variables, such as latency, to explain the propagation dynamics; (5) dealing with multiple cascades of information that might occur competitively.

Motivating example. Our above hypothesis is supported by the following example: consider a typical scenario occurring in a political campaign, where two candidates want to target the audience of potential electors. Let's assume, at the start of the political campaign, every elector has a complete unbiased opinion towards one of the two candidates. The ultimate decision about which candidate to vote it will likely be affected by both "exogenous" and "endogenous" influencing factors, i.e., one may be genuinely influenced by decisions taken by her/his social contacts — impact of homophily factors — but s/he may also have formed her/his own opinion outside the network of friends. In fact, not only friends, but also the network of *foes* has some degree of influence over the decision process of an individual. As a consequence of such negative influence received by foes, one may become more hesitant in taking a decision, which would be reflected by a *quiescence* state of the elector before being fully engaged in the promotion of the chosen candidate. Moreover, despite an elector may alternate her/his opinion in favor of one or the other candidate before the final endorsement,

it will be more difficult to induce this change over time. In this regard, a time-aware notion of *activation threshold* is needed to mimic the effects of the *confirmation bias*. Finally, all decisions must be taken before the time limit, i.e., the election day, which constrains the political campaign period.

Question **Q2** has been addressed by a relatively large corpus of research studies in the last few years. A variety of methods, mainly built upon classic information diffusion models such as Independent Cascade (IC) and Linear Threshold (LT) [97], have tried to explain realistic propagation phenomena in order to solve optimization problems related to influence propagation. As we shall discuss in Section 2, diffusion models have been developed to incorporate one or more of the following aspects: multiple, competitive cascades of information; time horizon for the unfolding of the diffusion process; time-dependent influence; delay in the propagation; and trustworthiness of the influence relations. However, to the best of our knowledge, *all* of the above aspects have never been unified into the same (LT-based) diffusion model.

Contributions. In this chapter, we propose a novel class of diffusion models, named *Friend-Foe Dynamic Linear Threshold Models* (F^2DLT). They are based on the classic LT model and are designed to deal with *non-competitive* as well as *competitive* time-varying propagation scenarios. In our proposed models, the information diffusion graph is defined on top of a *trust network*, so that the strength of trust and distrust relationships is encoded into the influence probabilities. The response of a user to the influencing attempts is described by the means of a time-varying activation function, depending on both the inherent activation-threshold of the user and her/his tendency of keeping or leaving the campaign-specific activation state over time. We also introduce a quiescence function to model the latency or delay in the propagation, which accounts for the involvement of the user's foes in the information diffusion. Remarkably, in our models, the trusted connections and distrusted connections play different roles: only friends can exert a degree of influence for activation/contagion purposes, whereas foes can only contribute to increase the user's hesitation to commit with the propagation process. For competitive scenarios, we define two models with clearly different semantics: a *semi-progressive* model, which assumes that a user, once activated, is only allowed to switch to a different campaign, and a *non-progressive* model, which instead requires a user to have always the support of her/his in-neighbors to keep the activation state with a certain campaign.

We provided several theoretical insights into the proposed models. In particular, we demonstrated how each of our models could be reduced to other LT-based models for which properties are known about whether monotonicity and submodularity hold for the corresponding activation function.

Another contribution of this work is the definition of four *seed selection* strategies, which mimic different, realistic scenarios of influence propagation. These strategies are central to our methodology of propagation simulation, since the development of optimization methods under our diffusion models is beyond the goals of this work. Notably, in competitive scenarios, we have focused on combinations of strategies (to associate with competing campaigns) that might be reasonably considered for a misinformation spread limitation problem.

Experimental evaluation conducted on four real-world networks, also including comparison with stochastic epidemic models and the dynamic linear-threshold (DLT) model, has provided interesting findings on the meaningfulness and uniqueness of our proposed models.

TABLE 3.1: Summary of related work based on optimization problem, basic diffusion model (**DM**), competitive diffusion (**C**), non-progressivity (**NP**), time-aware activation (**TA**), delayed propagation (**DP**), trust/distrust relations (**TD**).

Ref.	Problem	DM	C	NP	TA	DP	TD
[24]	rumor blocking	IC	✓				
[191]	rumor blocking	IC	✓				
[82]	rumor blocking	LT	✓				
[57]	rumor blocking	distrib.	✓				
[36]	positive influ. max.	IC	✓				
[135]	active time max.	IC		✓			
[58]	PTS min.	LT		✓			
[31]	positive influ. max.	Voter	✓				✓
[183]	positive influ. max.	LT	✓				✓
[201]	positive influ. max.	LT	✓				✓
[131]	time-constrain. influ. max.	IC				✓	
[34]	time-constrain. influ. max.	IC				✓	
[150]	positive influ. max.	IC	✓	✓		✓	✓
[130]	rumor blocking	LT	✓	✓	✓		
[137]	positive influ. max.	IC	✓				

3.2 Related work

We overview information diffusion models that, in the attempt of explaining realistic propagation phenomena, incorporate one or more of the following aspects: multiple, competitive cascades of information, time horizon for the unfolding of the diffusion process, time-dependent influence, delay in the propagation, trustworthiness of the influence relations. Table 3.1 provides a guide to our discussion.

Please note that here we refer to the vast literature on probabilistic models originally designed to explain stochastic processes of information diffusion, which include the classic *Independent Cascade* (IC) and *Linear Threshold* (LT) models [33], and relating optimization problems, such as influence maximization. By contrast, we will leave out of consideration deterministic models, such as the structural cascades specifically designed to model context/content-sensitive diffusion over an interaction network (e.g., [49, 103]).

Also, it is worth noting that the information diffusion modeling problem we tackle in this work is significantly different from the one addressed by *epidemic models*, such as SIS, SIR(S), and SEIR(S) [83], already for the non-competitive scenario. Standard epidemic models are originally defined as compartmental models, since the individuals of a population are divided in compartments that describe an epidemiological state. The parameters used to represent transition rates for changing states are absolute constants, which means that the infection process in compartmental models has a deterministic behavior. Also, standard epidemic models are of mass-action type, since individuals are represented as normalized fraction of a population which randomly interact with each other. As discussed in [51], even social contagion based on stochastic or generalized epidemic models (i.e., there is a probability distribution of rates to govern the infection process) is originally defined on random networks, and its revision to deal with social networks would lead to more complicated models. In this regard, one direction is taken by the *stochastic individual-contact, network models*, whereby SIS and related models are reformulated by considering a stochastic infection process and a network-based population of individually identifiable elements. In Section 3.5.3.1,

we present a stage of experimental evaluation devoted to a comparison with such models. However, even if epidemic models have also been used for social influence, they are not the most common approach to such topic [163]. This is manly due to the fact that identifying and modeling the causal mechanisms of the spread of ideas is more difficult than for the spread of diseases. By contrast, the threshold models for influence propagation (even the simplest ones) have two important features that are not clearly present in epidemic models. First, individuals have different behaviors, being such differences reflected in the distribution of activation thresholds associated with the individuals; by contrast, in stochastic epidemic models, the state-transition probabilities are drawn independently of the individuals' relations. Second, an individual's behavior also depends on the behavior of other individuals s/he is linked to: here, it is helpful to think about threshold models as an example of complex contagions, whereby an individual takes an action as a result of the exposure to multiple sources of influence; by contrast, epidemic models are more likely to represent simple contagion, in that a single source of influence (as social contact) may suffice to cause an individual's action. Moreover, while the transition to the recovered state assumes non-progressivity in stochastic diffusion models such as LT or IC, such a transition in SIR(S) is defined to happen spontaneously, discarding any influence that may result from the interaction with other individuals. For all such reasons, thresholds models are usually considered more appropriate in contexts like the adoption of new technologies or controversial ideas [33]. And, in our work, we indeed follow this line of research. One further point of divergence adheres to the notion of competitiveness that is somehow found in advanced epidemic models: this refers to the presence of two or multiple groups of individuals (with some distinguishing characteristics) which are however affected by the same, single disease [80, 85, 94], therefore it corresponds to a totally different notion than what is addressed in our work.

In the following, we briefly recall the definition of the LT model, which is at the basis of our proposal; then, we focus on related work that address the aforementioned aspects concerning complex propagation phenomena.

The classic Linear Threshold model. Given a directed graph representing a social network, with estimates of influence probabilities provided as edge weights, nodes can be "activated" (i.e., influenced) through an information cascade starting from an initially selected set of seed nodes (i.e., early-adopters). At the beginning of the information diffusion process, each node is assigned a threshold uniformly at random from $[0, 1]$. The diffusion process unfolds in discrete time steps and follows certain rules: nodes are either active or inactive; once activated, nodes cannot deactivate; an active node may trigger activation of neighboring nodes; a node can be activated at time $t + 1$ by its active neighbors if their total influence weight at time t exceeds the threshold associated to that node. The process runs until no more activations are possible.

Competitive diffusion. A number of studies have been devoted to model competitive diffusion; see, e.g., [33] for a general introduction to IC and LT classes of competitive diffusion models. Focusing on competitive diffusion and related optimization problems under the context of misinformation spread limitation, one of the earliest work is [24], which proposes a multi-campaign IC model to address the influence limitation problem, i.e., to find a seed set of size k for one, "good" campaign such that the number of nodes influenced by the other, "bad" campaign is minimized. In [191], the problem of rumor blocking is addressed under the competitive IC model and a randomized algorithm is developed for the selection of the seed set able to yield the maximum reduction in the number of bad-infected nodes. An influence blocking maximization problem is also addressed in [82], using competitive LT. In [36], the

two competing cascades correspond to opposite opinions, where the negative one may emerge spontaneously from any user in the network, e.g., a user got disappointed with a purchased item and decides to spread negative opinion among her/his contacts. Lu et al. [137] also address the aspect of complementarity between two competing campaigns, under the assumption that if the two information items are correlated then the adoption of one item might favor further adoption of the second item over time.

Non-progressive diffusion. While modeling the competitive nature of information cascades, the above works however refer to progressive models. On the contrary, a few studies have been proposed to model non-progressive diffusion. For instance, [135] introduces a deactivation function into a continuous non-progressive model, whereas an extension of LT is proposed in [58] to define a non-progressive strict majority model. However, both models are also non-competitive.

Social ties and temporal aspects. All of the aforementioned works discard two important aspects: (i) the nature of social ties and their impact on the influence propagation, and (ii) time aspects concerning the diffusion process. The dichotomy between opposite types of social ties (e.g., friend vs. foe relations) has been widely studied in OSN analysis (e.g., [113]), however its incorporation into diffusion models has been relatively little explored so far. For instance, two extensions of competitive LT with negative relations are defined in [183], to support positive opinion maximization, and in [201], to model the adoption of opinions from friends or opposite opinions from foes. All of such models are competitive but do not consider temporal aspects in the activation or propagation processes.

Several works have studied different types of temporal variables and their impact on spreading processes (e.g., [93, 129, 194]), mainly focusing on lags and delays due to the diverse response-time and heterogeneous susceptibility of users. In [131], the authors propose a latency-aware IC model inspired by [93], in which an influencing delay is introduced in the activation function. Under this model, a time-constrained IM problem is defined, i.e., to find a seed set of size k such that the expected number of nodes is activated before a given time limit. Another extension of IC is proposed in [34], where a notion of meeting probability is introduced to control the activation of neighbors. The models in [34, 131] are non-competitive. By contrast, the trust-based latency-aware IC model proposed in [150] features competitiveness, non-progressivity, temporal delay in propagation, and is also designed to deal with trust/distrust relations.

Dynamic behaviors. All of the previously mentioned works still lack aspects modeling the dynamic behavior of the users. In particular, according to recent studies about polarization of opinion in OSNs [2] and related works about misinformation reduction [104, 119], a crucial aspect is to intervene before a competing campaign can reach the users, or at least soon enough, so that a user does not have time to radicalize her/his thoughts. This idea was first captured in [130], where a dynamic LT model (DLT) is defined to deal with competitive information cascades. The influence weights temporally decay according to a Poisson distribution, and every node can be either positively or negatively activated at a given time depending on the absolute value of the cumulative influence of its neighbors, while the activation sign depends on the sign of the cumulative influence. Moreover, a dynamic behavior aspect lays on the update of the activation threshold whenever a user switches her/his belief.

The latter work shares with our proposal all features of competitiveness, non-progressivity (although deactivation is not allowed), time-aware propagation, dynamic influence behavior, and incorporation of opposite opinions in the influence probabilities; moreover, it is also based on LT. However, our competitive models differ from DLT in [130] since (i) we explicitly model trust and distrust relationships to define the

influence probabilities, (ii) our activation function takes into account only the trusted connections while (iii) distinguishing between the two information cascades; (iv) we introduce a quiescence function to model a delay in the information propagation depending on the strength of influence exerted by distrust relations (i.e., foe neighbors); finally, (v) the activation threshold in our models becomes stronger over time as a node is holding a particular belief. In Section 3.5.3.2 we shall compare our models with DLT.

3.3 Friend-Foe Dynamic Linear Threshold Models

In this section we describe our proposed class of Friend-Foe Dynamic Linear Threshold (F^2DLT) models, which is comprised of: the Non-Competitive F^2DLT (nC-F^2DLT), the Semi-Progressive Competitive F^2DLT (spC-F^2DLT), and the Non-Progressive Competitive F^2DLT (npC-F^2DLT). We first provide an overview of the framework based on F^2DLT. Next, we introduce key features common to all models, then we elaborate on each of them.

3.3.1 Overview

Figure 3.1 illustrates the conceptual architecture of a framework for information diffusion and influence propagation based on our proposed models. Given a population of OSN users, the framework requires three main inputs: (i) a *trust network*, which is inferred from the social network of those users to model their trust/distrust relationships; (ii) user behavioral characteristics that are intrinsic to each user (i.e., exogenous to an information diffusion scenario) and oriented to express two aspects: *activation-threshold*, i.e., the effort needed to activate a user through cumulative influence from her/his neighbors, and *quiescence*, i.e., the user's hesitation in being actively committed with the propagation process; and, (iii) one or multiple competing *campaigns*, i.e., information cascades generated from the agent(s) having viral marketing purposes. Moreover, the information diffusion process has a *time horizon*, and its temporal unfolding is reflected in the evolution of the information diffusion graph: this also depends on the dynamics of the users' behaviors in response to the influence chains started by the campaign(s), which admit that users may switch from the adoption of a campaign's item to that of another one. Putting it all together, our F^2DLT based framework embeds all previously discussed aspects that are required to explain complex propagation phenomena, i.e., competitive diffusion, non-progressivity, time-aware activation, delayed propagation, and trust/distrust relations.

Please note that inferring a trust network from a social network is *not an objective* of this work; rather, we assume that trust relationships between users of an OSN are available and, as we shall describe next in this section, they are exploited as key information to develop our proposed models. Several heuristics have been proposed to infer a trust network from social relations and interactions among users in an OSN. A common approach is to infer trust relationships based on the social influence exerted by users over the network and propagation of trust ratings [67, 79, 95, 157]. Other studies utilize users' activities in social media [66], or users' attributes and interactions [132], or combinations of aspects concerning user affinity, familiarity and reputation [205], social influence, social cohesion and the effective valence expressed by the users in the textual contents they produce [195]. The interested reader may refer to [174, 186] for an exhaustive overview on the topic.

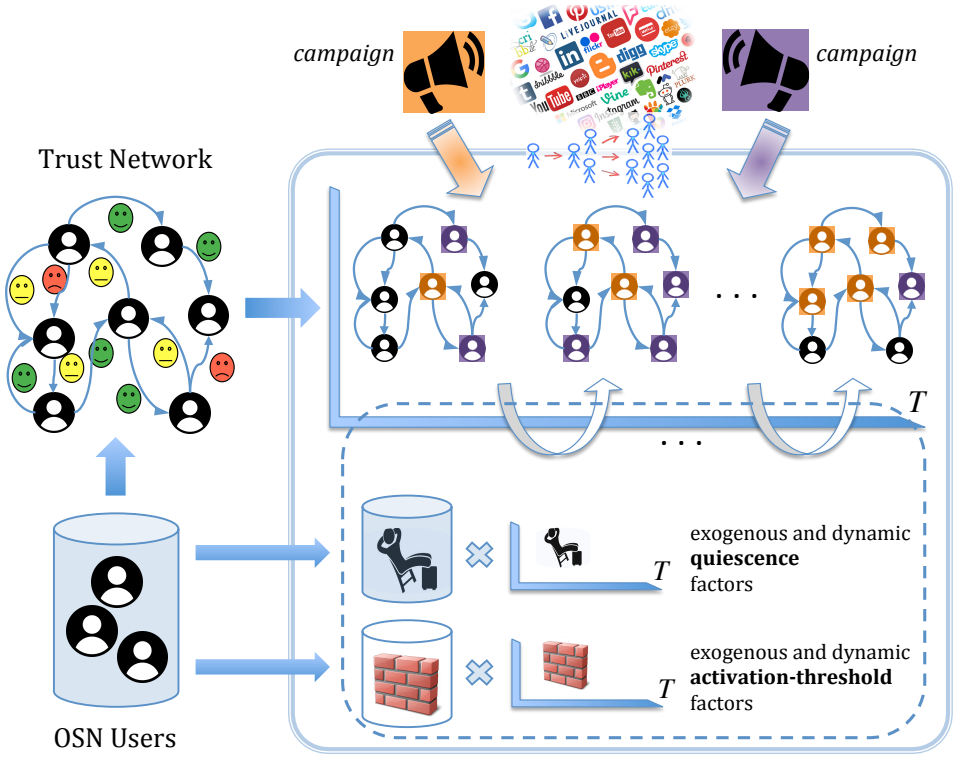


FIGURE 3.1: Illustration of the information diffusion framework based on our proposed F^2DLT

3.3.2 Basic definitions

We are given a *trust network* represented by a directed graph $G = \langle V, E, w \rangle$, with set of nodes V, set of edges E, and weighting function $w : E \mapsto [-1, 1]$ such that, for every edge $(u, v) \in E$, $w_{uv} := w(u, v)$ expresses how much v trusts its in-neighbor u. Positive, resp. negative, value of w_{uv} corresponds to a *trust*, resp. *distrust*, relation.

For every $v \in V$, we denote with $N^{in}_{+}(v)$ and $N^{in}_{-}(v)$ the set of neighbors trusted by v (i.e., *friends* of v) and the set of neighbors distrusted by v (i.e., *foes* of v), respectively. Moreover, as required in linear threshold models, the constraints $\sum_{u \in N^{in}_{+}(v)} w_{uv} \leq 1$ and $\sum_{u \in N^{in}_{-}(v)} |w_{uv}| \leq 1$ must be fulfilled.

Let $G = G(g, q, T) = \langle V, E, w, g, q, T \rangle$ be a directed weighted graph representing the LT-based *information diffusion* graph associated with trust network G, where T denotes a *time interval* for the diffusion process, g and q denote time-dependent *activation-threshold* and *quiescence* functions. These are introduced in G to model the aspects of *time-aware activation* and *delayed propagation*, respectively. We use symbol S_t to denote the *set of active nodes* at time t, and symbol $\widetilde{S}_t$ to denote the set of active nodes for which, at t, the quiescence time is not expired yet, i.e., the *quiescent nodes*.

Activation-threshold function. According to the LT model, every node $v \in V$ is associated with an exogenous activation-threshold, $\theta_v \in (0, 1]$, which corresponds to the a-priori effort needed in terms of cumulative influence to activate the node. We enhance this concept by defining an *activation-threshold* function, $g : V, T \mapsto \mathbb{R}^+$, such that for every $v \in V$ and $t \in T$:

$$g(v, t) = \theta_v + \vartheta(\theta_v, t),$$

i.e., the activation of v at time t depends both on the user's pre-assigned threshold, θ_v, and on a time-evolving activation term, $\vartheta(\cdot,\cdot)$, which models the dynamic response of a user towards the activation attempt exerted by her/his neighbors.

To specify $\vartheta(\cdot,\cdot)$, we devise two main scenarios for $g(\cdot,\cdot)$:

- A *biased* scenario, modeled as a non-decreasing monotone function, to capture the tendency of a user to consolidate her/his belief, according to the *confirmation-bias* principle [2].
- An *unbiased* scenario, modeled as non-monotone function, whereby we assume that a user could revise her/his uncertainty to activate over time, thus becoming more or less inclined to change her/his opinion on an information item. This is particularly meaningful in applications such as customer retention, or churn prediction (i.e., a decrease in the activation-threshold would correspond to the tendency of a user to churn in favor of another service).

Both variants $\vartheta(\cdot,\cdot)$ range within the interval $[0,1]$, for any $v \in V$.

Let us first consider the biased scenario, which is focused on the confirmation bias principle. We choose the following form for the activation-threshold function, by which the value increases by increasing the time a node keeps staying in the same active state:

$$g(v,t) = \theta_v + \vartheta(\theta_v,t) = \theta_v + \delta \times \min\left\{\frac{1-\theta_v}{\delta}, t - t_v^{last}\right\}, \tag{3.1}$$

where t_v^{last} denotes the last (i.e., most recent) time v was activated and $\delta \geq 0$[1] represents the increment in the value of $g(v,t)$ for consecutive time steps. Thus, the longer a node has kept its active state for the same information cascade (*campaign*), the higher its activation value, and as a consequence, it will be harder to make the node change its state, or even no more possible (i.e., $g(v,t)$ saturates to 1, as the difference $(t - t_v^{last})$ exceeds $(1-\theta_v)/\delta$).

In the unbiased scenario, we define the activation-threshold function such that, for each v, the value of the function is maximum (i.e., 1) just after the activation, i.e., at time $t = t_v^{last}+1$, then for subsequent time steps, the function exponentially decreases towards θ_v:

$$g(v,t) = \theta_v + \vartheta(\theta_v,t) = \theta_v + \exp(-\delta(t - t_v^{last} - 1)) - \theta_v \mathbb{I}[t - t_v^{last} = 1], \tag{3.2}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function, i.e., it equals 1 if $t - t_v^{last} = 1$, 0 otherwise. Note that δ is used differently w.r.t. the previous scenario, as it acts as a coefficient that controls the decrease of the activation-threshold function over time.

Quiescence function. Each node in G is also associated with a *quiescence* value, which quantifies the latency in propagation through that node. We define a *quiescence* function, $q : V, T \mapsto T$, non-decreasing and monotone, such that for every $v \in V, t \in T$, with v activated at time t:

$$q(v,t) = \tau_v + \psi(N_-^{in}(v), t),$$

where $\tau_v \in T$ represents an exogenous term modeling the user's hesitation in being fully committed with the propagation process, and $\psi(N_-^{in}(v), t)$ provides an additional

[1] We assume the second additive term in Equation (3.1) is zero if $\delta = 0$.

delay proportional to the amount of v's neighbors that are distrusted and active, by the time the activation attempt is performed by the v's trusted neighbors:

$$q(v,t) = \tau_v + \psi(N_-^{in}(v), t) = \tau_v + \exp\left(\lambda \times \sum_{u \in S_{t-1}} |w_{uv}|\right), \tag{3.3}$$

where $\lambda \geq 0$ is a coefficient modeling the average user sensitivity in the perceived negative influence. Intuitively, this coefficient would weight more the negative influence as the diffusing informative item is more "worth of suspicion". Note also that, in Equation 3.3, w_{uv} is a negative value, since u is a distrusted neighbor of v, i.e., $u \in N_-^{in}(v)$.

Rationale for activation and propagation. Our choice of using, on the one hand, friends for the activation of a user, and on the other hand, foes to impact on delayed propagation, represents a key distinction from related work [130, 183, 201]. Therefore, in our models, the trusted connections and distrusted connections play different roles: only friends can exert a degree of (positive) influence, whereas foes can only contribute to increase the user's hesitation to commit with the propagation process.

It should be noted that both activation and delayed propagation terms also include exogenous factors. We indeed take into consideration both the existence of environmental and personal factors of influence on an individual's behavior. Several studies in information diffusion and influence maximization have reported evidences that, apart from influence coming from social contacts, an individual may be affected by some external event(s) and/or personal reasons to adopt an information [73] as well as to delay the adoption of an information [90].

In our setting, we tend to reject as true in general, the principle "I agree with my friends' idea and disagree with my foes' idea" (which is also close to the adage "the enemy of my enemy is my friend"), since this would imply that the behavior of a user should be completely determined by the stimuli coming from her/his neighbors. Rather, according to most conceptual models developed in social science and human-computer interaction fields (see, e.g., [15, 190]), we believe that the individual's influenceability has a component based on personal characteristics.

3.3.3 Non-competitive model

We introduce the first of the three proposed models, which refers to a single-item propagation scenario. Figure 3.2 shows the life-cycle of a node in the diffusion graph under this model.

Definition 1. ***Non-Competitive Friend-Foe Dynamic Linear Threshold Model (nC-F^2DLT).*** *Let $G = \langle V, E, w, g, q, T\rangle$ be the diffusion graph of Non-Competitive Friend-Foe Dynamic Linear Threshold Model (nC-F^2DLT). The diffusion process under the nC-F^2DLT model unfolds in discrete time steps. At time $t = 0$, an initial set of nodes S_0 is activated. At time $t \geq 1$, the following rule applies: for any inactive node $v \in V \setminus (S_{t-1} \cup \widetilde{S}_{t-1})$, if $\sum_{u \in N_+^{in}(v) \cap S_{t-1}} w_{uv} \geq g(v,t)$, then v will be added to the set of quiescent nodes $\widetilde{S}_t$, with quiescence time equal to $t^* = q(v,t)$. Once the quiescence time is expired, v will be removed from $\widetilde{S}_t$ and added to the set of active nodes S_{t^*}. The process continues until T is expired or no more activation attempts can be performed.* □

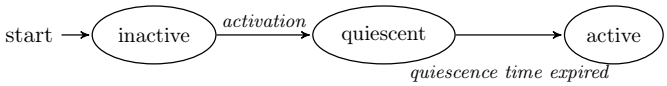


FIGURE 3.2: Life-cycle of a node in the nC-F^2DLT model.

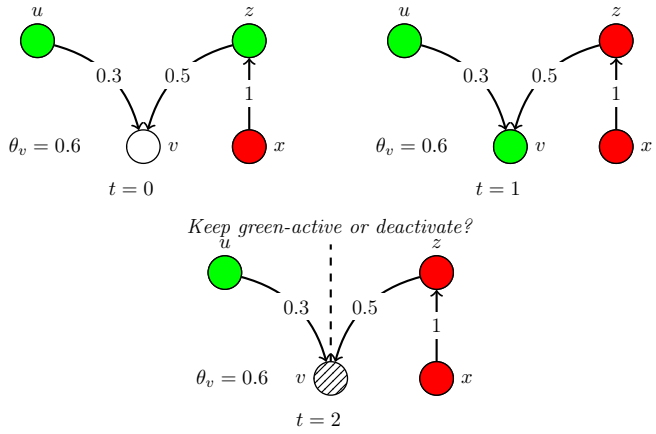


FIGURE 3.3: Uncertainty in an example two-campaign activation sequence.

3.3.4 Competitive models

Here we introduce the two competitive F^2DLT models. Let us first provide our motivation for developing two different competitive models: through the following example, we illustrate a particular situation that may occur when dealing with two campaigns competitively propagating through a network. Please note that, throughout the rest of this chapter, we will consider only two competing campaigns for the sake of simplicity; nevertheless, *our proposed models are generalizable to more than two competing campaigns.*

Example 1. *Figure 3.3 shows an example activation sequence in a competitive scenario between two information cascades, distinguished by colors red and green. At time $t = 0$, nodes u and z are green-active, and their joint influence causes green-activation of node v as well (since $0.3 + 0.5 \geq 0.6$). At time $t = 1$, as fully influenced by node x, node z has switched its activation in favor of the red campaign. After this switch, at time $t = 2$, it happens that v's activation state is no more consistent with the (joint or individual) influenced exerted by u and z. In particular, two mutually exclusive events might in principle happen at $t = 2$: either v is deactivated or v maintains its green-activation state.* ■

The uncertainty situation depicted in the above example prompted us to the definition of two models, namely *semi-progressive* and *non-progressive* F^2DLT: the former corresponds to the case of v keeping its current (i.e., green) activation state, whereas the latter corresponds to v returning to the inactive state. Clearly, the two models' semantics are different from each other: the semi-progressive model assumes that a user, once activated, cannot step aside, unlike the non-progressive one, which instead requires a user to have always the support of her/his in-neighbors to keep activation.

Given two information cascades, or *campaigns* C', C'', for every time step $t \in T$ we will use symbols S'_t and S''_t to denote the sets of active nodes, such that that $S'_t \cap S''_t = \emptyset$, and analogously symbols $\widetilde{S'}_t$ and $\widetilde{S''}_t$ as the sets of quiescent nodes, for C' and C'', respectively. Also, $S_t = S'_t \cup S''_t$ and $\widetilde{S}_t = \widetilde{S'}_t \cup \widetilde{S''}_t$.

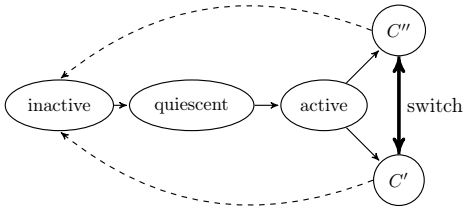


FIGURE 3.4: Life-cycle of a node in competitive models. Straight lines represent the
transitions common to both *spC-F²DLT* and *npC-F²DLT*, while dashed lines
refer to *npC-F²DLT* only.

It should also be noted that, while sharing the time interval (T) of diffusion, C' and C'' are not constrained to start at the same time t_0. Nevertheless, for the sake of simplicity, we hereinafter assume that $t_0 = t'_0 = t''_0$ (with $t_0 \in T$), unless otherwise specified (cf. Section 3.5).

Definition 2. ***Semi-Progressive Competitive Friend-Foe Dynamic Linear Threshold Model*** **(*spC-F²DLT*).** *Let $G = \langle V, E, w, g, q, T\rangle$ be the diffusion graph of Semi-Progressive Competitive Friend-Foe Dynamic Linear Threshold Model (spC-F²DLT), and C', C'' be two campaigns on G. The diffusion process under the spC-F²DLT model unfolds in discrete time steps. At time $t = 0$, two initial sets of nodes, S'_0 and S''_0, are activated for each campaign. At every time step $t \geq 1$, the following state-transition rules apply:*

R1. *For any inactive node $v \in V \setminus (S_{t-1} \cup \widetilde{S}_{t-1})$, if $\sum_{N^{in}_+(v)\cap S'_{t-1}} w_{uv} \geq g(v,t)$, then v will be added to $\widetilde{S'}_t$; analogously, if $\sum_{N^{in}_+(v)\cap S''_{t-1}} w_{uv} \geq g(v,t)$, then v will be added to $\widetilde{S''}_t$. If both conditions hold, i.e., v can be simultaneously activated by both campaigns, a* tie-breaking *rule will apply, in order to decide which campaign actually determines the node's transition in the quiescent state.*

R2. *When a node v enters the quiescent state corresponding to C' (resp. C'') for the first time, it will stay in the quiescent node-set $\widetilde{S'}_t$ (resp. $\widetilde{S''}_t$) until the quiescence time is expired. After that, v will be moved to S'_t (resp. S''_t), i.e., it will become active for C' (resp. C'').*

R3. *Given a node v active for C'', i.e., $v \in S''_{t-1}$, if $\sum_{N^{in}_+(v)\cap S'_{t-1}} w_{uv} \geq g(v,t)$ and $\sum_{N^{in}_+(v)\cap S'_{t-1}} w_{uv} > \sum_{N^{in}_+(v)\cap S''_{t-1}} w_{uv}$, then v will be removed from S''_t and added to S'_t; analogous rule holds for any node active for the first campaign.*

Every node for which none of the above transition-state rules is triggered at time t, it will keep its current state at time $t+1$.

□

The life-cycle of a node in *spC-F²DLT* is shown in Figure 3.4. Note that, once a node becomes active, it cannot turn back to the inactive state, but it can only change the activation campaign. Moreover, switch transitions occur instantly.

Definition 3. ***Non-Progressive Competitive Friend-Foe Dynamic Linear Threshold Model*** **(*npC-F²DLT*).** *Let $G = \langle V, E, w, g, q, T\rangle$ be the diffusion graph of Non-Progressive Competitive Friend-Foe Dynamic Linear Threshold Model (npC-F²DLT), and C', C'' be two campaigns on G. The diffusion process in npC-F²DLT evolves according to the same rules as in spC-F²DLT plus the following rule concerning the deactivation process of an active node:*

R4. *For any active node v at time $t-1$, if $\sum_{N_+^{in}(v)\cap S'_{t-1}} w_{uv} < \theta_v$ and $\sum_{N_+^{in}(v)\cap S''_{t-1}} w_{uv} < \theta_v$, then v will turn back to the inactive state at time t.*

Every node for which none of the transition-state rules is triggered at time t (including the ones defined for spC-F^2DLT), it will keep its current state at time $t+1$.

□

It should be noted that a node's deactivation rule depends on θ_v only (rather than on the whole function $g(v,t)$); otherwise, every node activated at a given time could deactivate itself in the next time step, due to the increase in its activation threshold. This would eventually lead to a configuration in which all nodes in the network, except the initially activated ones, are in the inactive state. The life-cycle of a node in the *npC-F^2DLT* is illustrated in Figure 3.4. Note that, unlike in *spC-F^2DLT*, transitions to inactive state are allowed.

3.3.5 Theoretical properties of the models

In this section we provide insights into the proposed models. Our main goal is to understand how the features introduced in each of our LT-based models impact on the models' spread behavior, particularly on monotonicity and submodularity properties. We organize our analysis into two parts: the first corresponding to *non-competitive* diffusion, and the second to *competitive* diffusion.

3.3.5.1 Non-competitive diffusion

We show that *nC-F^2DLT* can be reduced to *LT with quiescence time*, hereinafter denoted as *LTqt*. By proving the equivalence between the two models, we hence claim that both the monotonicity and submodularity properties hold for *nC-F^2DLT*. Note that since we deal with a progressive model, we assume without loss of generality that, for every node v, the activation-threshold function has a constant value for the whole duration of the diffusion process, i.e., $g(v,t) = \theta_v$.

Definition 4. ***Reduction of** nC-F^2DLT **to LTqt.*** *Given $G = \langle V, E, W, g, q, T\rangle$ for nC-F^2DLT, a diffusion graph $G_{LT} = \langle V_{LT}, E_{LT}\rangle$ can be derived, under LTqt, such that $V_{LT} = V$ and $E_{LT} = \{(u,v)|(u,v) \in E, w_{uv} > 0\}$. Every node $v \in V_{LT}$ is assigned a quiescence time equal to the maximum value of the quiescence function $q_v(\cdot)$, i.e., $\tau_v^{max} = \tau_v + \psi(N_-^{in}(v))$.* □

Definition 4 exploits the fact that the distrust connections are not involved in the activation process, but only in the calculation of the quiescence time. Therefore, we can assume this time to be the maximum possible value, and hence we can study the propagation under *LTqt*. The reduction of *nC-F^2DLT* to *LTqt* is meaningful since the two models are proved to be equivalent, as we report in the following theoretical result.

Proposition 1. *The Non-Competitive Trust Threshold Model (nC-F^2DLT) and the Linear Threshold Model with quiescence time (LTqt) are equivalent.* ◀

Proof. According to the definition of equivalence of two diffusion models in [33, 97], in order to prove the equivalence of *nC-F^2DLT* and *LTqt* we need to prove that the distribution of the *active sets* for any given seed set S_0 is the same under the two models. We provide a proof by induction, hence we consider the evolution of the active sets during the diffusion rounds.

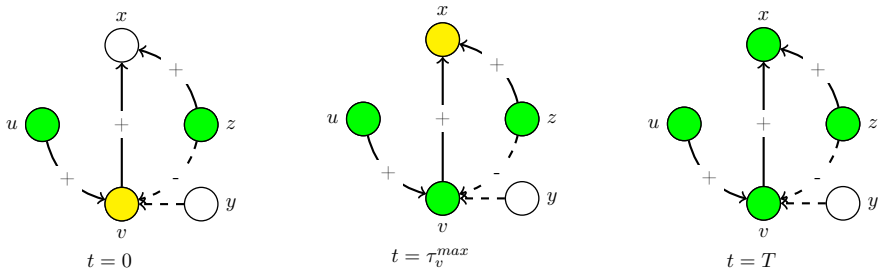


FIGURE 3.5: Activation sequence for the *LTqt* model.

For the *LTqt* model, the probability of a node to be activated exactly at time $t+1$ (with $t \geq 1$) is given by:

$$\begin{aligned} \Pr(v \in \widetilde{S_{t+1}} \mid v \notin S_t) &= \frac{\Pr(v \in \widetilde{S_{t+1}}, v \notin S_t)}{\Pr(v \notin S_t)} \\ &= \frac{\Pr(\sum_{u \in S_{t-1}} w_{uv} < \theta_v \leq \sum_{u \in S_t} w_{uv})}{\Pr(\sum_{u \in S_{t-1}} w_{uv} < \theta_v)} \\ &= \frac{\sum_{u \in S_t \setminus S_{t-1}} w_{uv}}{1 - \sum_{u \in S_{t-1}} w_{uv}} \end{aligned} \tag{3.4}$$

Above, it should be noted that the joint probability $\Pr(v \in \widetilde{S_{t+1}}, v \notin S_t)$ corresponds to the probability that the threshold associated with node v falls into the interval denoted by the influence received by v until the previous time step and the one received at the current time step. Moreover, $\Pr(v \notin S_t)$ is just the probability that, at time $(t-1)$, the influence received by v is still below its threshold. Finally, we derive the last equality in Equation 3.4, which intuitively denotes that the influence exerted by the nodes in $S_t \setminus S_{t-1}$, i.e., the nodes turning into the active state exactly in the current time step, is decisive to exceed the threshold θ_v.

For the *npC-F²DLT* model, the conditional probability $\Pr(v \in \widetilde{S_{t-1}} \mid v \notin S_t)$ can be derived starting from Equation 3.4 by constraining w_{uv} such that $u \in N_+^{in}(v)$, i.e., only trusted relations are considered. This leads to an equivalent definition of conditional probability, which holds for every time step t and seed set S_0. Therefore, we can conclude that the final active sets will be the same for both models.

□

It should be noted that, due to the quiescence times, the sets of active nodes in the two models may not be the same at every time step, but the two final active sets will match each other.

Since the introduction of quiescence time in LT does not have effect on the distribution of the final active nodes [33], we obtain the following equivalence: *LT* ≡ *LTqt* ≡ *nC-F²DLT*. Therefore, the activation function is still monotone and submodular under *nC-F²DLT*.

Example 2. *Consider Figure 3.5, where the propagation process unfolds according to the LTqt dynamics. Nodes u and z are chosen as initial seeds. Thresholds and weights are set such that* $\theta \leq w_{uv}$ *and* $\max\{w_{vx}, w_{zx}\} < \theta_x \leq w_{vx} + w_{zx}$*, therefore the combined influence of v and z is required for the activation of node x. The dashed edge denotes a distrust connection removed as a result of the reduction defined in Def. 4. In the initial time step* $(t = 0)$*, u activates v causing its transition from the inactive state to the quiescent state (in yellow). When* $t = \tau_v^{max}$*, v turns to the active state,*

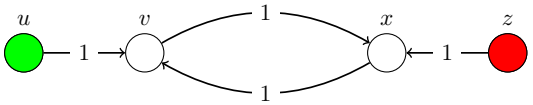


FIGURE 3.6: An example of non-terminating diffusion process

and together with z it becomes able to trigger the activation of node x (which will eventually become active by the time-horizon T).

It should be noted that the same dynamics holds for the nC-F²DLT model, apart from the difference that concerns the quiescence time of node v: this would be less than τ_v^{max} *since y, a foe of v, is not involved in the propagation process.* ■

3.3.5.2 Competitive diffusion

We focus here on *spC-F²DLT* and *npC-F²DLT*, and show that both models can be reduced to the *Homogeneous Competitive Linear Threshold* (*H-CLT*) with Majority Vote as tie-breaking rule [33]. This is a competitive, progressive model based on LT, for which it is known that its activation function is monotone but not submodular regardless of the particular tie-breaking rule.

To begin with, we might recall that the non-progressive LT-based diffusion can be reduced to the progressive case, using a particular form of *layered* graph []. Given a time interval T and a diffusion graph $G = \langle V, E\rangle$ for non-progressive LT, a new graph G^T can be derived such that every node $v \in V$ will have a replica v_t in every layer at time $t \in T$, and for every edge $(u, v) \in E$ there will be an edge (u_{t-1}, v_t) in G^T.

Unfortunately, this serialization technique cannot be directly applied to our models, since it is not designed to deal with competitive or non-progressive diffusion and it discards activation or delayed propagation aspects. In the following, we define serialization techniques that are suitable for our competitive models and treat one particular configuration at a time. One general requirement is related to the time horizon to bound the unfolding of the diffusion process. In fact, when dealing with competitive models, the termination guarantee is lost. A simple example is provided next to depict such a non-termination scenario.

Example 3. *In Figure 3.6, nodes u and z are chosen as seed for the green campaign and the red one, respectively. Nodes v and x become green-active and red-active, respectively, at time* $t = 1$. *Next, they will constantly switch their activation campaign, causing non-termination of the diffusion process.* ■

CONFIGURATION 1: No quiescence time, constant activation-threshold We assume that $q(v) = 0$ and $g(v, t) = \theta_v$, for all $v \in V, t \in T$. For both *spC-F²DLT* and *npC-F²DLT*, we claim their reduction to the *H-CLT* model with majority voting as tie-breaking rule.

Definition 5. ***spC-F²DLT graph serialization for reduction to H-CLT.*** *Given a time interval T, we define a layered graph* $G^T = \langle V^T, E^T\rangle$ *such that, for each layer at time* $t \in T$, *every node* $v \in V$ *will be represented in* V^T *as a tuple* $\langle v_t^1, v_t^2, v_t^3\rangle$. *Instances* v_t^1 *and* v_t^2 *have activation-threshold equal to 0, while* v_t^3 *has the same threshold as the original node* $v \in V$. *The set of edges is defined as* $E^T = \{(u_t^1, v_{t+1}^3) \mid (u, v) \in E, t, t+1 \in T\} \cup \{(v_t^3, v_t^2) \mid v \in V, t \in T\} \cup \{(v_t^2, v_t^1) \mid v \in V, t \in T\} \cup \{(v_t^1, v_{t+1}^2) \mid v \in V, t \in T\}$, *and the following constraint on edge weights must hold:* $\forall v_t^2 \in V^T,\ w(v_{t-1}^1, v_t^2) < w(v_t^3, v_t^2)$. □

In the above definition, triples act as *connectors* between two consecutive time-layers. The role of any connector component is as a sort of "switch" to enable a node choosing between its activation state in a layer and the one in the subsequent layer. In other words, node v_t^1 is the main instance of node v, since the activation state of v_t^1 reflects the state of v in the original graph, under *spC-F²DLT* at time t; node v_t^3 is the instance of v connected with other nodes from layer at $t-1$, therefore it reflects the influence received by v in the original graph, at time $t-1$; if the activation attempt to v_t^3 fails, node v_t^2 will be activated with the same state of v; otherwise, according to the edge weight constraint (cf. Def. 5), v_t^2 will switch to the other campaign, and then will propagate to instance v_t^1. Recall that v_t^1, v_t^2 have zero activation-threshold.

Figure A.1 in Appendix A.1 shows an example of serialization for a *spC-F²DLT* diffusion graph with time horizon set to 2.

It should be emphasized that, compared to the serialization method in [97], we require replication of each node in each layer, and additional edges connecting the replica-instances, in order to allow the maintenance of the activation state when no activation event occurs between two time-consecutive layers.

Analogous reduction technique can be defined for the *npC-F²DLT* model.

Definition 6. *npC-F²DLT* ***graph serialization for reduction to H-CLT.*** *Given a time interval T, we define a layered graph $G^T = \langle V^T, E^T \rangle$ such that, for each layer at time $t \in T$, every node $v \in V$ will be represented in V^T as a tuple $\langle v_t^1, v_t^2, v_t^3 \rangle$. Instances v_t^1 and v_t^2 have activation-threshold equal to 1 and 0, respectively, while v_t^3 has the same threshold as the original node $v \in V$. The set of edges is defined as $E^T = \{(u_t^1, v_{t+1}^3) \mid (u,v) \in E, t, t+1 \in T\} \cup \{(v_t^3, v_t^2) \mid v \in V, t \in T\} \cup \{(v_t^2, v_t^1) \mid v \in V, t \in T\} \cup \{(v_t^3, v_t^1) \mid v \in V, t \in T\} \cup \{(v_t^1, v_{t+1}^2) \mid v \in V, t, t+1 \in T\}$, and the following constraints on edge weights must hold: $\forall v_t^2 \in V^T,\ w(v_{t-1}^1, v_t^2) < w(v_t^3, v_t^2)$, and $\forall v_t^1 \in V^T,\ w(v_t^2, v_t^1) + w(v_t^3, v_t^1) = 1$.* □

It should be noted that the last condition in Def. 6 imposes nodes v_t^2 and v_t^3 to hold the same activation state in order to activate v_t^3.

Analogously to the reduction of *spC-F²DLT* to *H-CLT*, we can conveniently devise a notion of "connector" component between any two consecutive layers, which however in this case should also account for node deactivations.

Figure A.2 in Appendix A.1 shows an example of connector for the *npC-F²DLT* model.

Claim 1. *For any given diffusion graph G under spC-F²DLT (resp. npC-F²DLT), assuming constant activation-threshold and no quiescence time, every node v in G is active at time $t \in T$ if and only if its corresponding instance v_t^1 is active in the serialized graph G^T (resp. npC-F²DLT).* ◀

Configuration 2: Constant quiescence time, constant activation-threshold. We assume that $q(v) = \tau_v$ and $g(v,t) = \theta_v$, for all $v \in V$. For both *spC-F²DLT* and *npC-F²DLT*, we claim their reduction to *H-CLT* with majority voting as tie-breaking rule.

In this case, we need to consider that, whenever a node is activated, its quiescence time may not expire before the time horizon; for this reason, we will consider only nodes reachable from $S_0 = S_0' \cup S_0''$ within T, for any two given seed sets S_0' and S_0''. To identify such nodes, we define a *quiescence-aware distance* measure that accounts for the quiescence times along the path connecting any two nodes. Given nodes u, v, and the set $P(u,v)$ of all paths between u and v, the distance from u to v will be measured as $d(u,v) = \min_{p \in P(u,v)} \sum_{x \in p} \tau_x$. Moreover, we denote with $d(S_0, v)$ the

minimum distance between nodes $u \in S_0$ and v. By exploiting this distance, we will discard all nodes that cannot be "contagious" before the end of T, say t_{max}. Therefore, the node set V^T of the layered graph is defined as:

$$V^T = \{\langle v_t^1, v_t^2, v_t^3\rangle \mid \forall v \in V,\ t \in T,\ d(S_0, v) < t_{max}\}.$$

Each node $v \in V$ with quiescence time τ_v will have connections from the previous layers according to the following rule: for any layer at time t, if $t < d(S_0, v)$ then v will not have any incoming edges, otherwise all incoming edges of v will be from the layer at time $t - \tau_v - 1$.

Using the above settings in the serialization method previously presented, it can easily be demonstrated that both *spC-F²DLT* and *npC-F²DLT* can be reduced to an equivalent *H-CLT* model.

Claim 2. *For any given diffusion graph G under spC-F²DLT (resp. npC-F²DLT), assuming constant activation-threshold and constant quiescence time, every node v in G is active at time $t \in T$ if and only if its corresponding instance v_t^1 is active in the serialized graph G^T (resp. npC-F²DLT).* ◀

Configuration 3: Variable quiescence time, constant activation-threshold We assume that $q(v,t)$ is variable, while $g(v,t) = \theta_v$, for all $v \in V, t \in T$.

Like in the previous case, we need to specify the seed sets S_0', S_0''. However, note that the quiescence time of a node now depends on the actual activation state of its in-neighborhood (cf. Equation 3.3), which makes it unfeasible a direct serialization of the whole diffusion graph.

Starting from the original diffusion graph G, we derive an "intermediate" graph $\widehat{G}$, which is equivalent to G unless each node $v \in V$ is associated with a quiescence time interval $[\tau_v, \tau_v^{max}]$, where $\tau_v^{max} = \tau_v + \psi(N_-^{in}(v))$. Let us denote with G^{min} the instance of $\widehat{G}$ such that the quiescence time of every $v \in \widehat{G}$ is τ_v, and with G^{max} the instance of $\widehat{G}$ such that the quiescence time of every $v \in \widehat{G}$ is τ_v^{max}.

Although we cannot assert that *spC-F²DLT* and *npC-F²DLT* are equivalent to *H-CLT* under the layered graph obtained by applying the previously described serialization techniques, an important theoretical result can nonetheless be provided, as reported next.

Claim 3. *For any diffusion graph G under spC-F²DLT (resp. npC-F²DLT), with campaigns C', C'', assuming constant activation-threshold and variable quiescence time, for any seed sets S_0' and S_0'', it holds that:*

$$\sigma'_{H\text{-}CLTmax}(S_0', S_0'') \leq \sigma'(S_0', S_0'') \leq \sigma'_{H\text{-}CLTmin}(S_0', S_0''), \tag{3.5}$$

where σ' is the number of nodes activated by C' under spC-F²DLT (resp. npC-F²DLT), $\sigma'_{H\text{-}CLTmax}(S_0', S_0'')$ and $\sigma'_{H\text{-}CLTmin}(S_0', S_0'')$ are the number of nodes activated by C' under H-CLT in the layered graph obtained by serialization of spC-F²DLT (resp. npC-F²DLT) on G^{max} and G^{min}, respectively. ◀

Enabling variable quiescence time, i.e., $\psi(\cdot)$, means that the exact time required by each node to make a transition from the quiescent state to the active one cannot be established in advance at the beginning of the propagation process. Since for any node v the quiescent time ranges within $[\tau_v, \tau_v^{max}]$, we devise two opposite scenarios. In the first scenario, represented by the rightmost side of Equation 3.5, each node is assumed to wait the minimum amount of time, i.e., τ^{min}, before its activation; this leads to a higher fraction of nodes that could be activated before the time horizon

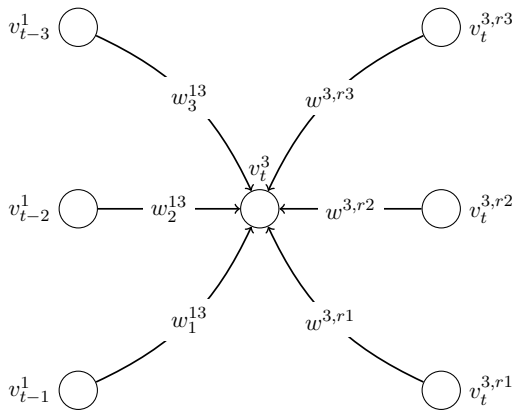


FIGURE 3.7: Example connector for modeling the time-varying activation-threshold
in the serialized graph under a competitive model.

T is reached. The second scenario, represented by the leftmost side of Equation 3.5, assumes that each node has to wait the maximum possible quiescence time, i.e., τ^{max}; as a consequence, a smaller fraction of nodes will be able to complete the activation process before the time limit, thus leading to a lower spread.

CONFIGURATION 4: No quiescence time, variable activation-threshold. We assume that $q(v) = 0$ and $g(v,t) = \theta_v + \vartheta(\theta_v, t)$, for all $v \in V, t \in T$. For both *spC-F²DLT* and *npC-F²DLT*, we claim their reduction to *H-CLT* with majority voting as tie-breaking rule. In the following, we refer to the biased activation-threshold function, although it is easy to show analogous considerations for the non-biased activation-threshold function.

Because of the dynamic behavior of the activation-threshold function, we cannot predict its value at any particular time step of the diffusion process; nevertheless, by specifying the value of coefficient δ in Equation 3.1, we can derive the value of t_v^{max}, which would suggest how many time-layers we have to look back in order to know the actual threshold value of v at a particular time t. In order to capture such dynamic aspect in *H-CLT*, we define a further serialization technique, built on top of the previously defined. We will restrict to a particular case, afterwards we provide some rules that apply to the general case.

Let us assume to focus on a particular node v, and at any two consecutive time steps of activation for the same campaign its threshold increases by δ. Again, node v will have *replicas* for any time-layer t, i.e., $\langle v_t^1, v_t^2, v_t^3 \rangle$, with the first replica, v_t^1, holding the actual state of v in the corresponding serialized graph for the competitive model. In addition, we introduce further replicas, in number equal to the value t_v^{max}; suppose, for the sake of simplicity, $t_v^{max} = 3$, we derive replica nodes $\langle v_t^{3,r1}, v_t^{3,r2}, v_t^{3,r3} \rangle$, such that each of them will have a threshold value in $[\theta_v, 1]$ with increment of δ.

Figure 3.7 illustrates this new component in the serialized graph.

Because this component is introduced as an extension of the previous techniques, the meaning of the nodes $v_{t-1}^1, v_{t-2}^1, v_{t-3}^1$ remains the same as in the previous cases. On the right side of Figure 3.7, each of the additional replicas has a different value of threshold and it is connected with nodes coming from the previous layers. Clearly, the overall behavior of this component depends on the weights attached to every edge in

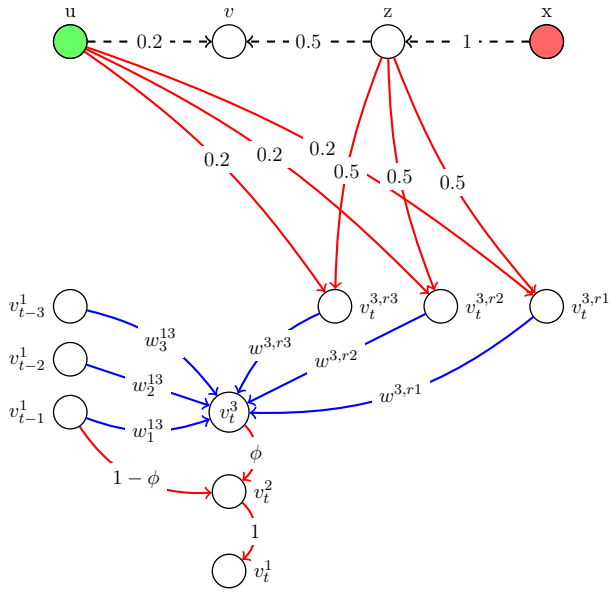


FIGURE 3.8: Serialization of a diffusion graph under a competitive model with time-varying activation-threshold.

the structure. In this regard, we define the following constraints on the edge weights:

$$\begin{cases} w^{3,r1} > w^{13}_1 & (a) \\ \forall i > 1 \quad w^{13}_i = w^{3,ri} & (b) \\ \forall i \geq 1 \quad w^{13}_i > \sum_{j>i}^{n} w^{13}_j & (c) \\ \forall i \geq 1 \quad w^{3,ri} > \sum_{j>i}^{n} w^{3,rj} & (d) \\ w^{3,r1} - w^{13}_1 < w^{13}_n & (e) \end{cases} \tag{3.6}$$

It should be noted that the activation attempts are performed directly on the replicas. Therefore, the above constraints on the edge weights control whether a node assumes the state derived as the outcome of the most recent activation attempts, or the one consistent with its personal history. as the outcome of the most recent activation attempts or the one consistent with its personal history. Each of the aforementioned inequality contributes to this decision process, following a different purpose. Equation 3.6(a) ensures that the state derived from the last activation attempt is always preferred to the one derived from the previous time step. Equation 3.6(b) ensures that the information coming from the previous time steps shall be given the same importance as the one derived from the current replicas. Equation. 3.6(c-d) ensures that the most recent information, i.e., the closest previous time steps, has higher priority than the earliest one. Equation. 3.6(e) ensures that there is consistency with respect to the state assumed in the closest previous time step and farthest involved time step (e.g., the third previous time step in the addressed scenario).

Moreover, the threshold of the "central" node in the component (v^3_t) is set to $w^{3,r1}$, to ensure sequentiality of the diffusion. By setting $\theta_{v^3_t}$ equal to $w^{3,r1}$, we avoid that v^3_t can be activated by its own replicas belonging to layers preceding the $t-1$-th layer.

Figure 3.8 shows how the above defined connector is integrated into a serialization technique. In the figure, only the connections incident on vertex v are expanded. The red edges are the ones connecting consecutive layers, therefore the replica $v^{3,r1}_t$ is connected with the previous layer, the replica $v^{3,r2}_t$ is connected with the second previous layer and so on. Blue edges represent the new connections due to the introduction of

TABLE 3.2: Summary of evaluation network data.

	Epinions	*Slashdot*	*Wiki-Conflict*	*Wiki-Vote*
#nodes	131 828	77 350	116 836	7 118
#edges	841 372	516 575	2 027 871	103 675
% distrusted/negative-edges	14.7%	23.3%	61.9%	21.6%
avg. out-degree	6.38	6.67	17.36	6.68
diameter	14	11	10	7
clust. coeff.	0.093	0.026	0.015	0.128
strong LCC #nodes	36 490	23 217	116 836	1 178
strong LCC #edges	602 722	243 600	2 027 871	31 572

this new component.

Claim 4. *For any given diffusion graph G under spC-F^{2}DLT (resp. npC-F^{2}DLT), assuming variable activation-threshold and no quiescence time, every node v in G is active at time $t \in T$ if and only if its corresponding instance v_t^1 is active in the serialized graph G^T (resp. npC-F^{2}DLT).* ◀

3.4 Evaluation methodology

3.4.1 Data

We used four real-world, publicly available networks, namely: *Epinions* [113], *Slashdot* [113], *Wiki-Conflict* [22] and *Wiki-Vote* [112]. Epinions is a "who-trust-whom" network of the homonymous review site. Slashdot models friend/foe relations between the users of the homonymous technology-related news website. Wiki-Conflict refers to Wikipedia users involved in an "edit-war", i.e., edges represent either positive or negative conflicts in editing a wikipage. Wiki-Vote models "who-vote-whom" relations between Wikipedia users that voted for/against each other in admin elections. Our choice of the evaluation datasets was mainly driven by two intents: (i) to provide a reproducible evaluation framework based on publicly available network data, and (ii) to test our models on a diversified set of real-world OSNs with suitable characteristics for information propagation processes.

Table 3.2 summarizes main structural characteristics of the networks. To favor meaningful competition of campaigns based on selected pairs of strategies, we limited the diffusion context to the largest strongly connected component in each evaluation network; note that, for Wiki-Conflict, the largest strongly connected component coincides with the whole graph. Also, the clustering coefficient corresponds to the definition of global transitivity in an undirected graph (the direction of the edges is ignored).

All networks are originally directed and signed; in addition, the two Wikipedia-based networks also have timestamped edges. In order to derive the weighted graphs of influence probabilities, we defined the following method: for every $(u, v) \in E$, the edge weight w_{uv} was sampled from a binomial distribution $\mathcal{B}(|N_+^{in}(v)|, p)$ if $u \in N_+^{in}(v)$ (i.e., v trusts u), otherwise $w_{uv} \sim -\mathcal{B}(|N_-^{in}(v)|, p)$, where the probability of success p is equal to the fraction of trust edges in the network;

the rationale is that for higher fraction of trusted connections in the network, the nodes will be more likely to trust each other, and hence each node is more likely to be involved in the propagation process.

We performed $1,000$ samplings of edge weights, for each of the four networks. Therefore, all presented results will correspond to *averages of* $1,000$ *simulation runs*.

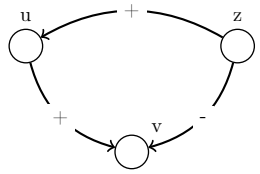


FIGURE 3.9: Stress configuration

3.4.2 Seed selection strategies

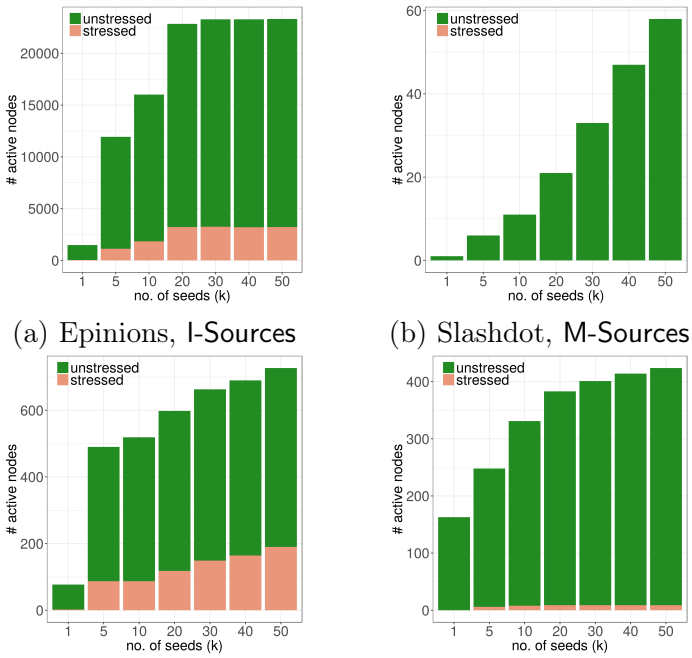
We defined four seed selection strategies, each of which mimics a different, realistic scenario of influence propagation.

Exogenous and malicious sources of information. This method, hereinafter referred to as M-Sources, aims at simulating the presence of multiple sources of malicious information within the network. Here, an exogenous source is meant as a node without incoming links, e.g., a user that is just interested in spreading her/his opinion: such a node is also regarded as malicious if a high fraction of outgoing influence exerted by the node is distrusted by out-neighbors. Formally, given a budget k, the method selects the top-k users in a ranking solution determined as $r(v) = (\bar{W}^-/(\bar{W}^- + \bar{W}^+)) \log(|N^{out}(v)|)$, for every v such that $N^{in}(v) = \emptyset$, where $\bar{W}^+, \bar{W}^-$ are shortcut symbols to denote the sum of trust (resp. distrust) weights, respectively, outgoing from v.

Exogenous and influential trusted sources of information. Analogously to the previous method, this one, dubbed I-Sources, searches for the "best" influential trusted sources. The ranking function is as $r(v) = (\bar{W}^+/(\bar{W}^- + \bar{W}^+)) \log(|N^{out}(v)|)$. Note that this still takes into account the negative weights, because even a highly trusted user might be distrusted by some other users (e.g., "haters").

Stress triads. This strategy is based on the notion of *structural balance* in triads [113]. Figure 3.9 shows an example of stress-triad configuration: node v has two incoming connections, the one from node z with negative weight, and the other from u with positive weight, and there is also a trust link from z to u. We say that z is a *stress-node* since, despite the distrusted link to v, it could also indirectly influence v through the trusted connection with u. Based on that, our proposed Stress-Triads strategy searches for all triads containing stress-nodes and selects as seeds the first k stress-nodes with the highest number of triads they participate to.

Newcomers. We call a node $v \in V$ as a *newcomer* if all of its incoming edges are timestamped as less recent than its oldest outgoing edge. The *start-time* of v is the oldest timestamped associated with its incoming edges. We divide the set of newcomers into two groups obtained by equal-frequency binning on the temporal range specific of a network. Upon this, we distinguish between two strategies, dubbed Least-New and Most-New, which correspond to the selection of k newcomers having highest out-degree among those with the oldest start-time and with the newest start-time, respectively. Both strategies were applied to Wiki-Vote and Wiki-Conflict, due to the availability of timestamped edges.



(a) Epinions, I-Sources (b) Slashdot, M-Sources

(c) Wiki-Vote, Stress-Triads (d) Wiki-Conflict, Most-New

FIGURE 3.10: Spread of nC-F^2DLT by varying seed set size (k) and selection strategy.

3.4.3 Settings of the model parameters

For every user v, the exogenous activation-threshold θ_v and quiescence time τ_v were chosen uniformly at random within [0,1] and [0,5]. Moreover, λ (used in the quiescence function) was varied between 0 and 5, while the coefficient δ (used in the activation-threshold function) was selected in $\{0, 0.1\}$ for the biased scenario (Equation. 3.1) and kept fixed to 1 for the unbiased scenario (Equation 3.2).

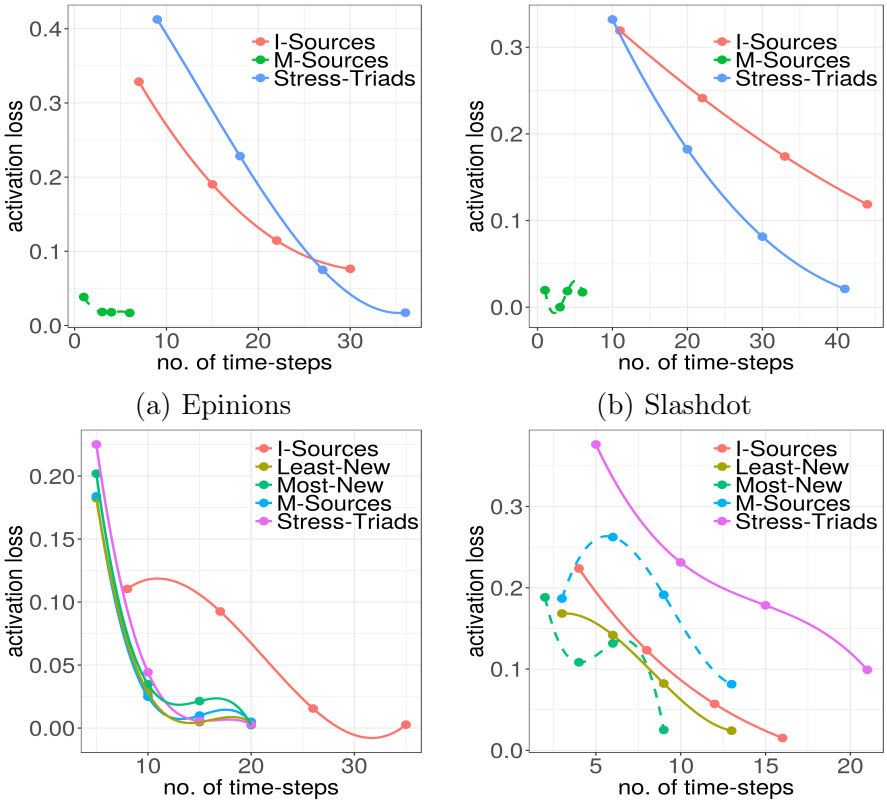
3.5 Results

We organize the presentation of our experimental results into three parts. The first part is devoted to the evaluation of the non-competitive model (Section 3.5.1), and the second part for the competitive models (Sect. 3.5.2). In the third part (Section 3.5.3), we present a comparative evaluation of our non-competitive model against IC and stochastic individual-contact epidemic models, whereas for the competitive scenario, we compare our models with the DLT model [130].

3.5.1 Evaluation of nC-F^2DLT

3.5.1.1 Spread, stressed users and negative influence

We analyzed the number of final activated users (i.e., *spread*) by varying the size (k) of seed set, for every seed selection strategy. In this analysis, we assumed constant activation thresholds (i.e., $\vartheta(\cdot,\cdot) = 0$) and constant quiescence times (i.e., $\psi(\cdot,\cdot) = 0$). Moreover, we distinguished between "*stressed*" and "*unstressed*" users, being the former regarded as active users having at least one distrusted active in-neighbor. As shown in Figure 3.10 for some representative cases, besides the expected growth in spread as k increases, we found the activation of stressed users lower in amount but



(a) Epinions (b) Slashdot

(c) Wiki-Conflict (d) Wiki-Vote

FIGURE 3.11: Activation loss due to time-varying quiescence (for $\lambda = 5$, $k = 50$) under the nC-F^2DLT model.

following similar trend as that corresponding to unstressed users. For both types of users, I-Sources revealed higher spread capability, followed by Stress-Triads, in all networks (with the exception of Wiki-Vote). The two newcomers-based strategies (where applicable) turned out to be effective as well, with Least-New prevailing on Most-New for lower k. By contrast, M-Sources was in general unable to yield a spread comparable to other strategies.

We further investigated the effect of distrusted connections on the spread during the unfolding of the diffusion process.

In this regard, Table 3.3 shows the amount of nodes that, at the time of their involvement in the propagation process, were *negatively influenced* by in-neighbors activated at any previous time, along with their perceived negative influence. The symbol "-" in Table 3.3 denotes that the corresponding seed-selection strategy does not apply to a particular network.

In general, we observed a significant presence of negative influence spread when using I-Sources and Stress-Triads. Considering Epinions and Slashdot, the former (resp. the latter) corresponded to a negative influence spread of the order of thousands (resp. hundreds), with average influence weight around 0.3 (resp. 0.2). The impact of these strategies was lower in the Wikipedia networks (one order of magnitude below). M-Sources yielded to null (in Epinions and Slashdot) but also non-negligible (in Wiki-Vote and Wiki-Conflict) spread. By contrast, the newcomers-based strategies had small (in Wiki-Vote) or negligible (in Wiki-Conflict) effect on the negative influence spread.

TABLE 3.3: Summary about negative influence spread ($k = 50$).

Network \ Strategy		M-Sources	I-Sources	Stress-Triads	Most-New	Least-New
Epinions	# nodes	0	2 117	847	-	-
	avg weight	0	0.30	0.22	-	-
Slashdot	# nodes	0	4 599	345	-	-
	avg weight	0	0.32	0.19	-	-
Wiki-Conflict	# nodes	13	829	26	1	0
	avg weight	0.22	0.05	0.01	0.02	0
Wiki-Vote	# nodes	45	27	175	10	12
	avg weight	0.21	0.13	0.22	0.04	0.07

3.5.1.2 Activation loss

As partially unveiled by the previous analysis, the users' involvement in the propagation process is affected by the behavior of the quiescence function, whose impact would increase with the amount of distrusted influence in the spread. This further prompted us to measure the *activation loss*, i.e., the percentage decrease of activated users, due to the enabling of the time-varying quiescence factor (i.e., $\lambda > 0$ in Equation 3.3) in the users' activation states. Figure 3.11 shows results corresponding to relatively large λ (set to 5) and k (set to 50). For each seed selection strategy, the curve is drawn by using polynomial splines,[2] where the marked points (from low to high time steps) refer to the 25%, 50%, 75% and 100% of the time horizon observed for the diffusion process under the chosen strategy without time-varying quiescence times. One general remark that stands out is a relatively high percentage of activation loss for the initial time steps; this holds in particular for Stress-Triads, which might be explained since the initial influenced users by means of this strategy tend to be subjected to a certain amount of distrusted influence. As the time steps get closer to the time horizon, the activation loss tends to significantly decrease, down to nearly zero in most cases, with few exceptions including the use of I-Sources in Slashdot and Epinions, and Stress-Triads and M-Sources in Wiki-Vote — note this is indeed consistent with the previous analysis on negative influence spread.

3.5.2 Evaluation of competitive models

To analyze the behavior of *spC-F²DLT* and *npC-F²DLT*, we aimed at simulating a scenario of *limitation of misinformation spread*, i.e., we assumed that one campaign, the "bad" one, has started diffusing, and consequently another campaign, the "good" one, is carried out in reaction to the first campaign.

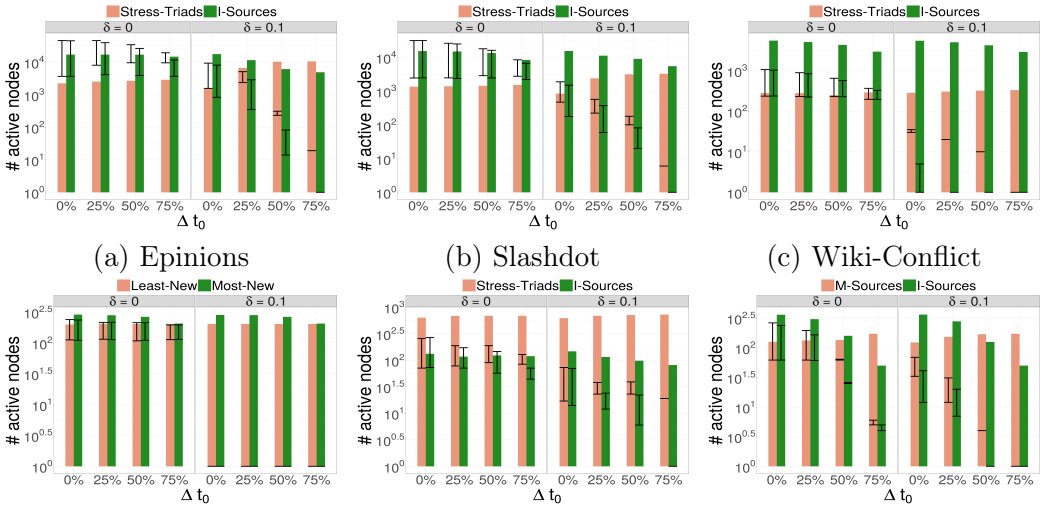
3.5.2.1 Combining seed selection strategies

Within this view, we preliminarily investigated about strategy combinations that might be reasonably considered for a misinformation spread limitation problem. Table 3.4 provides a number of statistics we collected to characterize selected pairs of strategies, for two campaigns carried out independently to each other, i.e., in a non-competitive scenario, with $k = 50$. Using Stress-Triads for the bad campaign and I-Sources for the good campaign was found to be significant for all networks, with sharing percentage close to 100% in Epinions and Slashdot and above 80% in Wiki-Conflict. Also, pairing M-Sources with I-Sources, and Least-New with Most-New, was well-suited in Wiki networks.

[2] We used the *splines2* R-package, available at `https://CRAN.R-project.org/package=splines2`.

TABLE 3.4: Statistics about selected pairs of strategies for two campaigns: the seed set $S_0^{(1)}$ (resp. $S_0^{(2)}$) computed for the first-started or "bad" (resp. second-started or "good") campaign SS_1 (resp. SS_2), the spread $|\Phi(S_0^{(1)})|(resp.|\Phi(S_0^{(2)})|)$, the Forest Wiener Index (FWI) [103] to measure the structural virality over the k seed-rooted diffusion trees, the fraction of spread of the bad campaign shared with the good campaign (*shared* column), the percentage of shared users that were activated first by the bad campaign (SS_1 *first* column), the average time of activation of the shared users, and the average time of activation of the shared users by the bad campaign before the good campaign, and vice versa.

network	SS_1	SS_2	$\lvert\Phi(S_0^{(1)})\rvert$	$\lvert\Phi(S_0^{(2)})\rvert$	FWI_1	FWI_2	shared	SS_1 first	avg. activation time		
									any	SS_1 first	SS_2 first
Epinions	Stress-Triads	I-Sources	10595	23321	8.63	8.61	0.99	28%	6.03	0.67	5.27
	M-Sources	I-Sources	59	23321	0.12	8.61	0.01	100%	4.0	3.0	0.0
Slashdot	Stress-Triads	I-Sources	3263	18671	6.54	10.43	0.98	40%	6.56	2.54	7.63
	M-Sources	I-Sources	58	18671	0.13	10.43	0.05	100%	4.0	5.0	0.0
Wiki-Conflict	Stress-Triads	I-Sources	344	5968	1.96	4.93	0.84	95%	2.43	1.43	4.93
	M-Sources	I-Sources	203	5968	0.0	4.93	0.75	98%	3.64	0	9.5
	Least-New	Most-New	216	424	0.07	0.86	0.7	100 %	3.77	0	0
Wiki-Vote	Stress-Triads	I-Sources	727	394	5.36	3.35	0.45	78%	4.32	0.54	5.87
	M-Sources	I-Sources	172	394	2.28	3.35	0.41	79%	4.04	0.05	3.93
	Least-New	Most-New	165	159	1.03	1.74	0.13	63%	2.72	0.0	4.37



(a) Epinions (b) Slashdot (c) Wiki-Conflict

(d) Wiki-Conflict (e) Wiki-Vote (f) Wiki-Vote

FIGURE 3.12: *spC-F²DLT*: Spread, number of switched users, and number of switches (log scale) by varying start-delay (Δt_0) of the "good" campaign (second bars),for $\delta = 0$ (left-most bar groups) and $\delta = 0.1$ (right-most bar groups), $k = 50$.

3.5.2.2 Setting and goals for the evaluation of competitive diffusion

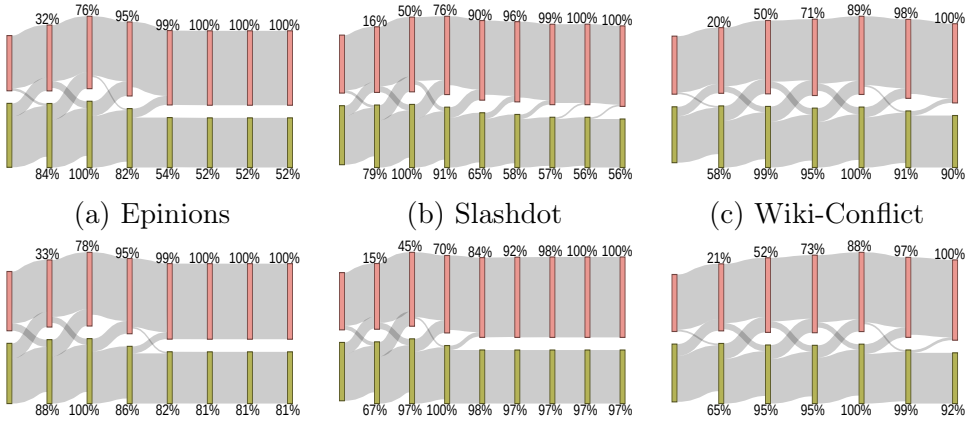
As previously mentioned, the seed selection strategies chosen for the two campaigns might not start at the same time, in which case we assume that the first-started one is the bad campaign. Moreover, we used fixed-probability as tie-breaking rule, with probability equal to 1 for the bad campaign. Also, we set the time horizon to the end-time of the (non-competitive) diffusion of the bad campaign.

Our main goal in the analysis of the two competitive models was to understand the effect of the setting of the activation-threshold function on the users' campaign-changes/deactivations, under the case of "real-time correction" or "delayed correction" by the good campaign against the bad one (cf. Introduction).

3.5.2.3 Evaluation of *spC-F²DLT*

We present results on the campaign spreads, the number of users activated for one campaign that *switched* to the other campaign, and the total number of switches; the latter two measurements are represented, in the barcharts shown in Figure 3.12, by the lower and upper whiskers, respectively, in the linerange vertically placed on each bar. Results correspond to start-delays Δt_0 of the good campaign w.r.t. the bad one (from 0 to 75% of the end-time of the bad campaign). For this analysis, we considered the biased definition of the activation-threshold function (Equation 3.1).

One general remark is that, for $\delta = 0, \Delta t_0 = 0$, the seed strategy that showed to be most effective in spread in the non-competitive case (cf. Table 3.4) confirmed its advantage against the other campaign's strategy. Nevertheless, for $\delta > 0, \Delta t_0 > 0$, the two campaigns would tend to an equilibrium, or even to invert their trend (e.g., in Epinions and Wiki-Vote). In particular, by accounting for (even little) confirmation bias and letting both campaigns start at the same time, I-Sources slightly increases its spread (which is explained since this strategy allows for activating first a high fraction of shared users, e.g., 70% in Epinions); but, as the start-delay increases at 50%, the good campaign is no more able to save users from being influenced by the bad campaign (i.e., Stress-Triads in Epinions, M-Sources in Wiki-Vote).



(a) Epinions (b) Slashdot (c) Wiki-Conflict

(d) Epinions (e) Slashdot (f) Wiki-Conflict

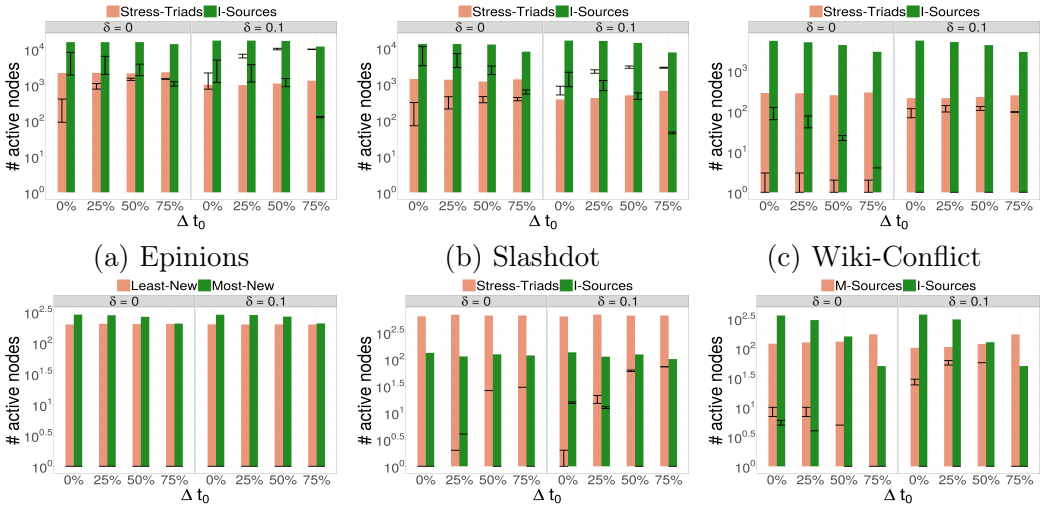
FIGURE 3.13: *spC-F^2DLT*: spread of I-Sources (red) vs. Stress-Triads (green),
for (a)-(c) the unbiased scenario and (d)-(f) the biased scenario of activation-threshold
function, with δ set to 1 and 0.1, respectively, and $k = 50$.

Interesting remarks were also drawn from the analysis of the transitions from one campaign to the other one. For $\delta = 0$, as the start-delay increases, the number of switched users follows a nearly constant trend in all networks (but Wiki-Vote, where we observed a drastic decrease for both campaigns), while the total number of switches is subjected to a more evident decreasing trend. Moreover, we observed a higher number of (unique and total) switches from the bad campaign to the good campaign, than vice versa, which occurred even when the spread of the bad campaign was higher than the good one (e.g., in Wiki-Vote, for both combinations of strategy choices). Setting $\delta = 0.1$ led to a general decrease in the switch measurements w.r.t. the corresponding previous case, and also to a substantial increase in "saved" users by the good campaign.

Biased vs. unbiased activation-threshold function. We also investigated how our proposed semi-progressive model behaves under the unbiased scenario corresponding to the activation-threshold function (Equation 3.2).

Figure 3.13 shows flow diagrams of the spread based on *spC-F^2DLT* for the selection of strategies I-Sources (red color) and Stress-Triads (green color), with the activation-threshold function defined either for the unbiased scenario (plots on the top) or for the biased scenario (plots on the bottom). In each plot, the height of a vertical bar along with the percentage displayed upon it, denote the number of active users at a particular time step and the ratio w.r.t. the maximum number of active users achieved by the corresponding selection strategy. The space between two consecutive bars corresponds to a time window, here set to 6 time steps for readability reasons. In each window we record two main events: (i) the number of active nodes that keep the same activation state, represented in base-2 logarithmic scale by the flow connecting two consecutive bars for the same campaign, and (ii) the number of users that switched from one campaign to the opposite one, represented in base-10 logarithmic scale by the flow connecting two consecutive bars with different colors.[3]

[3]The choice of two different logarithmic scales to represent the active users and the switched users is for the sake of readability of the plots, since the number of switched users is typically orders of magnitude smaller than the number of active users.



(a) Epinions (b) Slashdot (c) Wiki-Conflict

(d) Wiki-Conflict (e) Wiki-Vote (f) Wiki-Vote

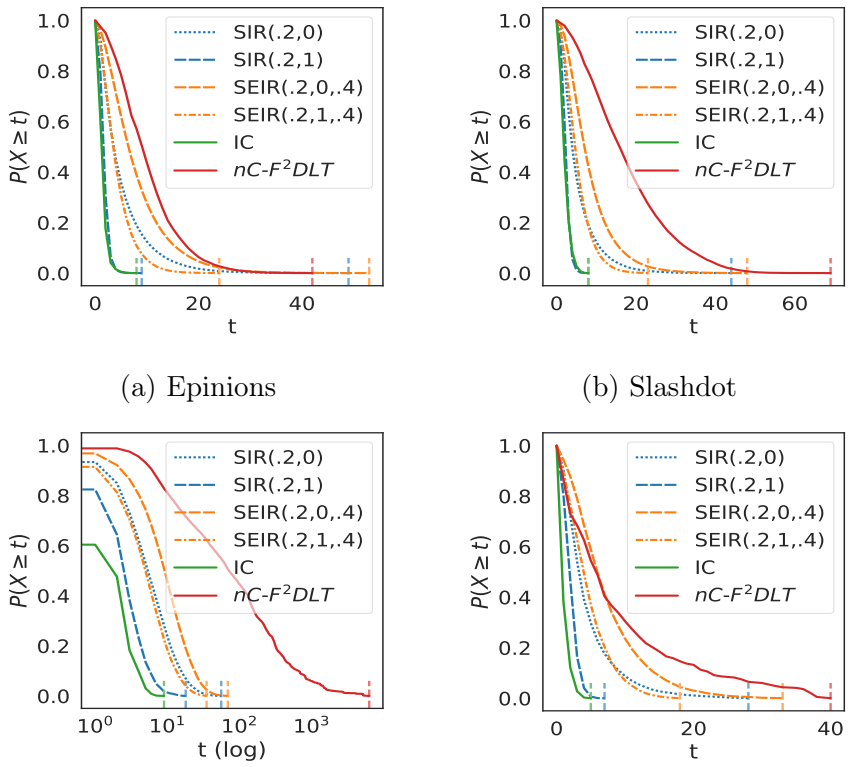
FIGURE 3.14: *npC-F²DLT*: Spread, number of deactivated nodes, and number of deactivations (log scale) by varying start-delay (Δt_0) of the "good" campaign (second bars), for $\delta = \{0, 0.1\}$, $k = 50$.

Note that for this analysis we discarded the start delay for the good campaign.

As expected, the number of switched users tends always to be in favor the good campaign, which has typically the best strategy of activation. However, and more importantly, the number of switched users in the unbiased scenario is significantly greater than in the biased scenario. Moreover, when the confirmation-bias effect is enabled, the majority of the switches are concentrated in the initial time-windows, then they follow a rapid decreasing trend until the time horizon. On the contrary, in the unbiased scenario, there is still a concentration of switches in the early stages of the propagation, but it becomes less evident and the number of switches tends to decrease more smoothly as opposed to the confirmation-bias scenario. This is particularly evident in Slashdot, where switches last until the latest time-windows, while in the confirmation-bias scenario the switches stop just after the third time-window. No significant differences can be observed on Wiki-Conflict, which is explained since the majority of shared nodes are activated in the early stage of the propagation, where the diffusion seems to behave in the same way regardless of the particular activation-threshold function.

3.5.2.4 Evaluation of *npC-F²DLT*

Compared to the evaluation of *spC-F²DLT*, the spread trends observed under *npC-F²DLT* showed no particular differences. However, more importantly, the occurrence of de-activation events, which are admitted by *npC-F²DLT*, appeared to favor the good campaign strategy, as shown in Figure 3.14. In particular, in Epinions and Slashdot, the number of user-unique and total deactivations tend to increase for the bad campaign and to decrease for the good one; moreover, although the spread of the good campaign remains higher, the deactivations for the good campaign are more frequent than those for the bad campaign as long as the start-delay remains zero or low, and the confirmation bias factor is not introduced. A few differences arise in Wiki-Conflict. As concerns Stress-Triads vs. I-Sources, although the 95% of shared users is activated first by the bad campaign, this advantage revealed not to be enough to avoid that the good



(a) Epinions (b) Slashdot

(c) Wiki-Conflict (d) Wiki-Vote

FIGURE 3.15: Complementary cumulative distribution functions of node activations/
infections for $nC\text{-}F^2DLT$, IC, SIR ($\beta = 0.2$ and $\gamma \in \{0, 1\}$), and SEIR ($\beta = 0.2$, $\gamma \in$
$\{0, 1\}$, $\sigma = 0.4$), using $k = 50$ and strategy I-Sources.

campaign will eventually activate more users. In fact, the number of deactivations with $\delta = 0.1$ increases for Stress-Triads and is always higher than for I-Source, in which the statistic remains nearly constant by increasing the start-delay. Similar situation was observed for the combination M-Sources with I-Sources. Also, using Least-New with Most-New led to no deactivations, which might be explained since the totality of shared users was reached first by the bad campaign (cf. Table 3.4).

3.5.3 Comparative evaluation

We conducted a twofold comparative evaluation, divided in two stages. The first one refers to the non-competitive scenario, whereby we compared $nC\text{-}F^2DLT$ to two epidemic models, i.e., SIR and SEIR, and the IC model. inspired by the studies in the epidemiology field. The second stage of our evaluation addresses the comparison between $spC\text{-}F^2DLT$ with the DLT model [130], which is the closest to our work, as we previously discussed in Section 3.2.

3.5.3.1 Comparison with the IC, SIR and SEIR models

We begin with briefly recalling the basic principles underlying the competitor models considered in this section. The independent cascade (IC) model is a stochastic discrete-time diffusion model like LT, such that once a node becomes active, in the

following time step of propagation it has a single chance of activating each of its out-neighbors. As concerns the epidemic models SIR and SEIR, the individuals of a population are divided in compartments that describe one of the following epidemiological states: susceptible (S), infective (I), latent-period or exposed (E), and recovered (R); therefore, individuals transition through those states. It is important to note that, since we need to treat each node in a network as an individual agent in order to enable a comparison with our diffusion model, we implemented SIR and SEIR based on a *stochastic individual-contact network* modeling as opposed to the standard, deterministic compartmental modeling (cf. Section 3.2). Within this view, the infection process in SIR is governed by two main parameters: (i) the transmission or contact rate (β), i.e., the probability of a susceptible node to be infected by any of its infected in-neighbors; and (ii) the recovery rate (γ), i.e. the probability of an infected node to transition to the recovery state with immunity, thus consequently stopping propagating the disease along the network. Moreover, in the SEIR model, the transition to the exposed state is governed by the incubation rate (σ), which defines the average duration of incubation as $1/\sigma$; note that the notion of exposed state somehow resembles the quiescent state of nodes in $nC\text{-}F^2DLT$.

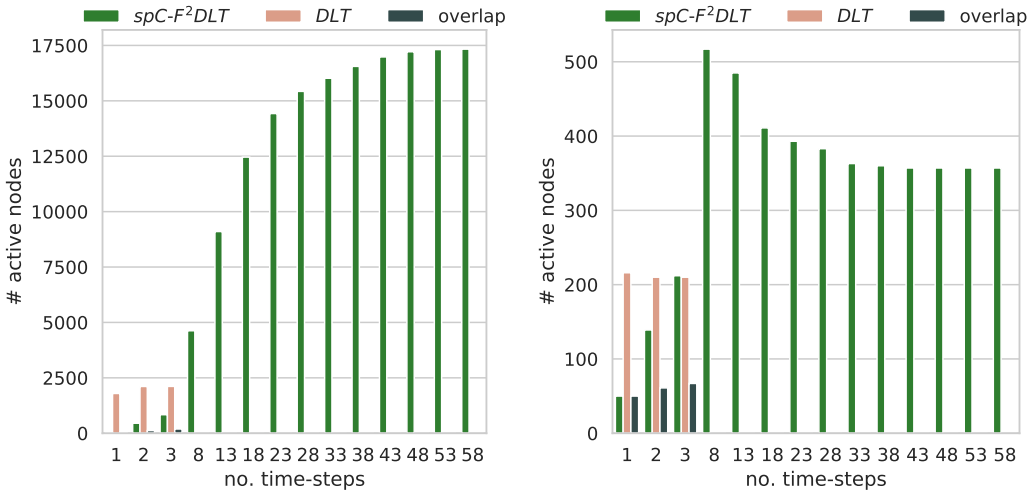
Figure 3.15 shows the complementary cumulative distribution function (CCDF) of the probability for a node of being active/infected from any given time step t to the termination of the process. The presented results correspond to β set to 0.2 and γ set to either 0 or 1: note that $\gamma = 1$ implies that any node recovers immediately after its activation, and hence similarly to the IC model it has a single chance for activating its susceptible out-neighbors; by constrast, setting $\gamma = 0$ implies that a node is unable to recover after its activation, therefore similarly to $nC\text{-}F^2DLT$ it will continue to contribute to the activation of its susceptible/inactive out-neighbors until the end of the process. (Further results for other settings of β and γ are reported in Appendix B.) Moreover, for SEIR, we set $\sigma = 0.4$, thus imposing an average incubation time equal to 2.5 time steps; this setting enables a fair comparison with our model, since each node will be expected to spend the same amount of time in the quiescent state for $nC\text{-}F^2DLT$ (cf. Section 3.4.3) as in the exposed state for SEIR.

Looking at the figure, as expected IC and SIR (with $\gamma = 1$) show an almost identical behavior, since most activations occur in the early stage of the propagation. On the contrary, when $\gamma = 0$, the SIR model tends to behave relatively closer to $nC\text{-}F^2DLT$ rather than IC, since the activations appear more uniformly distributed along the lifetime of the process. Also, the introduction of the exposed state in the SEIR model forces the dynamics of the propagation to be further more similar to the $nC\text{-}F^2DLT$ model, especially with $\gamma = 1$. One general remark that stands out is that $nC\text{-}F^2DLT$ tends to favor a slower diffusion, since the propagation process lasts consistently longer than IC and the epidemic models. Moreover, $nC\text{-}F^2DLT$ yields a smoother behavior in terms of time-decay of its CCDF than those corresponding to the other models.

In general, we can state that already for the non-competitive scenario, epidemic models even in their stochastic contact network formulation provide a differnt solution in terms of behavioral dynamics w.r.t. our proposed $nC\text{-}F^2DLT$.

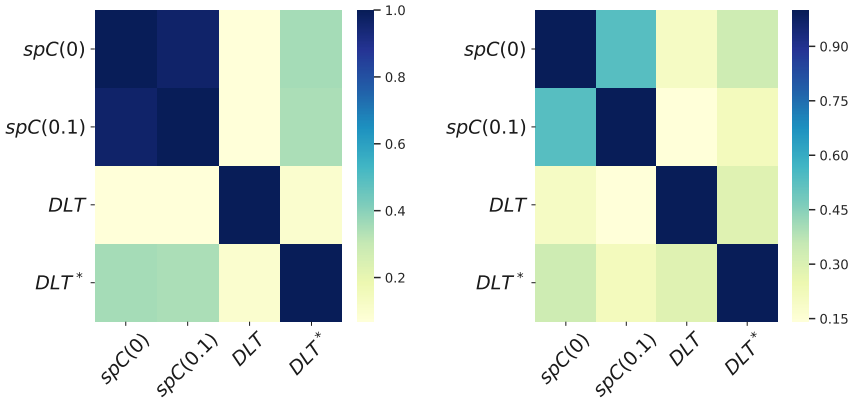
3.5.3.2 Comparison with the DLT model

We finally conducted a stage of comparative evaluation with the DLT method [130] (cf. Section 3.2). To this purpose, we analyzed the trends of spread and corresponding overlaps of activated nodes, under a competitive scenario. For DLT, we considered two cases: the one including the decay of influence probabilities (with Poisson decay



(a) I-Sources (b) Stress-Triads

FIGURE 3.16: $spC\text{-}F^2DLT$ vs. DLT: spread trends and overlaps, over time up to
convergence of $spC\text{-}F^2DLT$, on Slashdot.



(a) I-Sources (b) Stress-Triads

FIGURE 3.17: $spC\text{-}F^2DLT$ vs. DLT: overlap percentages at convergence of the two
models, on Slashdot.

coefficient set to 1), and the other one discarding the influence decay (as also studied in [130]), hereinafter dubbed DLT*; as concerns our models, we were forced to use $spC\text{-}F^2DLT$ since DLT does not allow node deactivation.

Figure 3.16 shows results obtained on Slashdot, for the choice of strategies I-Sources and Stress-Triads (similar trends were observed for Epinions and Wiki networks). A first remark is that, regardless of the seed selection strategy, the diffusion process under DLT terminates in a very few time steps, mainly due to the influence decay factor. Also, before convergence, DLT enables the activation of more nodes than $spC\text{-}F^2DLT$, though this actually corresponds to a small portion of the finally activated nodes by $spC\text{-}F^2DLT$ and, in any case, with an overlap that is generally below the 50%.

We explored more in detail the spread overlap between $spC\text{-}F^2DLT$ and DLT as well as its variant without decay (DLT*); for $spC\text{-}F^2DLT$, we considered the settings

$\delta = 0$ and $\delta = 0.1$. Figure 3.17 shows the heatmaps for the percentages of overlap of activated nodes at convergence of their respective models. As we observe in both plots, and regardless of δ, the overlap between DLT* and *spC-*F^2DLT is around 40-45%, which further drops to less than 10% when the influence decay is considered (the lighter, the lower is the overlap).

Overall, we can conclude that DLT, and even its variant DLT* without influence decay, behaves significantly different from our semi-progressive F^2DLT.

3.6 Discussion and usage recommendations

Our theoretical inspection of the proposed models, whose technical details have been presented in Section 3.3.5, revealed two important findings:

(F1): The non-competitive, progressive model, *nC-*F^2DLT, is proven to be equivalent to *LT with Quiescence Time*; therefore, the activation function in *nC-*F^2DLT is monotone and submodular.

(F2): The competitive, non-progressive models, *spC-*F^2DLT and *npC-* F^2DLT, can be reduced, via graph serialization, to *Homogeneous Competitive LT* [33], which is competitive and progressive, and has monotone, non-submodular activation function; therefore, the activation function in *spC-*F^2DLT and *npC-*F^2DLT is monotone but not submodular. It should be emphasized that the basic technique of graph serialization introduced in [97] to reduce the non-progressive LT-based diffusion to the progressive case, cannot be applied to our proposed models, since it is not designed to deal with competitive or non-progressive diffusion and it discards activation or delayed propagation aspects; to overcome this issue, we provided new serialization techniques and relating definitions of layered-graphs that are suitable for our competitive models, focusing on particular settings of the activation-threshold and quiescence functions.

The two findings clearly have different impact on the development, upon our F^2DLT models, of approximate solutions to influence maximization, rumor blocking, and related problems. On the other hand, in terms of expressiveness of our competitive F^2DLT models, it should be noted that the serialization techniques require the construction of layered graphs whose size easily grows with some of the models' parameters, making the application of such serialized graphs unfeasible at a large scale. Therefore, using our competitive F^2DLT models turns out to be essential in the representation of complex, dynamic propagation phenomena.

Our proposed class of trust-aware, dynamic models for non-competitive and competitive information diffusion offers a versatile solution for enhanced understanding of complex influence-propagation phenomena that occur in real-life network scenarios.

Our models are also unique, since they have significantly different behavioral dynamics w.r.t. epidemic models and the dynamic linear threshold model, according to theoretical considerations that were also clearly supported by empirical evidence in our experimental evaluation.

It should be noted that the setting of the dynamic activation-threshold function ϑ and quiescence function ψ, especially of their parameters δ and λ, respectively, plays a crucial role in the expressiveness of our models, thus differently impacting on positive-influence propagation and on negative-influence/misinformation limitation. We make the following recommendations for the usage of our models.

- The results of our evaluation revealed that the average user's sensitivity in the negative influence perceived from distrusted neighbors (which is controlled by λ) makes the seed identification process more aware of the negative influence

spread, thus considering the quiescence-biased contingencies by which a non-negligible fraction of users cannot be activated before the time limit.

- The confirmation-bias effect underlying δ may lead the "stronger" campaign (i.e., the one able to activate most users at the early steps of its diffusion) to increase its spread capability.
- As shown by simulations under the semi-progressive competitive model ($spC\text{-}F^2DLT$), the combined effect of increased δ with an increase in the delay of the beginning of the second-started (good) campaign may reduce its capability of "saving" users from the influence of the bad campaign; therefore, to limit misinformation spread, the good campaign should concentrate its (activation) efforts in the early stage of its diffusion.
- The non-progressive competitive model ($npC\text{-}F^2DLT$) appears to be less sensitive to the increase of δ. Yet, it tends to favor deactivation events (for users previously activated by the weaker campaign) over switched events.

In this regard, it would be interesting to study how to learn the various parameters in our models for the corresponding IM scenarios. Note that learning parameters for IM tasks is a challenging problem, which is still largely open, given the relatively little work done even under basic diffusion models [73, 168]. Major difficulties are in the assumption of availability of past propagation data from which the parameters would be learnt, which is in general difficult to obtain, and the large number of parameters, which poses efficiency issues. To overcome these aspects, the approach in [193] appears to be particularly promising, as it does not depend on the rules that control how the propagation unfolds over time. Nonetheless, we expect that the learning problem in our setting will easily become much more challenging, given the presence of other parameters than just the diffusion probabilities, and the need for coping with competitive influence scenarios. Therefore, we believe there will be much work to do in such a direction.

3.7 Chapter notes

In this chapter we proposed a novel class of trust-aware, dynamic LT-based models for non-competitive and competitive influence propagation in information networks. Evaluation on real-world, publicly available networks included simulations of scenarios of misinformation spread limitation, based on realistic strategies of selection of the initial influential users.

We believe that our proposed models can pave the way for the development of sophisticated methods to solve misinformation spread limitation and related optimization problems. Remarkably, our models can profitably be used in a variety of applications whereby there is an emergence *to predict the time required to debunk fake information*, or *to estimate how people are affected by the spread of competitive opposite opinions* through a social network. Also, we envisage an effective *support for fact-checking*, through a contextualization of the activation and quiescence functions to the production/consumption of contents in interaction networks.

Chapter 4

Topological Characterization of the Most Influential Nodes

Estimating the spreading potential of nodes in a social network is an important problem which finds application in a variety of different contexts, ranging from viral marketing to spread of viruses and rumor blocking. Several studies have exploited both mesoscale structures and local centrality measures in order to estimate the spreading potential of nodes. To this end, one known result in the literature establishes a correlation between the spreading potential of a node and its *coreness*: i.e., in a core-decompostion of a network, nodes in higher cores have a stronger influence potential on the rest of the network. In this chapter we show that the above result does not hold in general under common settings of propagation models with submodular activation function on directed networks, as those ones used in the influence maximization (IM) problem.

Motivated by this finding, we extensively explore where the set of influential nodes extracted by state-of-the-art IM methods are located in a network w.r.t. different notions of graph decomposition. Our analysis on real-world networks provides evidence that, regardless of the particular IM method, the best spreaders are not always located within the inner-most subgraphs defined according to commonly used graph-decomposition methods. We identify the main reasons that explain this behavior, which can be ascribed to the inability of classic decomposition methods in incorporating higher-order degree of nodes. By contrast, we find that a distance-based generalization of the core-decomposition for directed networks can profitably be exploited to actually restrict the location of candidate solutions for IM to a single, well-defined portion of a network graph.

4.1 Introduction

Measuring and understanding the spread of "contagion" has attracted tremendous attention as a universal phenomenon that is extensively studied in physical, biological, and social networks. Exemplary application domains are related to social influence, diffusion of information, misinformation or rumors, spread of viruses etc. In this context, a key problem is the identification of the most effective spreaders in a social network. In order to estimate the spreading potential of nodes in a social network, several heuristics have been studied: centrality measures, such as degree or PageRank, or mesoscale-structure-based properties of nodes, such as *core decomposition*. One important study by Kitsak et al. [101] showed that the influential spreaders are those located in the inner-most core of the network, in contrast to the fact that high-degree or high-betweenness nodes could have little effect on the extent of a spreading process. Since then, several studies have been proposed to improve the discriminating ability (i.e., monotonic ranking of spreaders) of the core decomposition (e.g., [7, 71, 143]).

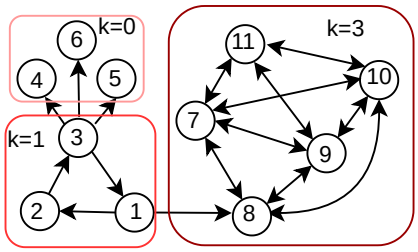


FIGURE 4.1: Core decomposition over a directed network. Cores are determined according to the nodes' *out-degree*.

In this line of research, the network is assumed to be *undirected*, and the empirical findings on the spreading process refer to standard epidemic models (e.g., SIR or SIS).

An alternative line of research corresponds to the widely studied *influence maximization* (IM) [97] problem: given a directed network, a (stochastic) diffusion model, and a budget on the number s of seeds (i.e., early-adopters or initial influencers), IM asks to find a s-sized seed-set S that maximizes the *influence spread* over the network, i.e., the expected number of nodes that are activated, starting from S, at the end of the diffusion process. The main distinction between finding a good seed-set and estimating the spreading potential of nodes in isolation, is that the former problem requires to take into account the cumulative effect of the influence spread. In fact, different nodes may exert influence on largely overlapping portions of the network, so that their cumulative spread would be wrongly estimated by just considering the sum of their spreading potential.

Besides the difference in the network (directed vs. undirected) and in the diffusion models, the difference between these two lines of research is better explained by the next example.

Example 4. Let us consider the example graph in Figure 4.1. Suppose we are required to select one seed of the propagation process (i.e., $s = 1$). It can be noted that node v_1 has a strategic location as it can reach all nodes in the graph. This is clearly an ideal situation for an IM which, depending on the setting of influence probabilities (here omitted for simplicity) and the diffusion model adopted, will likely select v_1 as seed. By contrast, most of the centrality measures will fail at capturing the spreading ability of that node in the network. In fact, none among out-degree, directed closeness and betweenness, and PageRank is able to rank v_1 at the top. Also, considering the outcomes of out-degree-based core decomposition of the example graph, any node in the inner-most core (i.e., core with $k = 3$), would be preferred as seed to any other node with lower core-index, including v_1, despite no nodes in the inner-most core can propagate outwards, thus they cannot be an optimal choice for an IM solution.

Motivated by the above observations, we aim at producing an extensive analysis of where the set of influential nodes extracted by state-of-the-art IM methods are located in a network w.r.t. different notions of graph decomposition. More specifically, we want to understand whether decomposition algorithms can support the identification of subnetworks where nodes have a good influence-spreading potential collectively, rather than as independent individuals. In this regard, our study reveals that a major limitation of classic decomposition algorithms in predicting the influence ability of nodes, is that they are traditionally based on first- or second-order node-degree information, and this may represent a myopic view on the topological properties that would make a node a good spreader. We raise the following research questions:

- In which cores do state-of-the-art algorithms for IM select their seeds in a directed network (e.g., an online social network built upon following relation)?
- How are the seed locations sensitive to any particular graph-decomposition notion?
- What are the internal/external connectivity characteristics that a portion of the network should have to support the most influence potential of their nodes?
- What are the main limitations that may lead a graph decomposition method to fail at determining the regions more densely populated of influential nodes?

Contributions. In this chapter we address the research questions above in a systematic way, through the following main steps. We first review a selection of representative notions of graph decomposition, and adapt their extraction methods in order to enable their applicability to directed networks in influence spread estimation tasks. We then empirically assess the effectiveness of those algorithms when it comes to detecting good spreaders, both as a group of users and individual ones, on a selection of real-world online social networks of different sizes and topological properties.

We evaluate IM algorithms in terms of their respective seed-selection strategies, i.e., how they identify the seeds w.r.t. the considered graph-decomposition methods. Moreover, since allocating seeds inside the inner subnetworks may not be the best choice for IM, we investigate the reasons underlying this contingency.

Finally, we provide evidence that a major limitation that prevents classic decomposition algorithms to find the most influential spreaders, is their inability to incorporate higher-order degree of nodes. Our analysis shows that *distance-based generalization of core decomposition* [17] provides a more informative characterization of how important nodes are in terms of their reachability, thus providing an effective approach to the identification of good spreaders.

4.2 Related work

We present related literature in the analysis of information propagation and influence maximization, as well as the graph-decomposition methods that we will use in this chapter.

Influence propagation. The analysis of social contagion, i.e., the spread of new practices, beliefs, technologies and products through a population, driven by *social influence*, is a very central theme in social sciences, and it has also attracted a lot of interest in the data science community [16]. Such phenomenon develops in two main subjects: the structure of the network and the actions or communications of the users over the network. Researchers have studied the role played by the network topology [200] and by several of its macroscopic characteristics, such as the level of homophily [206] and the modular structure of the network [11, 148], as well as node-level characteristics, such as their centrality, or their capacity of spanning structural holes, thus bridging communities and facilitating, or blocking, the spread of information. Other researchers have considered the social network and the log of past user-activities jointly, and studied important problems such as learning the parameters of the propagation model, i.e., the strength of influence along each edge [73, 168], or how to distinguish real social influence from "homophily" [3, 19, 48, 108]. Finally, a wide literature exists on the analysis of social influence in specific domains: for instance, studying person-to-person recommendation for purchasing books and videos [116, 118], telecommunications services [84], or studying information cascades driven by social influence in Twitter [8, 167].

Fueled by the seminal work by Kempe et al. [97], most of the attention has been devoted to exploiting social influence for "word-of-mouth" driven viral marketing applications: this is the case of the stochastic optimization problem known as influence maximization (IM). Given a social network, where each edge (u, v) is associated with a weight (or probability) $p_{u,v}$ representing the strength of influence that u exerts over v, IM requires to select the set of initial users that maximizes the *expected spread*, i.e., number of users in the social network that gets "infected", according to an assumed underlying *diffusion model*. IM is NP-hard under most standard diffusion models, such as *Independent Cascade* (IC) and *Linear Threshold* (LT) models, however, the simple greedy algorithm provides $(1 - 1/e)$ approximation guarantee, provided that the diffusion model is monotone and submodular (like in the cases of IC and LT). Since the expected spread cannot efficiently be evaluated exactly, most of the effort have been devoted to address this scalability issue by reducing the number of needed Monte Carlo estimations [115]. Alternatively, proxy-based methods have been developed to avoid running Monte Carlo simulations, by estimating the influence spread of the seed set through a reduced diffusion context; although, without ensuring theoretical approximation guarantee.

A significant study that overcomes the efficiency bottleneck of the simulation based methods, while preserving the theoretical approximation guarantee, is proposed in [20], which introduces the *Reverse Influence Sampling* (RIS) framework for IM. The key idea is that the expected spread can be estimated by taking into account a number of pre-computed sketches, i.e., realizations drawn from the distribution induced by influence graph according to the diffusion model. This breakthrough result paved the way for a variety of sketch-based algorithms. Tang et al. in [189] are the first to design a practically efficient solution, TIM/TIM+, whose improvement over RIS consists in keeping the same theoretical complexity as [20] with significantly fewer sketches, bounded by the influence of the unknown optimal set (OPT). More recent RIS-based IMM [188] and SSA [154] algorithms share the common motif of estimating OPT with a fewer number of sketches. IMM improves over TIM/TIM+ through a martingale analysis, while SSA takes an orthogonal perspective, as the number of sketches needed by the algorithm is determined at runtime via an iterative approach. The TIM/TIM+, IMM, and SSA methods will be used in our evaluation (Section 4.4–4.6).

Graph decomposition. *Cores* in a graph were first studied in [172] for characterizing tightly-knit groups in social networks. Since then, core decomposition has been used as a tool for several applications related to the understanding of mesoscale structural characteristics of a network, but also to capture the centrality or influential status of nodes. Among its advantages, core decomposition for an input graph is unique, and hence well-defined, and it can be computed efficiently in linear time w.r.t. the number of edges in the graph.

As mentioned in the Introduction, [101] is one of the earliest studies exploring relations between the spread of influence in undirected networks and core decomposition. The study shows that, under the SIR epidemic model, nodes with the best spreading potential are likely not those with the highest degree or betweenness centrality, but are in the most internal core of the network.

Following the lead of [101], in [145] a similar analysis is carried out in terms of *truss* decomposition [196]. Nodes selected within internal regions of a network according to the truss decomposition, tend to produce infections that are significantly more viral in the early steps of propagation as opposed to the one obtained started from the most internal cores, though this advantage becomes less evident as the propagation approaches to its termination.

The *k-peak* decomposition proposed in [71] aims to find robust decomposition when a network has distinct and independent regions of different edges density. Following the same setting as [101], it is shown that when the initial spreaders are chosen among those with the highest k-peak number, the size of the information cascade may be up to 50% greater than the size based on k-core decomposition. In [7], the *coreness centrality* is defined on top of the classic core decomposition, by aggregating the core-index of all neighbors of a given node. Again under the SIR model and for undirected and unweighted networks, this method has shown to produce better rankings than those based on k-core decomposition.

A recent study [17] extends the k-core decomposition to account for a neighbor-distance threshold h. Differently from [7], the proposed notion of (k, h)-core redefines the coreness property based on a higher-order degree of nodes, i.e., the core-index of a node is function of the number of nodes reachable up to a given distance h.

The notion of k-core decomposition was also extended to probabilistic graphs [18]. The (k, η)-core is defined as a subgraph such that each of its nodes has at least degree k with confidence at least η. Notably, in an IM evaluation scenario, where the edge probabilities are assumed to be influence propagation probabilities, the greedy algorithm could in principle exploit the computation of (k, η)-cores in order to locate the seeds in the inner most η-cores. Another bivariate-core notion is proposed in [65], where the (k, l)-D-core is defined to account for nodes with in-degree at least k and out-degree at least l. The significance of this approach was mainly assessed over collaboration networks where, unlike social influence-driven networks, both the inward and outward connectivities of nodes might be explicitly parametrized.

4.3 Decomposition of directed graphs

In this section, we present the graph decomposition methods examined in our study. One notable point is that, since these methods are *originally conceived for undirected networks* (cf. Section 4.2), we first need to revise their definitions in order to make these methods amenable to support an IM task, which requires a directed network as input context of influence propagation. Also, *our choice of decomposition methods was guided by two main factors*: (i) they are able to scale to large networks; (ii) they can be meaningfully extended to directed networks; and (iii) they are representatives of the most widely used decomposition strategies and variants.

Let $G = \langle V, E \rangle$ be a *directed* graph, with set of nodes V and set of edges $E \subseteq V \times V$. Given any subset $S \subset V$, we denote with $G[S] = \langle S, E[S] \rangle$ the subgraph of G induced by S, where $E[S] = \{(u, v) \mid (u, v) \in E \wedge u, v \in S\}$. Also, for each $v \in V$, $deg^{in}_G(v)$, resp. $deg^{out}_G(v)$, denote the in-degree, resp. out-degree, of v in G.

k-Core decomposition. Given $k \geq 0$, the k-core of a directed graph $G = \langle V, E \rangle$ is the maximal subgraph (denoted as G_{k-core}) corresponding to $G[C_k] = \langle C_k, E[C_k] \rangle$ such that each node $v \in C_k$ has out-degree at least k, i.e., $deg^{out}_{G[C_k]}(v) \geq k$. The degeneracy of the graph, hereinafter denoted as $K^C(G)$, is the highest value of k s.t. $C_k \neq \emptyset$. The core associated with the graph degeneracy is also called the *inner most core*. The *core-index*, or *coreness*, of a node v is the largest k such that $v \in C_k$ and $v \notin C_{k+1}$.

It is easy to show that the well-known $O(|E|)$ algorithm in [13] can straightforwardly be adapted to a directed network: nodes are ordered by increasing out-degree, then nodes u with lowest out-degree are iteratively removed from the graph and each incoming neighbor of u decreases its out-degree, and the process continues until no node remains. The core-index of a node is the out-degree at the moment of its removal.

k-Peak decomposition. It is conceived on top of k-core decomposition, based on the notion of *k-contour*. Given a graph $G = \langle V, E \rangle$, with degeneracy $K{=}K^C(G)$, a k-contour ($k \geq 0$) is the maximal subgraph recursively defined as the k-core of the graph $G \setminus \bigcup_{j=k+1}^{K} G_j$ for all $k < K$, where G_j is the j-contour, and the same as the k-core of G for $k = K$. The *peak-number* of a node is the value k such that the node belongs to the *k-contour*. The peak-degeneracy of the graph, hereinafter denoted as $K^P(G)$, is the highest value of k s.t. there is a non-empty k-contour; it is straightforward to note that $K^P(G) {=} K^C(G)$, for any graph G.

The k-peak decomposition algorithm assigns each node to exactly one contour. Unlike core decomposition, k-peak decomposition does not account for connections starting from outer cores (i.e., lower k cores) towards inner cores of the network. To compute the k-peak decomposition, we iteratively apply our core-decomposition algorithm for directed networks, over the subgraph obtained by removing all the nodes belonging to the inner most core and assigning those nodes the peak number equal to the value of the degeneracy before the removal.

k-Truss decomposition. In our setting, given any three nodes u, v, w, a triangle $\triangle_{uvw}$ is defined as a directed cycle between those nodes. The support $sup(e, G)$ of an edge $e = (u, v) \in E$ in G is defined as $|\triangle_{uvw} : \triangle_{uvw} \in \triangle_G|$, where $\triangle_G$ denotes the set of all triangles in the network. The *k-truss* of G ($k \geq 2$), denoted by T_k, is the largest subgraph of G such that $\forall e \in E_{T_k}$, $sup(e, T_k) \geq (k-2)$. The *truss-index* of an edge is the largest k-truss it belongs to.

Once the support of each edge is computed, we apply the algorithm proposed in [196] to obtain the decomposition. However, since the k-truss decomposition is defined with respect to the edges of the graph, we eventually assign a score to each node that is equal to the average truss-index of the node's outgoing edges. Also, we denote with $K^T(G)$ the highest of the node truss-indexes.

Neighbor-coreness aggregation. Adapting from [7], each node v is assigned with a *neighbor-coreness* score given by $C_{nc}(v) = \sum_{u \in N^{out}(v)} c(u)$, where $c(u)$ denotes the core-index assigned to node u and $N^{out}(v)$ is the set of v's out-neighbors. We also denote with $K^{NC}(G)$ the maximum neighbor-coreness score.

The algorithm for computing this score function extends the one used for directed k-core: once computed the core-indexes, we apply the function $C_{nc}(\cdot)$ to account for the out-neighbors' contribution, for every node in the network.

Distance-based generalization of core decomposition. Given $v \in V$, a subset $S \subseteq V$, and a neighbor-distance threshold $h > 0$, the h-neighborhood of v w.r.t. the subgraph $G[S]$ is $N_{G[S]}(v, h) = \{u \in S | u \neq v \wedge d_{G[S]}(v, u) \leq h\}$, where $d_{G[S]}(v, u)$ denotes the shortest path distance *from v to u* in the subgraph of G induced by S. The *h-outdegree* of a node w.r.t. S is defined as $deg^h_{G[S]} = |N_{G[S]}(v, h)|$. Given $k \geq 0$, a *(k, h)-core* represents the maximal subgraph $G[C_k] = (C_k, E[C_k])$ such that every node $v \in C_k$ has h-outdegree at least k, i.e., $deg^h_{G[C_k]}(v) \geq k$. Also, for any given h, the distance-generalized degeneracy, $K^{DGC}_h(G)$, is the maximum k such that $C_k \neq \emptyset$.

To compute the (h, k)-cores, we adapted Algorithm 1 in [17] by specializing the notion of h-neighborhood for out-neighbors.

Example 5. Let us consider again the example shown in Figure 4.1, to check whether the various graph-decomposition algorithms are able to assign v_1 with the highest score. We have already observed that this is not the case when using the k-core decomposition (cf. Example 4). Similar outcome holds also for the k-peak decomposition — two distinct contours are found, with v_1 having peak-number 0 along with nodes $v_2, \ldots, v_6$, and the remaining nodes with peak-number 3 — the k-truss decomposition and the neighbor-coreness aggregation method — which assign the highest

TABLE 4.1: Summary of evaluation network data.

network	#nodes	#edges	avg. in-deg.	avg. path len.	dens.	diam.	#sources	#sinks
DBLP - DB	317K	1M	3.31	7.89	$1\,05e^{-5}$	31	127K	12K
Epinions - Ep	116K	722K	6.2	4.79	$5\,3e^{-5}$	16	28K	43K
Nethept - Net	15K	62K	4.1	5.83	$2.7e^{-4}$	5	0	0
Twitter - Tw	21K	227K	10.38	6.28	$4.7e^{-4}$	32	3K	3K
Instagram - Ig	17K	617K	35.25	4.24	$2e^{-3}$	15	0	0
FriendFeed - FF	493K	19M	38.85	3.82	$7.8e^{-5}$	32	42K	292K

score to nodes $v_7, \ldots, v_{11}$. By contrast, the distance generalized core decomposition is able to detect, for $h = 2$, three cores, where the inner-most one does contain node v_1 (along with $v_7, \ldots, v_{11}$).

4.4 Evaluation methodology

We used 6 real-world online social network datasets, whose properties are summarized in Table 4.1. Our choice of these network data is justified as they can be regarded as benchmarks in IM or graph-decomposition studies. In particular, *Epinions, DBLP, Nethept* networks were used in the original works proposing the three IM methods under examination (i.e., TIM/TIM+, IMM, and SSA); the *Twitter* dataset was used in [18] to assess the significance of uncertain graph decomposition for IM; *Instagram* and *FriendFeed* were studied in [26] for targeted IM in a user engagement context.

We considered the two most commonly used diffusion models in IM, namely Independent Cascade (IC) and Linear Theshold (LT) models [97]. The results presented in the remainder of this chapter are only based on IC. The experimental results using LT — which can be found in the Appendix B.1 — are consistent with the findings for IC, reported in this chapter.

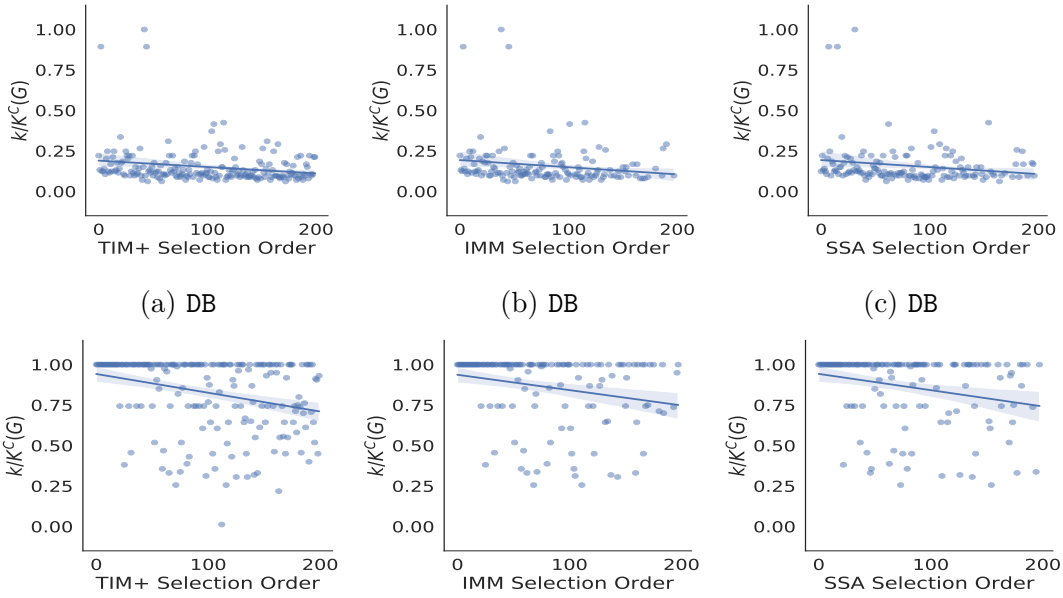
IC considers each node can be activated by each of its incoming neighbors independently. Based on the influence probabilities $p_{u,v}$ for each edge (u, v), and given a seed set S at time step 0, any diffusion instance of the IC model unfolds in discrete steps. Each active node u at step t will attempt to activate, with probability $p_{u,v}$ each of its outgoing neighbors v that is inactive at step t-1. Note that u has only one chance to activate its outgoing neighbors. If the attempt is successful, v becomes active at step $t+1$, otherwise v stays inactive. The diffusion instance terminates when no more nodes can be activated. For specifying the influence probability of the edges we adopt a widely-used strategy: each edge (u, v) is associated with a probability $1/deg^{in}(v)$, where $deg^{in}(v)$ is the number of in-neighbors of v.

The main goal of the experimental evaluation is to characterize the *coreness* of those nodes considered to have a strong spreading potential. More specifically, we want to investigate the capability of each graph-decomposition algorithm to locate the most influential nodes within its inner-most regions.

Results are organized into two main sections: first, we focus on those methods that rely only on first-order node-degree information (Section 4.5), then we evaluate the impact of the adoption of a distance-aware generalization of the core-decomposition (Section 4.6).

4.5 Degree-based cores

We investigate where the most influential nodes selected by state-of-the-art IM algorithms — TIM/TIM+ [189], IMM [188], and SSA [154] (cf. Section 4.2) — are



(a) DB (b) DB (c) DB

(d) FF (e) FF (f) FF

FIGURE 4.2: Normalized core-index ($k/K^C(G)$) of the first 200 seeds computed by (a,d) TIM+, (b,e) IMM, and (c,f) SSA, under the IC model.

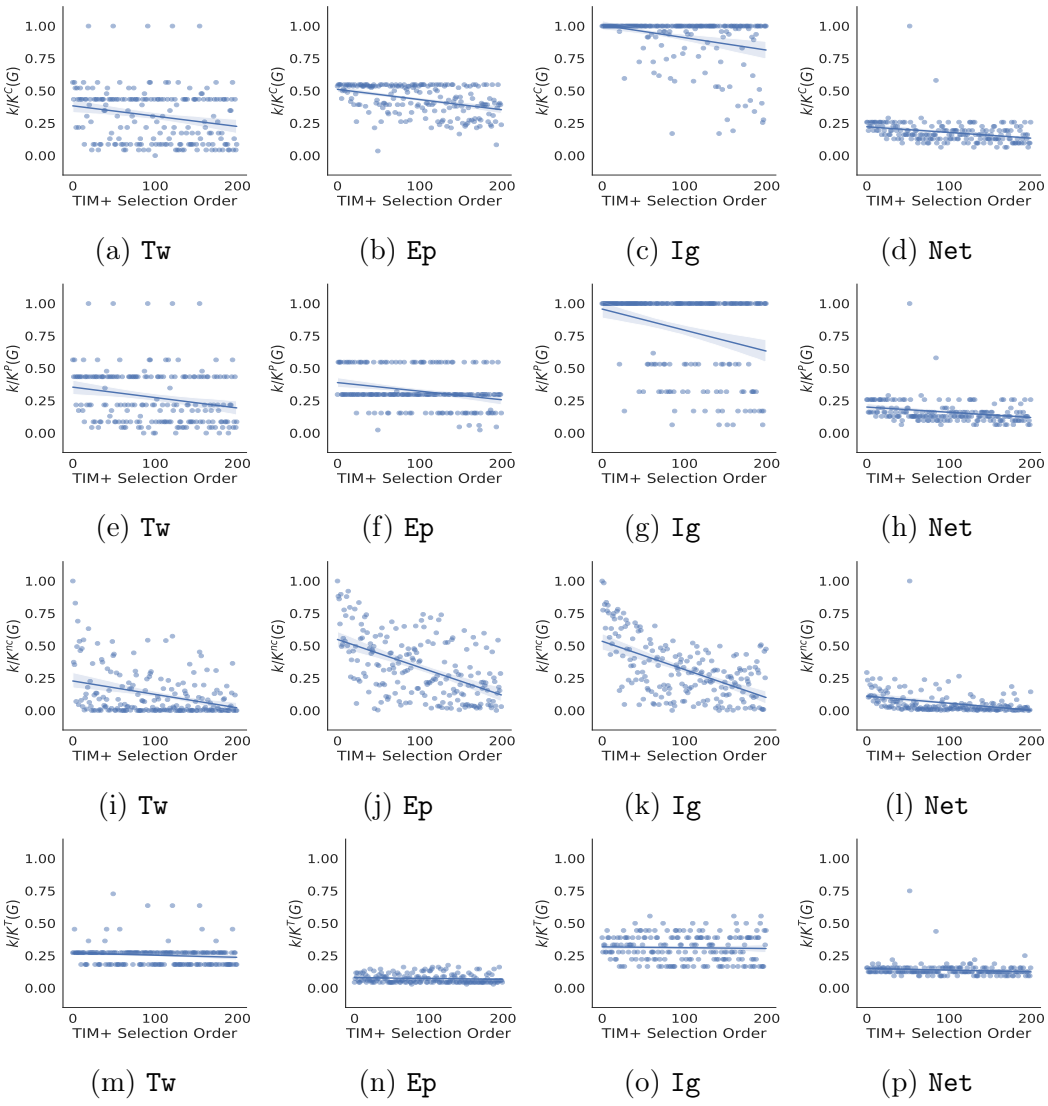
located in the network w.r.t. different graph-decomposition methods (Section 4.5.1). Prompted by the results obtained in this early step of evaluation, we will delve into the features that could be used as proxies for identifying a "good" subnetwork for locating IM-(near)optimal influential spreaders (Section 4.5.2).

4.5.1 Seed selection order

To begin with, we analyzed the selection order of seeds discovered by each IM method, under the IC model, in relation to their core index as produced by the classic core decomposition.

Figure 4.2 reports on the y-axis the normalized core index (i.e., the core index of the node divided by the degeneracy of graph) for the first 200 seeds — computed by TIM+, IMM, and SSA, respectively — ordered on the x-axis according to their selection order, i.e., the iteration corresponding to the insertion of a node into the seed set. For this analysis, we report only results corresponding to two datasets; nonetheless, these results are representative of a general scenario encompassing all remaining networks. We refer the reader to the Appendix B.1 associated with this chapter.

One remark that stands out is that the three IM methods exhibit a very consistent behavior, which seems to depend mostly on the network. This is not really surprising, since all such algorithms share the state-of-the-art RIS-based approach in their algorithmic scheme (cf. Section 4.2). While on dataset DB most of the seeds, with few notable exceptions among the first seeds, are in peripheral cores (the majority of the seeds have core-index between the 5-th and 25-th percentiles), for FF the situation is slightly different: many seeds are selected in high cores, although a good portion of seeds are identified in lower cores. What is common to both datasets (and to the others not reported in Figure 4.2) is that, as hinted by the regression line in each plot, as the selection progresses the various IM methods are more likely to identify



(a) Tw (b) Ep (c) Ig (d) Net

(e) Tw (f) Ep (g) Ig (h) Net

(i) Tw (j) Ep (k) Ig (l) Net

(m) Tw (n) Ep (o) Ig (p) Net

FIGURE 4.3: From top to bottom, normalized core-index ($k/K^C(G)$), peak-number ($k/K^P(G)$), neighbor-coreness ($k/K^{NC}(G)$), and truss-index ($k/K^T(G)$) of the first 200 seeds computed by TIM+, under the IC model.

the seeds among those with lower core index, i.e., in the periphery of the network. This can be explained with the fact that our evaluation methods work under the IC and LT diffusion models, whose activation functions are monotone and submodular: after the earlier stages of seed selection, the IM methods would start exploring the periphery of the network, since therein it will likely reside the nodes whose marginal gain is potentially less affected by the earliest selections.

This is in contrast with the findings in [101, 145], according to which the most influential nodes should reside in the inner-most core of the network. This difference is due to the fact that those works consider a SIR propagation model, whereas we use IC/LT, and on the fact that they focus on the spreading potential of nodes in isolation, while our analysis considers the cumulative expected spread of the seed set of the IM problem.

Results drawn from the previous analysis were confirmed by analogous evaluation extended to the other graph decomposition methods and networks. As shown in Figure 4.3, in most networks (e.g., Tw, Ep, Net), the majority of the seeds are located

in subnetworks that correspond to mid/low values of each particular decomposition method. One exception is represented by `Ig`, where most seeds are located in the inner subnetworks, provided that k-core or k-peak decomposition is used. Among the various decomposition techniques, it can be noted that neighbor-coreness provides high-variance, hence poorly meaningful results for our analysis. This is explained since neighbor-coreness was originally conceived as a proxy solution for ranking nodes w.r.t. their individual influence, rather than for achieving coarser-grain graph-decompositions; this also prompted us to ignore it in the remainder of our study. Another interesting remark regards the k-truss decomposition. In fact, identifying the seeds within the inner-most subnetworks induced by this method appears to be a disadvantageous choice for our purposes, as most of the seeds are located within the outer subnetworks (i.e., those containing nodes with lower truss-index values).

The above results, coupled with the ones discussed in the previous section (Section 4.5.1), provide evidence that allocating seeds in the inner-most regions of a network may turn out to be a poorly effective strategy for IM. This contingency may be ascribed to the fact that concentrating the selection of nodes within the same subnetwork induced by a graph-decomposition technique, would prevent us to exploit the submodularity of the activation function of the IM methods. Intuitively, it may happen that the propagation remains trapped inside the densest regions of a network, and consequently it will not be able to involve other parts of the network; this particularly holds for the k-truss decomposition, which considers the number of triangles a particular node is involved in.

Notably, our findings totally fit the LT model as well (results corresponding to LT can be found in the Appendix B.1).

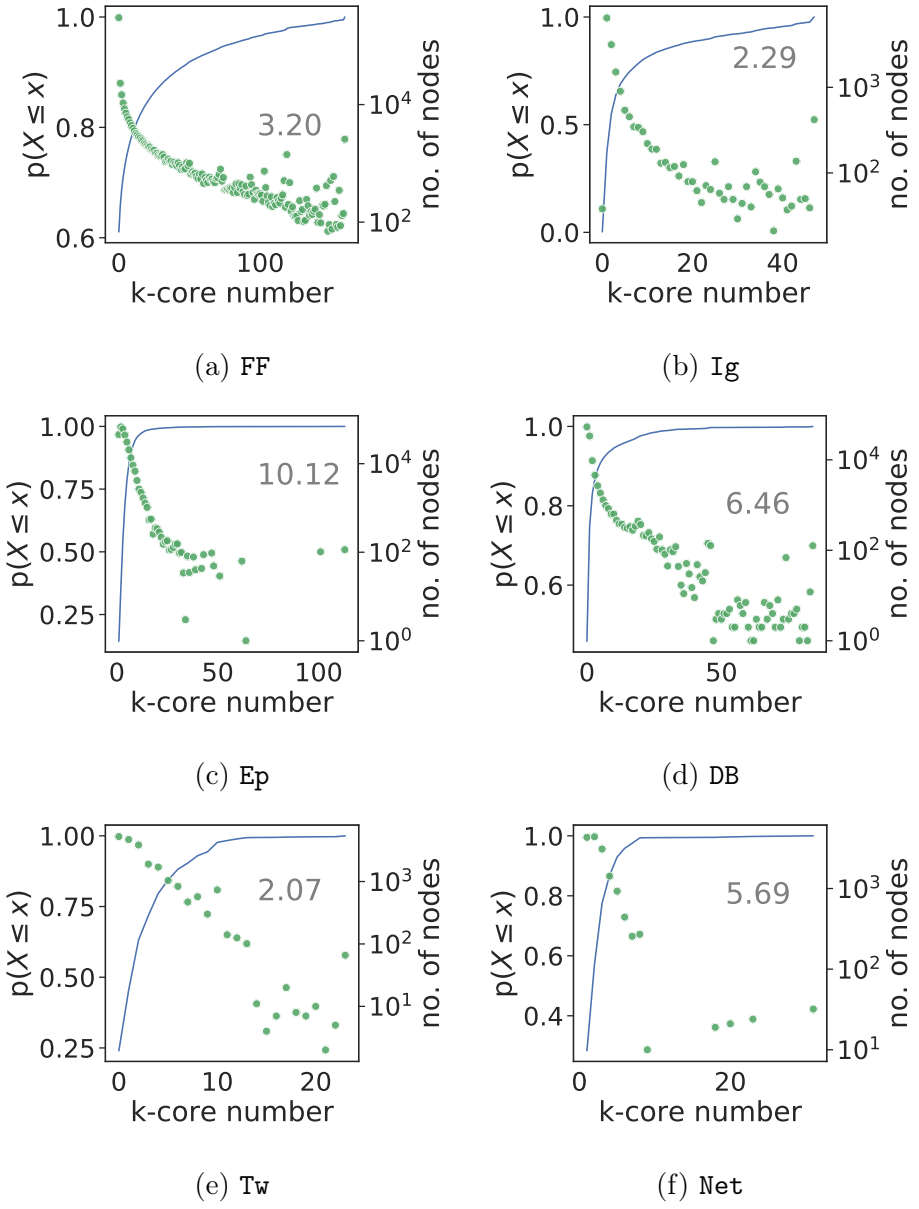
4.5.2 Characterization of the cores/contours

Based on the results obtained so far, we can recognize two main groups in the evaluation data: the one corresponding to `FF` and `Ig`, and the other one including all the remaining networks, where influential spreaders were found to be located in the "outer" portions of the network, as opposed to the former group.

Hereinafter, we restrict our attention to the k-core and k-peak decomposition, since they turned out to be the most promising and reliable ones to support our next analysis aiming at understanding how to estimate the nodes' influence-spread potential. Thus, we will devote our attention to two main aspects: (i) how nodes are distributed within the different cores/contours of the network, and (ii) how the cores/contours are connected to each other.

Core/Contour distribution. Figure 4.4 shows how nodes are distributed over the different cores of the network. If we compare these results in light of the previous findings (Section 4.5.1), we observe that a lower skewness in the distribution would correspond to the identification of seeds within the inner cores. In fact, the distributions for `FF` and `Ig`, exhibit a much lower skewness (i.e., 3.2 and 2.3, resp.) as compared to the one corresponding to the other networks, however with the exception of `Tw`. Albeit the skewness could serve as a moderately good indicator of how effective it will be to allocate seeds within the inner cores of a network, we need to further investigate the characteristics of the cores.

As regards the k-peak decomposition (results shown in the Appendix B.1), we observe it tends to favor skewer distributions than the core-decomposition ones. In particular, although $K^C(G) = K^P(G)$, the number of distinct contours in the evaluation networks is found to be consistently smaller than the number of distinct cores in the network. This implies that the k-peak decomposition may provide a coarser view



(a) FF (b) Ig

(c) Ep (d) DB

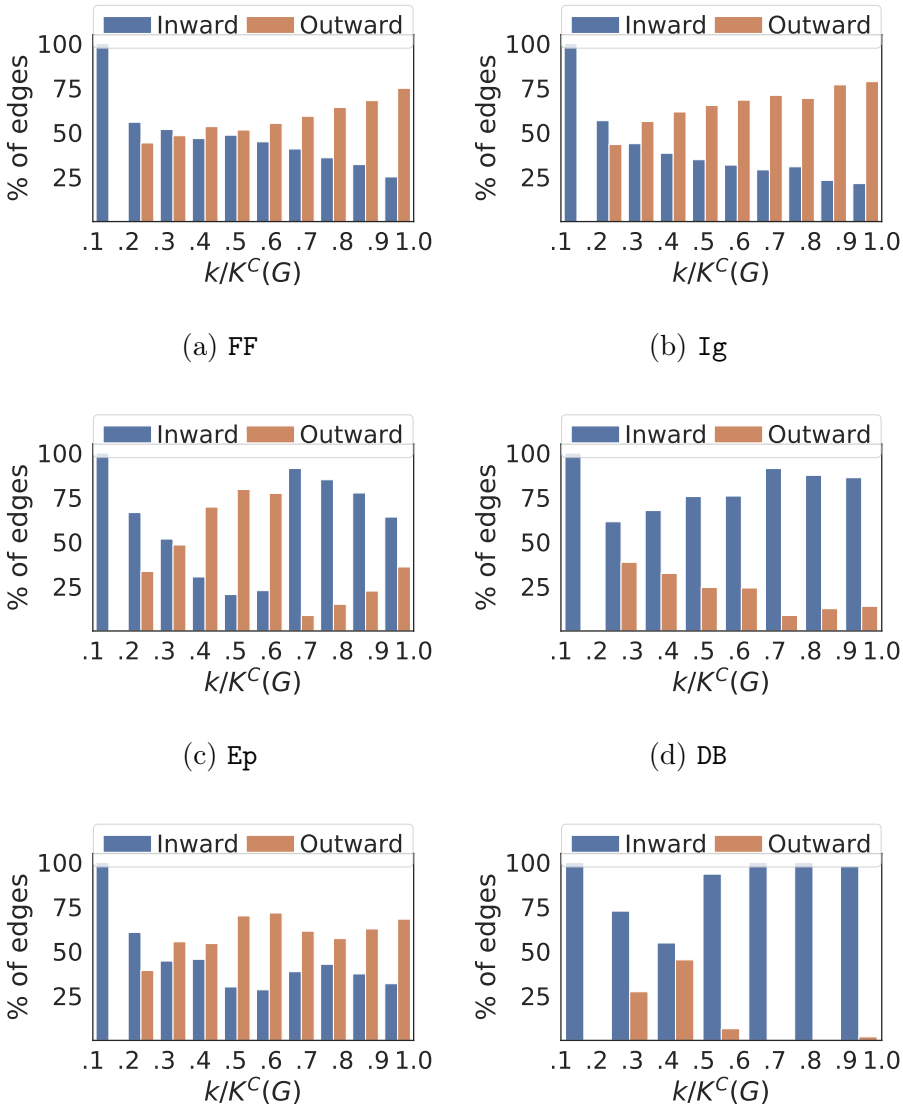
(e) Tw (f) Net

FIGURE 4.4: Distribution of nodes over the cores of the network. Each plot shows, for every core-index k (x-axis), the number of nodes with core-index at most k on the rightmost y-axis, and the cumulative distribution of core-index on the leftmost y-axis. Also, the skewness of the distribution is reported inside each plot.

on a network structure, where most nodes are concentrated in the subnetworks with lower peak number, thus hindering the ability of this technique to discriminate the influence-spread potential of nodes.

Core/Contour connectivity. Here we focus on the connectivity from a core/contour perspective. More specifically, we categorized edges into two separate classes, namely: **outward** edges, if the source node has a core-index/peak-number equal to or greater than the target node, and **inward** edges otherwise.

Figure 4.5 shows the fraction of edge-set that belongs to each of the two classes, based on core-indexes of their sources — very similar behaviors were also found in results corresponding to peak-numbers (shown in Appendix B.1). We recognize three



(a) FF (b) Ig

(c) Ep (d) DB

(e) Tw (f) Net

FIGURE 4.5: Percentage of inward and outward edges vs. normalized core-index $k/K^C(G)$. The i-th percentage bar ($i = 1..9$) corresponds to edges such that the source node has normalized core-index in $(x_i, x_{i+1}]$, upon a segmentation of the x-axis values into ten intervals $(x_1, x_2], \ldots, (x_9, x_{10}]$.

types of characteristics in the inward percentage-bars, as the normalized core-index increases: (i) a roughly decreasing trend, for FF and Ig, (ii) a roughly constant trend, for DB, and (iii) a roughly bimodal decreasing trend, for the remaining networks. For the former group, while the inward percentage remains much higher than the outward one until mid-high regimes in the x-axis, this gap tends to become small for the highest cores, showing that nodes in the inner-most core (i.e., rightmost side of a plot) also have a good connectivity towards the periphery of the network. Quite differently from FF and Ig, Ep and Tw show a roughly bimodal decreasing behavior, which appears to have a break-point around half of the degeneracy. Notably, this corresponds to the core where most of the seeds are actually found according to the results discussed earlier in this section (Section 4.5.1). However, while the second decreasing trend

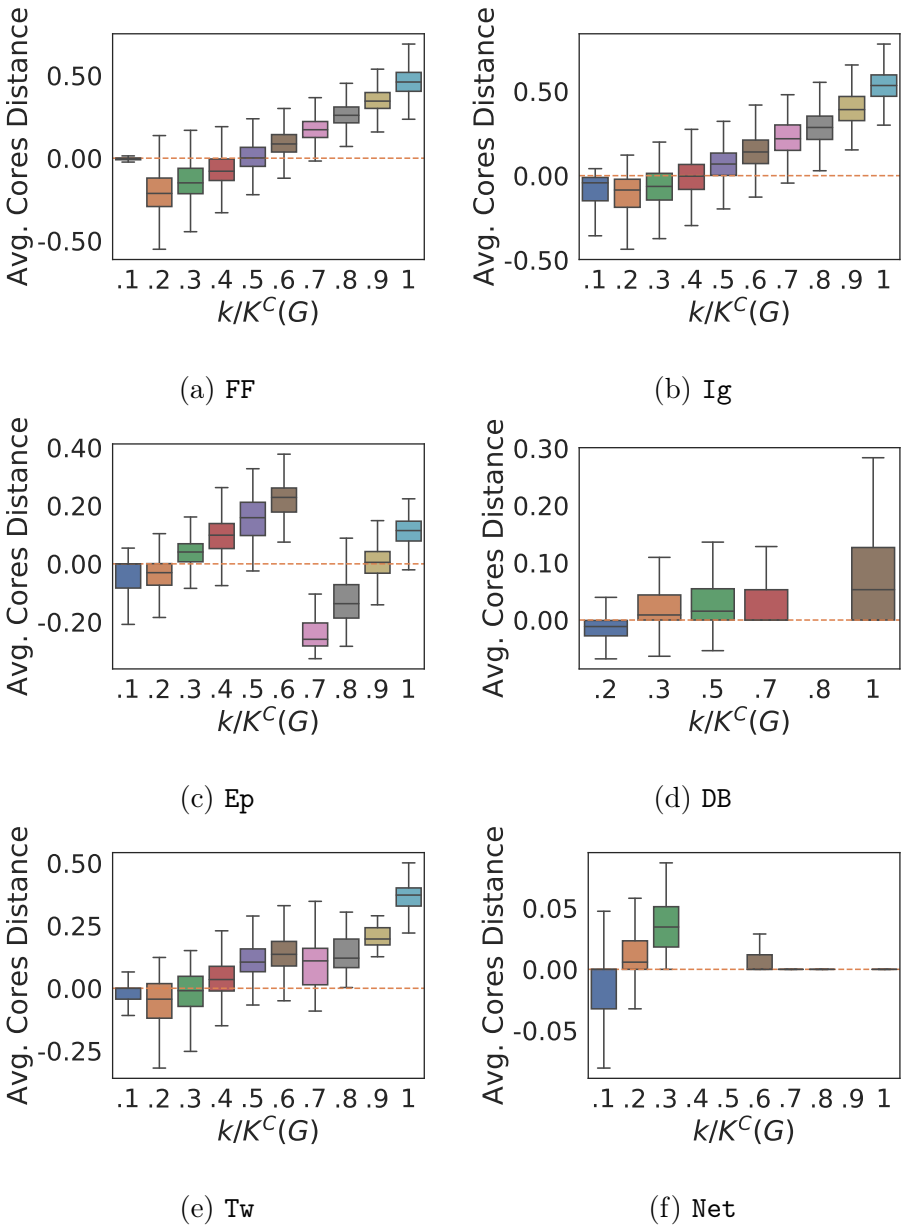


FIGURE 4.6: Distribution of the node's average normalized core distance vs. normalized core-index $k/K^C(G)$. For each core-index k, the corresponding boxplot represents the distribution of the average normalized core distances computed for each node having core-index k.

ends up with a 60% inward edges for the first and second inner-most cores in Ep, a further interesting scenario occurs in Tw. Here, the nodes within the inner-most core are mostly connected to each other, since a considerably high fraction of edges (above 80%) are inward. In DB, the inward edges are the large majority, regardless of the core-indexes of their nodes, which might be ascribed to a relatively high percentage of source nodes.

Pairwise core distances. We consider here a more robust measure than the inward/outward property of edges, which accounts for the difference of core-index values of two linked nodes. Given any edge (u, v), we define the *pairwise normalized core distance* as $dist(u, v) = (k_u - k_v)/K^C(G)$, with k_u and k_v the core-index assigned to u and v, respectively. Upon this, for each node u we compute the average normalized

core distance over its out-neighbors. A positive value means that u is mostly connected with nodes belonging to outer cores, and the greater the value, the more u's out-neighbors can be considered as peripheral w.r.t. the u's location.

Figure 4.6 shows the boxplot distributions of average normalized core distance w.r.t. the normalized core-index values. The analysis of such plots allows us to integrate and enrich the results observed in Figure 4.5. Considering first `Ig` and `FF`, where most of the seeds have the maximum core-index (Figsures 4.2–4.3), we observe a clearly increasing trend of the nodes' average normalized core distance. With corresponding boxplot median around 0.5, nodes within the highest core-index show to be well connected with nodes located in mid-level outer cores. A different situation is observed on `Tw`, `Ep`, and `DB`, where the maximum average normalized core distance mostly remains below 0.4, 0.3, and 0.1, respectively. Remarkably, in `Ep` (Figure 4.6(c)), where most seeds have mid/low core-index (Figure 4.3), we observe again a breakpoint in the distribution around half of the degeneracy, where the peak of average normalized core occurs, while the second increasing trend almost remains below positive values in the y-axis, with the inner-most boxplot having very low median (around 0.1). Also, on `DB` (Figure 4.6(d)), the values of range of each boxplot (always below 0.1) indicate that the edges tend to connect nodes that have very close core-index, which is also consistent with the fact that nearly all seeds are not located within the inner-most core (Figure 4.3).

4.5.3 Discussion

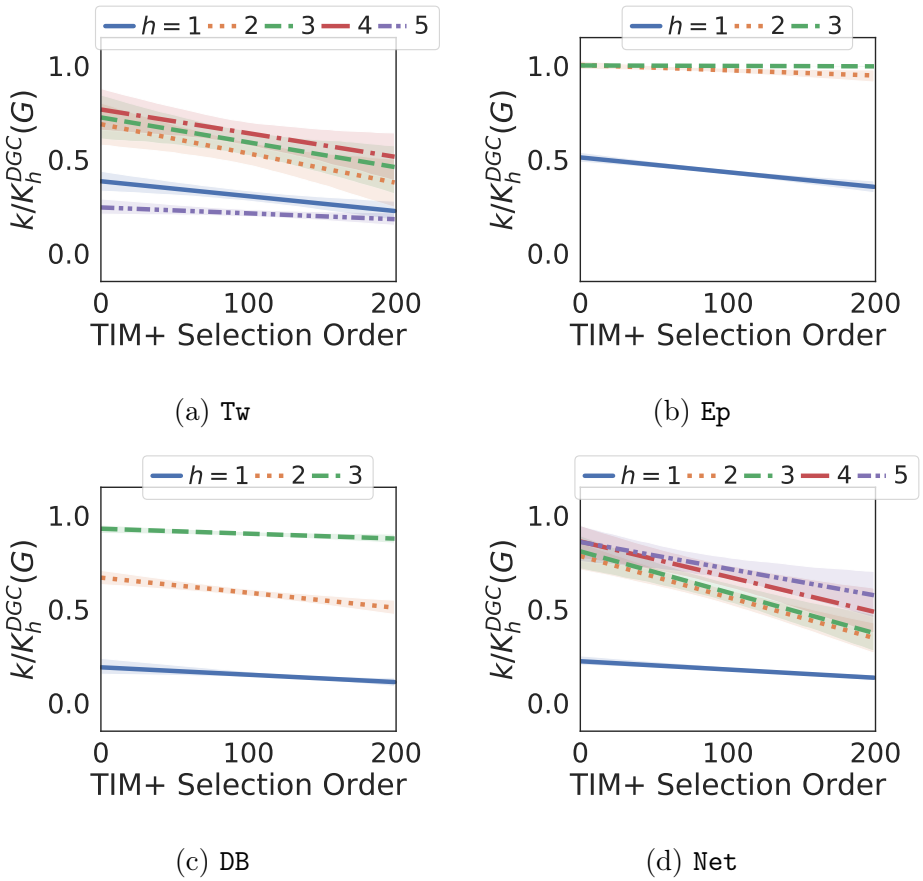
In this first stage of evaluation, we have learned that searching for influential spreaders within the inner subnetworks (based on any particular decomposition method) does not ensure to find the best seeds for an IM problem. Indeed, it should not be surprising that topological properties of the networks take a crucial role in determining whether or not nodes in the inner-most cores have the best influence-spreading potential. In fact, `Ig` and `FF`, where most of the seeds were found in the inner-most core, are also the networks that exhibit a significantly higher average in-degree and a network density that is slightly higher than the other networks (Table 4.1). The remaining networks, where seeds were mostly identified outside the inner-most cores, show a substantially sparser structure, as indicated by their values of average path length, density, and diameter.

We also found out that, when nodes in the inner subnetworks are mostly connected with each other rather than towards nodes in outer subnetworks, IM methods tend to select seeds among the set of nodes that couple a mid/low core-index with good connectivity towards the inner subnetworks. We conjecture that a major limitation of the decomposition methods considered so far, relies on their inability to leverage higher-order degree of nodes. The next stage of evaluation is conceived around this argument.

4.6 Higher-order cores

This section is dedicated to the evaluation of the only existing decomposition algorithm based on higher-order degree, i.e., (k, h)-core decomposition.

Results are organized into three parts. In the first part, we replicate the same setting adopted in the early step of the previous evaluation (cf. Section 4.5), in order to assess the relation of (k, h)-core decomposition with the outcomes of an IM algorithm (Section 4.6.1). Next, we assess the sensitivity of the decomposition w.r.t. the value of the neighbor-distance threshold h (Section 4.6.2). Finally, we also investigate the



(a) Tw (b) Ep

(c) DB (d) Net

FIGURE 4.7: Linear regression of the normalized distance-generalized core-index ($k/K_h^{DGC}(G)$) of the first 200 seeds computed by TIM+, under the IC model.

individual influential-spreading potential of nodes, and put this in relation with the decomposition outcomes (Section 4.6.4).

Please note that we shall focus our analysis on those networks where, by using all the previously analyzed graph-decomposition methods, the seeds were mostly identified outside the respective inner-most cores.

4.6.1 Seed selection order

Analogously to the analysis presented in the first phase of Stage 1 (cf. Section 4.5.1), we first investigated the relations between the (k,h)-core-index values and the selection order of the discovered seeds.

Looking at the plots in Figure 4.7, it stands out that *a significant fraction of seeds is now found to be located in the inner-most (k,h)-core(s)*. This is particularly evident in Ep and DB, where all top-200 seeds (i.e., not only the early-selected ones corresponding to a small budget s) are in the inner-most core or immediately outer one, with $h \in \{2,3\}$ and $h = 3$, respectively. A further important finding is that while regression lines tend to rise up for higher h, with major gain from $h = 1$ (i.e., equivalent to core decomposition) to $h = 2$, this trend is not monotone in general. Indeed, it may happen that an overly high value of h (typically higher than 4) could lead to decreased performance, even worse than the corresponding core decomposition (as observed for Tw, where the regression line for $h = 5$ lays on about 0.25).

TABLE 4.2: Maximum (k,h)-core-index (leftmost) and number of distinct (k,h)-cores (rightmost), for varying h.

	$h=1$	$h=2$	$h=3$
DB	113 / 47	343 / 234	2135 / 1957
Ep	85 / 85	909 / 902	5357 / 5053
Net	31 / 13	69 / 69	389 / 384
Tw	24 / 24	270 / 270	1349 / 1250

4.6.2 Sensitivity to h

Here we delve into the characteristics of the (k,h)-cores detected by differently setting h. In particular, we want to understand how nodes are distributed within the different (k,h)-cores of the network, by varying h.

First, as reported in Table 4.2, we observe that the number of cores and the maximum core-index grow significantly as h increases — recall that $h=1$ corresponds to the classic k-core decomposition — which suggests how the (k,h)-core decomposition can enable a fine-grain micro/mesoscale structure analysis.

In Figure 4.8, we observe that, when $h>1$, the number of nodes in the subnetworks with lower (k,h)-core-index is significantly smaller than for $h=1$. This is clearly due since nodes tend to be more connected to each other as h increases. More interestingly, the inner-most generalized cores (i.e., tail of the distributions) are consistently more populated than for $h=1$. Nonetheless, as displayed in the insets of Figure 4.8 for all networks, the inner-most generalized core covers a fraction of the whole node-set that is relatively small, yet meaningful for a seed-set selection task.

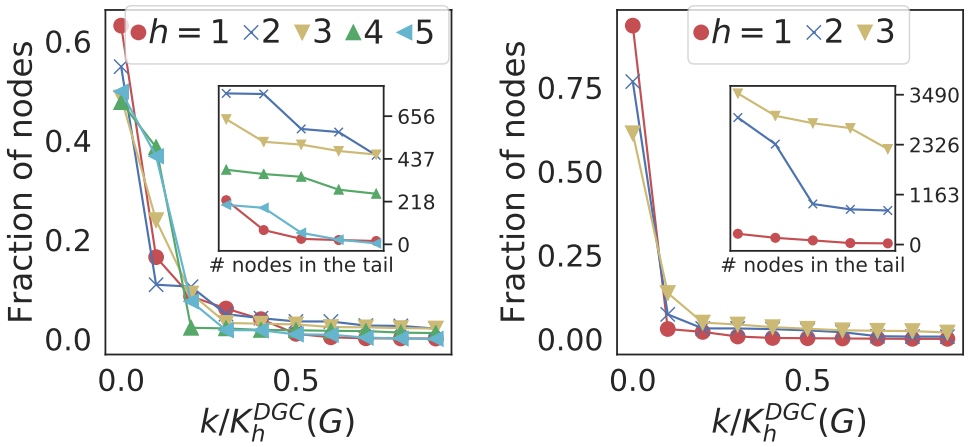
4.6.3 Discussion

We have unveiled that the best-influential spreaders can actually be located within one or very few inner-most core(s) of a network provided that a higher-order graph-decomposition method is used. The neighbor-distance threshold (i.e., h) plays a key role in the decomposition, since too large values of the parameter may in principle lead, at the cost of increased computational overhead, to few cores covering most nodes in the network, thus reducing the benefits of solving the identification of seeds within a small subnetwork; this would mostly happen when the chosen h approaches the average path length of the network, therefore more nodes fall into the same cores. However, in practice, $h \in \{2,4\}$ turned out to be the most effective choice to concentrate the identification of a relatively large seed-set within the inner-most generalized core. As one rule-of-thumb, a proper setting h is the one leading to observe the tail in the distribution of generalized core-index as corresponding to a fraction of nodes comparable with the budget for the seed-set to be discovered. Nonetheless, it emerges an interesting opportunity for a theoretical investigation of relations between h and structural characteristics of the network, which we leave as future work.

4.6.4 Individual influence-spreading ability

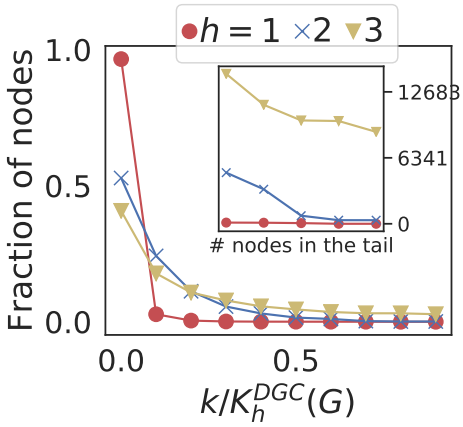
The above findings prompted us to further investigate whether the nodes assigned to the inner-most core by the distance-generalized core decomposition have also *individual spreading* ability. More specifically, we want to determine the nodes' individual influential-spreading potential, i.e., the spread of each node as a singleton seed-set, estimated through Monte Carlo simulation with 10 000 runs.

Figure 4.9 shows that a high (h,k)-core-index is in general a more reliable indicator of the influence a node can individually produce. In fact, in many cases, nodes having

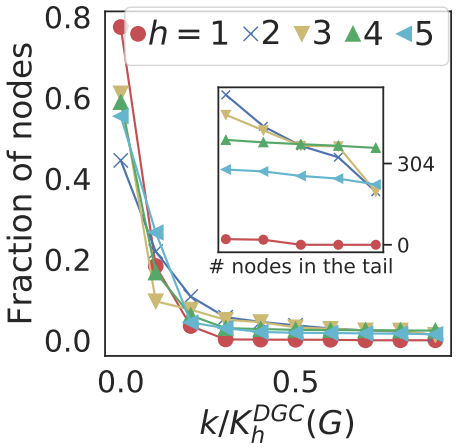


(a) Tw

(b) Ep



(c) DB

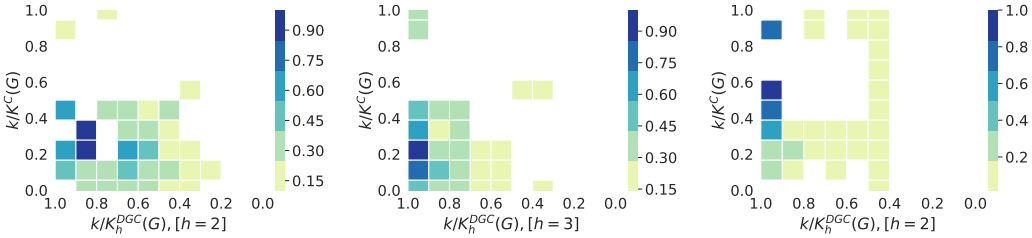


(d) Net

FIGURE 4.8: Fraction of nodes per normalized distance-generalized core-index $k/K_h^{DGC}(G)$, for varying h. Insets zoom in the tail of each distribution, showing the exact number of nodes in the last quartile of $k/K_h^{DGC}(G)$.

higher (h,k)-core-index exhibit higher influence potential. By contrast, such nodes are not necessarily those with the highest core-index according to k-core decomposition. Also, it should be noted the inner cores detected by k-core decomposition ($h = 1$) are very different from the ones corresponding to higher values of h. In fact, many nodes with low/mid core-index turn out to have a very high (h,k)-core-index.

To sum up, for an appropriate value of h, nodes in the inner-most cores are always the ones having the highest influential-spread potential, either as singletons and as groups (Section 4.6.1). This outstanding result clearly highlights the opportunity of exploiting a distance-aware core decomposition for effectively solving top-influencer identification problems that, while not being necessarily under the IM framework, would avoid trapping into an under/over estimation of cumulative spread of a set of nodes that is a typical of any top-s search centrality-based heuristic approach.



(a) DB (b) DB (c) Ep

FIGURE 4.9: Average spread of nodes w.r.t. selected combinations of k-core-index (y-axis) and (h,k)-core-index for a particular choice of h (x-axis). The expected spread of each node is computed by considering the node as a singleton seed-set. Darker colors correspond to higher normalized spread.

4.7 Chapter notes

In this we assessed for the first time the opportunity of leveraging on graph-decomposition methods to simplify the problem of identification of the most influential spreaders in directed network, under an influence maximization framework. We initially found out that the correlation between the influential spreading power and the indexing of nodes according to several graph-decomposition methods, is weaker than expected, as we demonstrated that state-of-the-art IM algorithms do not generally locate their seeds in the inner-most regions of a network, especially in networks with a sparse structure. We showed that one major flaw of any of the classic decomposition algorithm is related to the inability of integrating a notion of higher-order degree into the decomposition scheme. By contrast, we found out that leveraging on a distance-generalized core decomposition enables the desired outcome of detecting the most influential spreaders in the inner-most generalized-core portion of the network.

This work opens several paths of further investigation. Our empirical assessment of the relation between influence spread and different notions of graph-decomposition paves the way to the opportunity of embedding advanced, distance-based generalized decomposition methods in an IM-based influence analysis framework, with the purpose of narrowing the search space of the best seeds only to specific portions of the network, without even estimating in advance the influence probabilities. A related research direction concerns the challenge of understanding what are the theoretical properties underlying the relations between the neighbor-distance threshold h in the generalized core decomposition method, and the structural characteristics of the input network, in order to determine the minimum value of h that implies the detection of the most influential nodes within the inner-most generalized core.

Chapter 5

Topology-based Diversity-sensitive Targeted Influence Maximization

Research on influence maximization has often to cope with marketing needs relating to the propagation of information towards specific users. However, little attention has been paid to the fact that the success of an information diffusion campaign might depend not only on the number of the initial influencers to be detected but also on their *diversity* w.r.t. the target of the campaign. Our main hypothesis is that if we learn seeds that are not only capable of influencing but also are linked to more diverse (groups of) users, then the influence triggers will be diversified as well, and hence the target users will get higher chance of being engaged. Upon this intuition, we define a novel problem, named *Diversity-sensitive Targeted Influence Maximization (DTIM)*, which assumes to model user diversity by exploiting only topological information within a social graph. To the best of our knowledge, we are the first to bring the concept of topology-driven diversity into targeted IM problems, for which we define two alternative definitions. Accordingly, we propose approximate solutions of DTIM, which detect a size-k set of users that maximizes the diversity-sensitive capital objective function, for a given selection of target users. We evaluate our DTIM methods on a special case of user engagement in online social networks, which concerns users who are not actively involved in the community life. Experimental evaluation on real networks has demonstrated the meaningfulness of our approach, also highlighting the usefulness of further development of solutions for DTIM applications.

5.1 Introduction

Online social networks (OSNs) are nowadays the preferred communication means for spreading information, generating and sharing knowledge. One central problem is the identification of influential individuals in an OSN such that, starting with them, one can trigger a chain reaction of influence driven by "word-of-mouth", which allows for reaching a large portion of the network with a relatively little effort in terms of initial investment (budget). This is commonly referred to as *viral marketing* principle, which is the underlying motivation for a classic optimization problem in OSNs, namely *influence maximization* (IM). The general objective of an IM method is to find a set of initial influencers which can maximize the spread of information through the network (e.g., [62, 72, 97, 121, 189, 211]).

Most of existing works in IM and related applications focus on the entire social network through which the spread of influence is to be maximized. However, thinking in terms of viral marketing, an organization often wants to narrow the advertisement of its products to users having certain needs or preferences, as opposed to targeting the whole crowd. Also, in an OSN scenario, some events or memes would be of interest

only to users with certain tastes or social profiles. Our work fits into research on this problem, hereinafter referred to as *targeted IM*.

Leveraging diversity for enhanced IM. While maximizing the advertising of a product, an organization also needs to minimize the incentives offered to those users who will reach out the target ones. This obviously raises the necessity of choosing a proper number k of seed users (i.e., initial influencers) to be detected, which corresponds to the budget constraint. Surprisingly, an important aspect that is often overlooked is that the success of a viral marketing process might depend not only on the size of the seed set but also on the *diversity* that is reflected within, or in relation to, the seed set. Intuitively, individuals that differ from each other in terms of kind (e.g., age, gender), socio-cultural aspects (e.g., nationality, race) or other characteristics, bring unique opinions, experiences, and perspectives to bear on the task at hand; moreover, in an OSN context, members naturally have different knowledge, community experience, participation motivation and shared information [160, 166, 169]. It is worth noticing that diversity has been generally recognized as a key-enabling dimension in data analysis, which is essential to enhance productivity, develop wiser crowdsourcing processes, improve user satisfaction in content recommendation based on novelty and serendipity, avoid information bubble effects, and ultimately have legal and ethical implications in information processing [53, 162].

Bringing this picture into targeted IM scenarios, let us focus on the problem of *user engagement* [4, 89, 146, 160]. Users that have not yet experienced community commitment (i.e., they are not actively involved in the community life) often hail for different background and motivation, and communicate on diverse topics, which makes engaging them difficult. One effective strategy of user engagement should account for the support and guidance from elder, active members of the community [179]. Therefore, by identifying the most diverse, active members, the triggering stimuli will also be diversified. Since diverse individuals tend to connect to many different types of members, the likelihood of effective engagement would be higher.

The challenge of diversity in targeted IM. Existing targeted IM methods are not designed to embed a notion of diversity in their objective function. In this work, we aim to overcome this limitation, using an unsupervised approach. That is, our research relies on taking a perspective that does not assume any side-information or a-priori knowledge on user attributes (e.g., personal profile, topical preference, community role) that can enable diversification among users. By contrast, we assume that *a user's diversity in a social graph can be determined based on topological properties related to her/his neighborhood*. Remarkably, this finds justifications from social science, particularly from theories of *social embeddedness* [81] and *boundary spanning* [1, 176]. In particular, the latter explains how OSN users acquire knowledge from some of their social contacts and then spread (part of) it to other contacts that belong to one or more components of the social graph, e.g., topically-induced communities, as found in [91].

Our main hypothesis is that if we learn seeds that are not only capable of influencing but also are linked to more diverse (groups of) users, then we would expect that the influence triggers will be diversified as well, and hence the target users will get higher chance of being engaged.

Example 1. To advocate the above hypothesis, consider the example social graph shown in Figure 5.1, where nodes represent individuals and edges express influence relationships. Suppose this graph corresponds to the context of a diffusion process, captured at a given time step, where for the sake of simplicity we omit to indicate both the influence probabilities as edge weights and the active/inactive nodes. Let us focus our attention on the square border node t, which represents a target node,

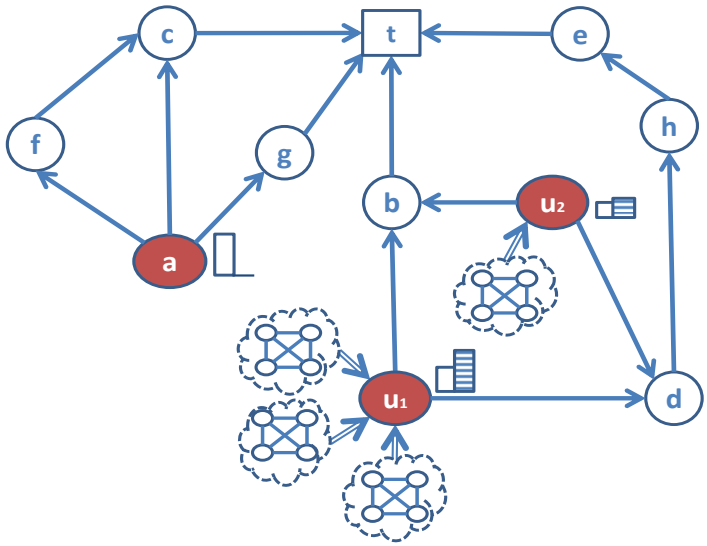


FIGURE 5.1: Effect of topological diversity on the outcome of targeted IM.

and assume that the colored nodes a, u_1, u_2 correspond to candidate seeds, for which we know the individual cumulated spreading influence towards t and the individual topological diversity according to some diversity function; in the figure, these scores are displayed by the leftmost bar and the rightmost bar, respectively, associated to each of the candidate seeds.
A conventional targeted IM method would add node a to the seed set, since it has the highest capability of spread among the candidate seeds; however, a's location has two characteristics that, as we shall explain later, would imply poor topological diversity: it does not receive any incoming connections from other components in the graph, and it diffuses towards nodes that are all in the same subgraph having t as sink. By contrast, the location of nodes u is strategical in terms of topological diversity, since they could be influenced by one or more groups of nodes (in the figure indicated as components enclosed within dashed clouds), thus potentially acquiring a wider spectrum of varied information and perspectives. Selecting nodes u would hence be favored by a diversity-aware targeted IM method as they might be more effective in increasing node t's engagement.

Two main research questions here arise concerning how to leverage users' social diversity in order to enhance the performance of a targeted IM task: **(R1)** how to determine diversity at a large-scale, when we have no a-priori knowledge on user attributes; and **(R2)** how the seed users should be learned by also considering diversity w.r.t. a target set.

Contributions. In this work we contribute with the definition of a novel problem, named ***D**iversity-sensitive **T**argeted **I**nfluence **M**aximization* (DTIM). To the best of our knowledge, we are the first to bring the concept of topology-driven diversity into targeted IM problems. More specifically, to answer **R1**, we provide two alternative ways of modeling topology-driven diversity for targeted IM, which depend on the approach adopted to exploit structural information from the diffusion subgraph specific to a given target node. (Loosely speaking, a target-specific diffusion subgraph corresponds the portion of the diffusion graph involved, at a given time step, in the unfolding of the diffusion towards a particular target node.) The first method, dubbed *local diversity*, is designed to compute node diversity at each step of the expansion of a target-specific diffusion subgraph. The *local diversity* of a node captures

the likelihood of reaching it from nodes outside the currently unfolded target-specific diffusion subgraph. Our second method of topology-driven diversity, dubbed *global diversity*, exploits the structural information of the fully unfolded target-specific diffusion subgraph, and determines the diversity of nodes that lay on the *boundary* of the subgraph, i.e., nodes that can receive influence links from nodes external to the subgraph. Intuitively, this would allow us to capture a boundary-spanning effect of external sources of influence coming from the rest of the social graph.

To address question **R2**, we capitalize on the *local diversity* and *global diversity* definitions to develop alternative algorithms for the DTIM problem, dubbed L-DTIM and G-DTIM. Both algorithms follow a greedy approach that exploits the search for shortest paths in the diffusion graph, in a backward fashion from the selected target set.

We evaluate our DTIM methods on a special case of user engagement in OSNs, which concerns the crowd of users who do not actively contribute to the production of social content. Such silent users, a.k.a. *lurkers*, might have great potential in terms of *social capital*, i.e., acquired knowledge through the observation of user-generated communications. Therefore, it is highly desirable to encourage (a portion of) silent users to more actively participate and give back to the community. Note that while we previously addressed this problem of user engagement in OSNs via a targeted IM approach in [91, 92], in this work we further delve into understanding such a challenging problem under the new perspective of diversity of the seeds to be identified for maximizing the engagement of silent users.

Experimental evaluation using three real-world OSN datasets was conducted to assess the meaningfulness of our approach, mainly in terms of characteristics of the identified seeds and the activated target users, and how they are affected by tuning the input and model parameters of our methods. We also included comparison with two of the most relevant existing IM methods, namely TIM+ [189] and KB-TIM [127], based on the state-of-the-art *RIS* approach. While this comparison has highlighted the uniqueness of our methods, it also suggested to improve their efficiency. In this respect, a further important contribution is the revisiting of RIS-based approximation theory to our diversity-sensitive targeted IM problem.

Organization of the chapter. The rest of the chapter is organized as follows. Section 2 discusses related work, focusing on diversity and targeted IM. Sections 3 presents our diversity-sensitive targeted IM problem, defines two alternative formulations of topology-driven diversity, and presents the L-DTIM and G-DTIM algorithms. In Section 4, we introduce a case study of user engagement for the evaluation of our proposed framework. Experimental evaluation methodology and results are reported in Section 5 and Section 6, respectively. Section 7 describes a RIS-based formulation of DTIM. Section 8 draws conclusions and provides pointers for future research.

5.2 Related work

Diversity in information spreading. Most existing notions of diversity have been developed around structural features of the network, or alternatively based on user profile attributes. This broad categorization applies to various contexts, such as, e.g., web searching and recommendation [122, 171, 203], and information spreading. Focusing on the latter aspect, the authors in [107] propose a measure of controllability, defined as the number of nodes able to spread an opinion through the whole network. In [10], the IC model is extended to take into account the structural diversity of nodes' neighborhood. Main difference between the above mentioned approaches and

our work, relies on the fact that they do not take into account any optimization problem. Other works deal with the problem of estimating the spreading ability of a single node in a network [60, 88]. Node diversity into the IM task has been introduced in [184]. This work shares with ours the linear combination of spread and diversity in the definition of objective function. However, our approach does not depend on user characterization based on topic-biased or categorical distributions.

Targeted influence maximization. Research on targeted IM has gained attention in recent years. A few studies have assumed that the target is unique and a-priori specified. In [77], the authors address the problem of finding the top-k most influential nodes for a specific target user, under the IC model. In [76], the authors investigate optimal propagation policies to influence a target user. In [204], the authors consider the problem of acceptance probability maximization, whereby a selected user (called initiator) wants to send a friendship invitation to a selected target which is not socially close to the initiator (i.e., the two nodes have no common friends). The goal is to find a set of nodes through which the initiator can best approach the target. Unlike the above single-target IM methods, our DTIM approach aims at maximizing the probability of activating a target set which can be arbitrarily large, by discovering a seed set which is neither fixed and singleton nor has constraints related to the topological closeness to a fixed initiator.

In [127], the authors describe a keyword-based targeted IM method, named KB-TIM. This assumes that each user is associated with a weighted term vector to capture her/his preference on advertisements. A user with keywords in common with the advertisement will belong to the target set. KB-TIM relies on a state-of-the-art approach for the classic IM problem, named *reverse influence sampling* (RIS) [20, 189], which provides theoretical guarantees on the solutions. RIS consists of two main steps: (i) computing, for a fixed number θ of nodes selected uniformly at random, the *reverse reachable* sets, i.e., the sets of nodes that can reach them, and (ii) selecting k nodes that cover the maximum number of reverse reachable sets. In [189], the authors show that, when θ is large enough, this set has high probability of being a near-optimal solution to IM. More in detail, they propose the TIM+ algorithm which derives the parameter θ as function of a lower bound of the maximum expected spread among all size-k node sets. The steps of KB-TIM are similar to TIM+, but as the former takes into account only users relevant to an advertisement, it defines a different lower bound for θ. Moreover, while in [20, 189] the random reverse reachable sets are sampled online, KB-TIM allows the sampling procedure to be performed offline by building a disk-based reverse reachable index for each keyword. Other targeted IM approaches for target-aware viral marketing purposes are described in [110, 120, 133, 142].

It is worth emphasizing that, except KB-TIM and TIM+, *all the above works focus on the IC diffusion model.* Note also that the study in [133], which is in principle suited to any diffusion model, actually does not take into account the effect of multiple spreaders (i.e., the diffusion process is considered only for computing the potential influence of each node at a time).

5.3 Targeted influence maximization with topology-driven diversity

5.3.1 Problem statement

Let $G = G_0(b, \ell) = \langle V, E, b, \ell \rangle$ be a directed weighted graph representing the information diffusion graph associated with the social network $G_0 = \langle V, E \rangle$, where V is the set of nodes, E is the set of edges, $b : E \to \mathbb{R}^*$ is an edge weighting function,

and $\ell : V \to \mathbb{R}^*$ is a node weighting function. The edge weighting function b corresponds to the parameter of the *Linear Threshold* (LT) model [97, 199], which we adopt as information diffusion model in this work. Under the LT model, each node can be "activated" by its active neighbors if their total influence weight exceeds the threshold associated to that node. More formally, for any edge (u, v), the weight $b(u, v)$ resembles a measure of "influence" produced by u to v and it is such that $\sum_{u \in N^{in}(v)} b(u, v) \leq 1$, where $N^{in}(v)$ is the in-neighbor set of node v. At the beginning of the diffusion process, each node v is assigned a threshold uniformly at random from $[0, 1]$. Given a set $S \subseteq V$ of initial active nodes, an inactive node v becomes influenced or active at time $\tau \geq 1$, if the total weight of its active neighbors is greater than its threshold. The process runs until no more activations are possible. We denote with $\mu(S)$ the *final active set*, i.e., the set of nodes that are active at the end of the diffusion process starting from S.

Given $G = \langle V, E, b, \ell \rangle$, the node weighting function ℓ determines the status of each node as a *target*, i.e., a node toward which the information diffusion process is directed. More specifically, for any user-specified threshold $L \in [0, 1]$, we define the *target set* TS for G as:

$$TS = \{v \in V \mid \ell(v) \geq L\}. \tag{5.1}$$

The objective function of our targeted IM problem is comprised of two functions. The first one, we call *capital*, is determined as proportional to the cumulative status of the target nodes that are activated by the seed set S.

Definition 17 (Capital)**.** *Given $S \subseteq V$, the capital $C(\mu(S))$ associated with the final active set $\mu(S)$ is defined as:*

$$C(\mu(S)) = \sum_{v \in (\mu(S) \cap TS) \setminus S} \ell(v) \tag{5.2}$$

The capital function corresponds to the cumulative amount of the scores associated with the activated (target) nodes, i.e., $C(\mu(S))$. Remarkably, in Equation (5.2) we do not consider nodes that belong to the seed set S, in order to avoid biasing the seed set by nodes with highest scores.

The second measure is introduced to capture the overall *diversity* of the nodes in set S w.r.t. the target set. We define it in terms of a function div_t that is in turn designed to measure the diversity of a node with respect to each of the target nodes separately.

Definition 18 (Diversity)**.** *Given $S \subseteq V$, the diversity $D(S)$ associated with the target set $TS \subseteq V$ is defined as:*

$$D(S) = \sum_{s \in S} \sum_{t \in TS} div_t(s) \tag{5.3}$$

As previously mentioned, our approach is to measure node diversity in relation to the structural context of the information diffusion graph. In Section 5.3.2 we shall elaborate on different ways of computing *topology-driven diversity*, and provide alternative formulations for the div_t function.

We now formally define our proposed problem of targeted IM, named ***D**iversity-sensitive **T**argeted **I**nfluence **M**aximization* (DTIM).

Definition 19 (Diversity-sensitive Targeted Influence Maximization)**.** *Given a diffusion graph $G = \langle V, E, b, \ell \rangle$, a budget k, and a threshold L, find a seed set $S \subseteq$*

V with $|S| \leq k$ of nodes (users) such that, by activating them, we maximize the Diversity-sensitive Capital (aDC):

$$\begin{aligned} S = & \underset{S' \subseteq \mathcal{V} \ s.t. \ |S'| \leq k}{\text{argmax}} \quad aDC \\ = & \underset{S' \subseteq \mathcal{V} \ s.t. \ |S'| \leq k}{\text{argmax}} \quad \alpha C(\mu(S^{'})) + (1-\alpha) D(S^{'}) \end{aligned} \tag{5.4}$$

where $\alpha \in [0,1]$ is a smoothing parameter that controls the weight of capital C with respect to diversity D.

The objective function of the problem in Equation 5.4 is defined in terms of linear combination of the two functions, capital and diversity. The problem in Def. 19 preserves the complexity of the IM problem and, as a result, it is computationally intractable, i.e., it is still NP-hard. However, as for the classic IM problem, a greedy solution can be designed since that the natural diminishing property holds for the considered problem, as stated in the following.

Proposition 2. *The capital function defined in Equation (5.2) is monotone and submodular under the LT model.*

Proof (sketch). By exploiting the equivalence between LT and the live-edge model shown in [97], for any set $A \subseteq V$ we can express the expected capital of the final active set $\mu(A)$ in terms of reachability under the live-edge graph:

$$C(\mu(A)) = \sum_{\forall X} \Pr(X) C(R^X(A)) \tag{5.5}$$

where $\Pr(X)$ is the probability that a hypothetical live-edge graph X is selected from all possible live-edge graphs, and $R^X(A)$ is the set of nodes that are reachable in X from A. Since for all $v \in V$, $\ell(v)$ is a non-negative value, $C(R^X(A))$ is clearly monotone and submodular. Thus, the expected capital under LT is a non-negative linear combination of monotone submodular functions, and hence it is monotone and submodular, which concludes the proof. □

Proposition 3. *The diversity function defined in Equation (5.3) is monotone and submodular.*

Proof (sketch). As in both the formulations of topology-driven diversity provided above, $div_t(v)$ returns a non-negative value for all $v \in V$, $D(\cdot)$ is clearly monotone. To see that is also submodular, we have to verify that, $\forall S, T \subseteq V$ with $S \subseteq T$ and $\forall v \in V \setminus T$, $D(S \cup \{v\}) - D(S) \geq D(T \cup \{v\}) - D(T)$. For definition of diversity, the above expression can be written as $D(S) + D(\{v\}) - D(S) \geq D(T) - D(\{v\}) - D(T)$, hence it is nondecreasing submodular, which concludes the proof. □
□

In light of these theoretical results, *aDC* is also monotone and submodular as it corresponds to a non-negative linear combination of monotone and submodular functions.

5.3.2 Topology-driven diversity

Our perspective in modeling user diversity is to utilize only structural information given by the topology of a social network graph. Therefore, we take the advantage of a completely *unsupervised* process to avoid requiring any side-information or a-priori

knowledge on user attributes that can enable diversification among users. Instead, we draw inspiration from social science, in that the way a user is connected to others within the OSN (a.k.a. *social embeddedness*) is recognized as a manifestation of diversity of the individual in that online social environment [81]. This is also strictly related to the theory of *boundary spanning* [1], which essentially states that OSN users may naturally get knowledge from some of their social contacts and then spread (part of) it to other contacts through one or more components of the social graph (e.g., topically induced communities). Boundary spanning has also been recognized as an important aspect to consider in order to adequately characterize those users that can show different behaviors in terms of information-production and information-consumption when considering them laying on the boundary of graph components [1, 180]. Upon the above intuitions, we start from the following basic assumption:

Principle 1. *The diversity of a user in a social graph can be determined based on topological properties of her/his neighborhood.*

Definition 20 (Target-specific information diffusion subgraph)**.** *Given the diffusion graph $G = \langle V, E, b, \ell \rangle$, defined over the social graph $G_0 = \langle V, E \rangle$, a target node $t \in TS$, and a time step τ, we define the target-specific diffusion subgraph as the directed acyclic graph $G_t^{(\tau)} = \langle V_t, E_t \rangle \subseteq G_0$, rooted in t, that corresponds to the portion of G involved in the unfolding of the diffusion towards t, at time τ.*

Definition 21 (Boundary set)**.** *Given a target-specific information diffusion subgraph $G_t^{(\tau)}$, its boundary set is defined as the set of nodes having at least one incoming connection from nodes in G outside $G_t^{(\tau)}$:*

$$B_t^{(\tau)} = \{v \in V_t \mid \exists (u, v) \in E \setminus E_t\} \tag{5.6}$$

It is worth noticing here that, while the diffusion starts from a set of seed nodes and follows the directed topology of G, a widely adopted way of modeling the search for nodes that could reach target ones is to use the *backward* or *reverse* depth-first search (e.g., [20, 72, 189]).

Definition 22 (Expansion of target-specific diffusion subgraph)**.** *Given a target-specific information diffusion subgraph $G_t^{(\tau)}$ at time τ, its expansion at time $\tau + 1$ is defined as the graph $G_t^{(\tau+1)}$ resulting from the reverse unfolding of $G_t^{(\tau)}$ such that $G_t^{(\tau+1)}$ contains nodes in G that can reach nodes in the boundary set of $G_t^{(\tau)}$. Moreover, a target-specific diffusion subgraph is said fully expanded if no further backward unfolding over G is possible.*

For the sake of simplification, we hereinafter use symbols G_t, B_t instead of $G_t^{(\tau)}, B_t^{(\tau)}$ as the association with a particular time step τ is assumed to be clear from the context. Moreover, for any $v \in B_t$, we denote with $N^{in}_{\neg E_t}(v) = N^{in}(v) \setminus \{u \mid \exists (u, v) \in E_t\}$ the set of in-neighbors of v that are not linked to v in G_t.

We provide two alternative ways of modeling topology-driven diversity for targeted IM, which depend on the strategy adopted to construct G_t:

- the first method is designed to compute node diversity at each step of the expansion of the information diffusion subgraph for a given target t.Since the method does not require information on the fully expanded diffusion subgraph for t, it is referred to as *local diversity*.
- the second method, named *global diversity*, is instead designed to compute node diversity on the fully expanded target-specific diffusion subgraph.

In the following, we will provide a complete specification of each of the above introduced diversity methods.

5.3.2.1 *Local Diversity*

Our notion of *local diversity* of node is designed to account for the progressive expansion of the information diffusion graph for a given target node.

Given the currently unfolded G_t and a node $v \in B_t$ with $N^{in}_{\neg E_t}(v) \neq \emptyset$, our goal is to determine the *local diversity* for every node u in $N^{in}(v)$ based on two main criteria:

Principle 2. *The diversity of node u should be proportional to the likelihood of reaching it from nodes outside the currently unfolded target-specific diffusion subgraph G_t, i.e., proportional to the number of u's in-neighbors in G not already in G_t.*

Principle 3. *The diversity of node u should be proportional to the increment contributed by that node to the number of incoming links not already included in G_t.*

Accordingly, we first characterize the diversity in the boundary set of G_t, and its incremental update due to the insertion of a new node to G_t, then we provide our definition of *local diversity*.

Definition 23 (Boundary diversity of set)**.** *Given the currently unfolded G_t, the boundary diversity δ_t of G_t is defined as the number of nodes in $N^{in}_{\neg E_t}(v)$ averaged over nodes v in B_t:*

$$\delta_t = \frac{1}{|B_t|} \sum_{v \in B_t} |N^{in}_{\neg E_t}(v)| \tag{5.7}$$

Note that the above definition is simple yet convenient to use in incremental computations. Moreover, it is directly related to the amount of possible paths to diffuse towards a particular target node. The study of alternative definitions of boundary diversity could be an interesting direction as future work.

For each $u \in N^{in}(v)$, with $v \in B_t$, if u is inserted in G_t, the boundary diversity will change accordingly, since B_t is updated to contain u. The boundary diversity w.r.t. B_t being updated with u, denoted with δ_t^{+u}, is straightforwardly determined as follows:

$$\delta_t^{+u} = \frac{|B_t|\delta_t + |N^{in}_{\neg E_t}(u)|}{|B_t|+1} \tag{5.8}$$

Definition 24 (Local diversity)**.** *The local diversity of u is defined as the ratio of the boundary diversity conditional on inclusion of u in G_t, to the actual boundary diversity:*

$$div_t(u) = \frac{\delta_t^{+u}}{\delta_t} = \frac{|B_t|}{1+|B_t|}\left(1 + \frac{|N^{in}_{\neg E_t}(u)|}{\sum_{v \in B_t} |N^{in}_{\neg E_t}(v)|}\right) \tag{5.9}$$

Intuitively, the *local diversity* applies to any node u that is in-neighbor of some node that lays on the boundary of the currently unfolded G_t, and expresses the increment due to node u to the overall likelihood of being reached from more different portions of the diffusion graph G.

5.3.2.2 *Global Diversity*

Our second method of topology-driven diversity computation relies on the availability of structural information of the fully expanded target-specific diffusion subgraph. While this solution loses the advantage of incremental computation, it also opens to the opportunity of using more structural features to measure the diversity of a node.

Given a target node t, G_t is here meant as the fully expanded diffusion subgraph for t. Moreover, the definition of *boundary* given in Equation 5.6 as well as the definition of *boundary diversity* given in Equation 5.7 do not change; however, we will exploit them at a "node level" rather than a "set-level" as for the *local diversity*.

First, the boundary diversity here assumes a slight different meaning with respect to the *local diversity* case. It still captures the strength of the flow potentially spanning over portions of the diffusion graph not already unfolded, which makes Principle 2 hold; however, since the target-specific diffusion subgraph G_t is considered as definitively unfolded, we conceptualize that:

Principle 4. *The boundary spanning should be regarded as exogenous to the diffusion process for a specific target, and hence intuitively associated to external sources of influence coming from the rest of the social graph.*

Definition 25 (Boundary diversity of node)**.** *Given a node $v \in B_t$, the boundary diversity of v is defined as the contribution of v to the boundary diversity δ_t:*

$$div_t^B(v) = \frac{|N^{in}_{\neg E_t}(v)|}{|B_t|} \tag{5.10}$$

Boundary diversity is set to zero for any $v \in V_t \setminus B_t$.

While the concept of boundary diversity is essential to characterize the connectivity of boundary nodes from outside G_t, we also consider here to measure their *outward* connectivity within G_t as the contribution a node gives to the average number of out-neighbors of nodes in B_t that belong to G_t. We denote the latter as $|N^{out}_{E_t}(v)|/|B_t|$. Moreover, we observe that, from the perspective of maximizing diversity of nodes that propagates towards a given target, the overall measure of diversity of node should be not only obviously proportional to its boundary diversity, but also proportional to its outward internal span. The above considerations lead to the following definition.

Definition 26 (Global diversity)**.** *The global diversity of node v is defined as:*

$$div_t(v) = div_t^B(v) \times \mathrm{f}\left(\frac{|N^{out}_{E_t}(v)|}{|B_t|}\right) \tag{5.11}$$

where f *is a smoothing function to assign the outward internal span a weight at most equal to the boundary diversity term.*

In the following, we will refer to a logarithmic smoothing, i.e., $\mathrm{f} = \log(1 + |N^{out}_{E_t}(v)|/|B_t|)$, since we want the outward internal span of node has an impact lower than the boundary diversity on the overall value of diversity.

5.3.3 The DTIM algorithms

In this section, we show our algorithmic solutions to the proposed **D**iversity-sensitive **T**argeted **I**nfluence **M**aximization problem. According to the *local diversity* and *global diversity* criteria previously introduced in Section 5.3.2, we provide two methods,

named L-DTIM and G-DTIM, respectively; due to space limits of this chapter, they are concisely reported in Algorithm 4.

Following the lead of the study in [72], L-DTIM and G-DTIM exploit as well the search for shortest paths in the diffusion graph, however in a backward fashion. Along with the information diffusion graph G, the budget integer k, the minimum score L and a parameter $\alpha \in [0, 1]$ which controls the balance between capital and diversity, L-DTIM and G-DTIM take in input a real-valued threshold η. This parameter is used to control the size of the neighborhood within which paths are enumerated: in fact, the majority of influence can be captured by exploring the paths within a relatively small neighborhood; note that for higher η values, less paths are explored (i.e., paths are pruned earlier) leading to smaller runtime but with decreased accuracy in spread estimation.

Algorithm 4 DTIM- **D**iversity-sensitive **T**argeted **I**nfluence **M**aximization

Input: A graph $G = \langle V, E, b, \ell \rangle$, a budget (seed set size) k, a target selection threshold $L \in [0, 1]$, a path pruning threshold $\eta \in [0, 1]$, a smoothing parameter $\alpha \in [0, 1]$.
Output: Seed set S.

1: $T \leftarrow V$ {*nodes that can reach target nodes*}
2: **for** $u \in V$ **do**
3: **if** $\ell(u) \geq L$ **then**
4: $TS \leftarrow TS \cup \{u\}$ {*identifies the target nodes*}
5: **end if**
6: $u.Dset \leftarrow \{\}$ {*initializes a data structure that keeps track of node diversity w.r.t. any target*}
7: **end for**
8: **while** $|S| < k$ **do**
9: **for** $u \in T \setminus S$ **do**
10: $u.C, u.D \leftarrow 0$ {*initializes each node's capital and diversity to zero*}
11: **end for**
12: $T \leftarrow \emptyset$
13: **for** $t \in TS \setminus S$ **do**
14: $G_t = \langle V_t, E_t \rangle \leftarrow \langle \{t\}, \emptyset \rangle$ {*initializes DAG rooted in t*}
15: backward$(\langle t \rangle, 1, t)$
16: **if** $|S| = 0$ **then**
17: updateDiversity(t)
18: **end if**
19: **end for**
20: $S \leftarrow S \cup \{bestSeed\}$
21: **end while**
22: **return** S

23: **procedure** backward$(\mathcal{P}, pp, t)$
24: $v \leftarrow \mathcal{P}.last(), \ T \leftarrow T \cup \{u\}$
25: **while** $u \in N^{in}(v) \ \wedge \ u \notin S \cup \mathcal{P}.nodeSet()$ **do**
26: $pp \leftarrow pp \times b(u, v)$ {*updates the path probability*}
27: **if** $pp \geq \eta$ **then**
28: $u.C \leftarrow u.C + pp \times \ell(t)$ {*updates the overall node capital*}
29: **if** $|S| = 0$ **then**
30: $u.inf \leftarrow u.inf + pp$ {*increases the overall influence of node u on the current target*}
31: (*) $u.Dset(t) \leftarrow div_t(u)$ {*computes the current node diversity w.r.t. the target by Eq.5.9*}
32: $G_t = \langle V_t \cup \{u\}, E_t \cup \{(u, v)\} \rangle$ {*adds the edge (u, v) to the explored DAG*}
33: **else**
34: $u.D \leftarrow u.D + pp \times u.Dset(t)$
35: **if** $u.aDC > bestSeed.aDC$ **then**
36: $bestSeed \leftarrow u$ {*sets the current best seed node as u*}
37: **end if**
38: **end if**
39: backward$(\mathcal{P}.append(u), pp, t)$
40: **end if**
41: **end while**

42: **procedure** updateDiversity(t)
43: **for** $v \in V_t$ **do**
44: (**) $v.Dset(t) \leftarrow div_t(v)$ {*computes node diversity w.r.t. the target t by Eq. 5.11*}
45: $v.D \leftarrow v.D + v.inf \times v.Dset(t)$ {*updates the overall node diversity*}
46: $v.inf \leftarrow 0$
47: **if** $v.DIC > bestSeed.aDC$ **then**
48: $bestSeed \leftarrow v$ {*sets the current best seed node as v*}
49: **end if**
50: **end for**

(*) Instruction at line 31 is performed by L-DTIM only.
(**) Instruction at line 44 is performed by G-DTIM only.

In order to yield a seed set S of size at most k, during each iteration of the main

loop (lines 8-21), both the variants of Algorithm 4 compute the set T of nodes that reach the target ones and keep track, into the variable *bestSeed*, of the node with the highest marginal gain (i.e., diversity-sensitive capital aDC).

The *bestSeed* node is found at the end of each iteration upon calling the subroutine backward over all nodes in TS that do not belong to the current seed set S. This subroutine takes a path $\mathcal{P}$, its probability pp and the target t from which the visit has started, and extend $\mathcal{P}$ as much as possible (i.e., as long as pp is not lower than η). Initially, a path is formed by one target node, with probability 1 (line 15). Then, the path is extended by exploring the graph backward, adding to it one, unexplored in-neighbor u at time, in a depth-first fashion. Path probability is updated (line 26) according to the LT-equivalent "live-edge" model [72, 97], and so the capital (line 28). The process is continued until no more nodes can be added to the path.

Both G-DTIM and L-DTIM compute the node diversity only at the first iteration of the main loop, i.e., when the seed set S is empty. Indeed, for each node, we keep track of its diversity w.r.t. each target it can reach, by using data structure $Dset$. A major difference between the two variants is that in G-DTIM the node diversity is computed (through the subroutine updateDiversity) only when the whole subgraph rooted in t has been completely built (line 44). In L-DTIM, instead, the node diversity is updated every time the node has been reached (line 31). Note that the instruction at line 31 (resp. 44) is performed by L-DTIM (resp. G-DTIM) only. The value of diversity of a node v is, in both the variants, smoothed with the influence that v might exert on t, contributing to the overall diversity D of v (line 45).

Note that both the numerical values yielded by both global diversity and local diversity functions div_t might be subject to scaling in order to enable a fair comparison with the numerical value yielded by the capital.

Example 2. Consider the example in Figure 5.2, where the target set includes the square border node $\{t\}$. Let's assume for simplicity we set $k = 1$, $\alpha = 0.5$, $\eta = 0$ and we ignore the spread computation for nodes inside the other components of G_t (represented within clouds in the figure). Moreover, the double arrows connecting these components to nodes u_1 and u_2 count as two edges each. In the following, we denote with $pp[x \to \cdots \to y]$ the probability of the path from x to y, and with $x.inf$ the overall influence exerted by node x to the target.

The target node t can be reached through a (with $a.inf = pp[a \to f \to c \to t] + pp[a \to c \to t] + pp[a \to g \to t] = 0.098 + 0.06 + 0.24$), b (with $b.inf = pp[b \to t] = 0.35$), c (with $c.inf = pp[c \to t] = 0.2$), d (with $d.inf = pp[d \to h \to e \to t] = 0.045$), e (with $e.inf = pp[e \to t] = 0.15$), f (with $f.inf = pp[f \to c \to t] = 0.14$), g (with $g.inf = pp[g \to t] = 0.3$), h (with $h.inf = pp[h \to e \to t] = 0.09$), u_1 (with $u_1.inf = pp[u_1 \to d \to h \to e \to t] + pp[u_1 \to b \to t] = 0.0135 + 0.21$), and u_2 (with $u_2.inf = pp[u_2 \to d \to h \to e \to t] + pp[u_2 \to b \to t] = 0.0315 + 0.14$). Node a has the largest chance of success in activating t, which results in the highest capital C. However, since a does not have in-neighbors, its diversity is equal to zero for both the diversity formulations.

Let us first focus on the behavior of G-DTIM. According to Equation 5.6, the set of boundary nodes is $B_t = \{u_1, u_2\}$. By definition of *global diversity* (Equation 5.11), G-DTIM computes the following values: $u_1.D = 2.08$ (as $div_t^B(u_1) = 6/2$ and $div_t(u_1) = 3 \times \log(1+2/2)$), $u_2.D = 0.69$ (as $div_t^B(u_2) = 2/2$ and $div_t(u_1) = 1 \times \log(1+2/2)$). By applying the max-normalization to the node diversity, the final values are $u_1.D = 1$ and $u_2.D = 0.33$. As a result, for G-DTIM node u_1 is chosen as seed node since it has diversity-sensitive capital ($aDC = 0.22 \times 0.5 \times (0.5 + 1) = 0.165$) higher than that of a ($aDC = 0.4 \times 0.5 \times (0.5 + 0) = 0.1$) and u_2 ($aDC = 0.13 \times 0.5 \times (0.5 + 0.33) = 0.05$).

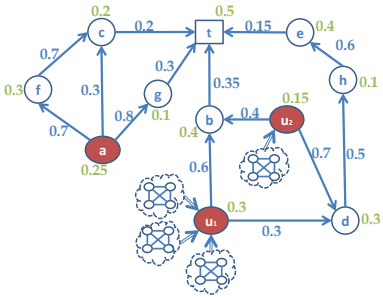


FIGURE 5.2: Targeted IM vs. diversity-sensitive targeted IM. Edge weights (values in blue) and node weights (values in green) are computed by functions b and ℓ. To avoid cluttering of the figure, the node activation thresholds used by LT model here coincide with the node weights.

The values of node diversity computed by L-DTIM depend on the order in which nodes are reached during the backward visit. Assume to visit first the branch starting from node e. According to Equation 5.9, L-DTIM computes the following values of node diversity: $e.D = 0.625$ ($div_t(e) = 1/2 \times (1 + 1/4)$ as $B_t = \{t\}$), $h.D = 0.83$ ($div_t(h) = 2/3 \times (1+1/4)$ as $B_t = \{t, e\}$), $d.D = 1$ ($div_t(d) = 2/3 \times (1+2/4)$ as $B_t = \{t, h\}$), and, assuming to visit u_1 before u_2, $u_1.D = 1.47$ ($div_t(u_1) = 2/3 \times (1+6/5)$ as $B_t = \{t, d\}$), $u_2.D = 0.9$ ($div_t(u_2) = 3/4 \times (1+2/10)$ as $B_t = \{t, d, u_1\}$). Analogously, it proceeds in computing the node diversity through branches c and g, whose values of diversity are lower than 0.9 (not reported for the sake of readability). L-DTIM eventually computes the following diversity: $b.D = 0.92$ ($div_t(b) = 3/4 \times (1 + 2/9)$ as $B_t = \{t, u_1, u_2\}$), $u_1.D = 1.2$ ($div_t(u_1) = 3/4 \times (1 + 6/10)$ as $B_t = \{b, u_1, u_2\}$), and $u_2.D = 0.92$ ($div_t(u_2) = 3/4 \times (1+2/9)$ as $B_t = \{b, u_1, u_2\}$). Upon max-normalization to the values so obtained, L-DTIM will choose b as seed node since it has diversity-sensitive capital ($aDC = 0.35 \times 0.5 \times (0.5 + 0.77) = 0.22$) higher than that of u_1 ($aDC = 0.22 \times 0.5 \times (0.5 + 1) = 0.165$).

5.4 Using DTIM to engage silent users in social networks

We evaluate our framework of targeted IM with topology-driven diversity on a special case of user engagement in OSNs, which refers to the problem of *how to turn silent users into more active contributors* in the community life.

All large-scale OSNs are characterized by a participation inequality principle: the crowd does not take an active role in the interaction with other members, rather it takes on a silent role. Silent users are also referred to as *lurkers*, since they gain benefit from information produced by others, by observing the user-generated communications at all stages (e.g., reading posts, watching videos, etc.), but without significantly giving back to the community [56, 179].

Social science and human-computer interaction research communities have widely investigated the main causes that explain lurking behaviors, which include subjective reticence (rather than malicious motivations) to contribute to the community wisdom, or a feeling that gathering information by browsing is enough without the need of being further involved in the community. Moreover, lurking can be expected or even encouraged because it allows users (especially newcomers) to learn or improve their understanding of the etiquette of an online community [56].

Regardless of their motivations, lurkers might have great potential in terms of

social capital, because they acquire knowledge from the OSN. They can become aware of the existence of different perspectives and may make use of these perspectives in order to form their own opinions, but they are unlikely to let other people know their value. In this regard, it might be desirable to engage such users, or *delurk* them, i.e., to develop a mix of strategies aimed at encouraging lurkers to return their acquired social capital, through a more active participation to the community life.

Engagement actions towards silent users can be categorized into four types [179]: reward-based external stimuli, providing encouragement information, improvement of the usability and learnability of the system, guidance from elders/master users to help lurkers become familiar with the system as quickly as possible. It is worth emphasizing that *our approach is independent on the particular strategy of delurking being adopted*. The goal here is how to instantiate our DTIM algorithms in a user engagement scenario where *lurkers are regarded as the target users* of the diffusion process. Therefore, our goal becomes: Given a budget k, to find a set of k nodes that are capable of maximizing the diversity-sensitive capital, i.e., the likelihood of activating the target silent users through diverse seed users.

A key aspect of our approach in this scenario is that the selection of target users is based on the solution produced by a *lurker ranking* algorithm [180–182] applied to the social network graph G_0. In Section 5.4.1 we provide a summary of the lurker ranking method we used in this work, and in Section 5.4.2 we describe how the input diffusion graph for DTIM is modeled, following our early work in [92].

5.4.1 Identifying target users through LurkerRank

Lurker ranking methods, originally proposed in [180, 182], are designed to mine silent user behaviors in the network, and hence to associate users with a score indicating her/his lurking status. Lurker ranking methods rely upon a *topology-driven definition of lurking* which is based on the network structure only. Upon the assumption that lurking behaviors build on the *amount of information a node receives*, the key intuition is that the strength of a user's lurking status can be determined based on three basic principles: overconsumption, authoritativeness of the information received, non-authoritativeness of the information produced.

The above principles form the basis for three ranking models that differently account for the contributions of a node's in-neighborhood and out-neighborhood. A complete specification of the lurker ranking models is provided in terms of PageRank and AlphaCentrality based formulations. For the sake of brevity here, we will refer to only one of the formulations described in [180, 182], which is that based on the full *in-out-neighbors-driven lurker ranking*, hereinafter dubbed simply as LurkerRank (LR).

Given the directed social graph $G_0 = \langle V, E \rangle$, where any edge (u, v) means that v is is "consuming" or "receiving" information from u, the LurkerRank $LR(v)$ score of node v is defined as:

$$LR(v) = [\mathcal{L}_{\text{in}}(v)\ (1 + \mathcal{L}_{\text{out}}(v))] + (1 -)p(v) \tag{5.12}$$

where $\mathcal{L}_{\text{in}}(v)$ is the in-neighbors-driven lurking function:

$$\mathcal{L}_{\text{in}}(v) = \frac{1}{out(v)} \sum_{u \in N^{in}(v)} \frac{out(u)}{in(u)} LR(u) \tag{5.13}$$

and $\mathcal{L}_{\text{out}}(v)$ is the out-neighbors-driven lurking function:

$$\mathcal{L}_{\text{out}}(v) = \frac{in(v)}{\sum_{u \in N^{out}(v)} in(u)} \sum_{u \in N^{out}(v)} \frac{in(u)}{out(u)} LR(u) \tag{5.14}$$

where: $in(v)$ (resp. $out(v)$) denotes the size of the set of in-neighbors (resp. out-neighbors) of $v, is a damping factor ranging within [0,1] (usually set to 0.85), and p(v) is the value of the personali$ and $out(\cdot)$ are Laplace add-one smoothed.

5.4.2 Modeling the diffusion graph

In Section 5.3.1, we introduced symbol $\ell(v)$ to denote the weight of node v that quantifies its status as target. In this application scenario, the higher is the lurker ranking score of v the higher should be $\ell(v)$.

We define the node weighting function ℓ upon scaling and normalizing the stationary distribution produced by the LurkerRank algorithm over G_0. The scaling compensates for the fact that the lurking scores produced by LurkerRank, although distributed over a significantly wide range (as reported in [180]), might be numerically very low (e.g., order of 1.0e-3 or below). Moreover, we introduce a small smoothing constant in order to avoid that the highest lurking scores are mapped exactly to 1. Formally, for each node $v \in V$, we define the *node lurking value* $\ell(v) \in [0,1)$ as follows:

$$\ell(v) = \frac{\widetilde{\pi_v} - min_r}{(max_r - min_r) + \epsilon_r} \tag{5.15}$$

where $\widetilde{\pi}$ denotes the stationary distribution of the lurker ranking scores (π) divided by the base-10 power of the order of magnitude of the minimum value in π, $\widetilde{\pi_v}$ is the value of $\widetilde{\pi}$ corresponding to node v, $max_r = \max_{u \in V} \widetilde{\pi_u}$, $min_r = \min_{u \in V} \widetilde{\pi_u}$, and ϵ_r is a smoothing constant proportional to the order of magnitude of the max_r value.

In order to define the edge weights so that they express a notion of strength of influence from a node to another (as normally required in an information diffusion model), we again exploit information derived from the ranking solution obtained by LurkerRank as well as from the structural properties of the social graph. Our key idea is to calculate the weight on edge $(u,v) \in E$ proportionally to the fraction of the original lurking score of v given by its in-neighbor u:

$$b_0(u,v) = \left[\sum_{w \in N^{in}(v)} \frac{out(w)}{in(w)} \pi_w \right]^{-1} \frac{out(u)}{in(u)} \pi_u \tag{5.16}$$

Using Equation (5.16), we finally define the edge weight as:

$$b(u,v) = b_0(u,v) \times e^{\ell(v)-1} \tag{5.17}$$

Note that Equation (5.17) meets the requirement $\sum_{u \in N^{in}(v)} b(u,v) \leq 1$, and accounts for $\ell(v)$ such that the resulting weight on (u,v) is lowered for higher $\ell(v)$, i.e., the more a node acts as a lurker, the more active in-neighbors are needed to activate that node.

5.5 Evaluation methodology

5.5.1 DTIM settings

We experimentally varied the input and model parameters in DTIM methods, namely: the size of seed set (k), the target selection threshold (L), the path pruning threshold (η), and the parameter α to control the contribution of diversity versus capital in the objective function of DTIM methods. Note that, to simplify the interpretation of L, we will instead use symbol L-*perc* to denote a percentage value that determines the setting of L such that the selected target set corresponds to the top-L-*perc* of the distribution of scores yielded by function ℓ; particularly, we set L-*perc* $\in \{5\%, 10\%, 25\%\}$. As concerns η, though $\eta = 1.0e$-03 is the default as used in other IM algorithms (e.g.,[72]), we set it to a lower value, $\eta = 1.0e$-04, to impact even less on the unfolding of the information diffusion process; moreover, we will not present results corresponding to $\eta = 0$ (i.e., no path-pruning), since we observed this negatively affects the runtime by several orders of magnitude while yielding nearly identical results to those corresponding to $\eta = 1.0e$-04.

5.5.2 Competing methods

We considered comparison with TIM+ [189] and KB-TIM [127], which are state-of-the-art solutions to the IM (resp. targeted IM) problem, based on the RIS approach (cf. Section 5.2).

Comparing DTIM with a non-targeted IM algorithm like TIM+ required to evaluate the quality of seed sets produced by the competing algorithm under a *targeted* scenario. To this purpose, we simply let TIM+ compute a size-k seed set over the entire graph and then we estimated the capital over different target sets in accord with the setting of DTIM. We considered two opposite settings for the main parameter (ϵ) in TIM+: (i) the default $\epsilon = 0.1$, which provides strong theoretical guarantees yet is adversarial to the algorithm's memory consumption, and (ii) $\epsilon = 1.0$, which conversely provides no approximation guarantees but high empirical efficiency; note that the latter setting was also used by the TIM+'s authors in [189] for the comparison with SimPath. We used default settings for the other parameters in TIM+.

As concerns KB-TIM, we modified the keyword-based target selection stage to make it equivalent to the target selection adopted in DTIM. KB-TIM requires two main input files to drive the target selection: (i) a sort of document-term sparse matrix, such that each node (document) in the graph is assigned a list of *keyword*, *#occurrences* pairs, and (ii) a list of keyword-queries, so that each query corresponds to the selection of a subset of nodes in the graph. To prepare these input files, we defined three queries corresponding to the setting L-*perc* $\in \{5\%, 10\%, 25\%\}$, and accordingly created the sparse matrix so that each node was assigned a keyword for each of the top-ranked subsets it belongs to (e.g., a node in the top-10% set of lurkers will be assigned two keywords, as it is also in the top-25% set); moreover, the *#occurrences* associated with any keyword for a given node v was calculated as the node lurking value $\ell(v)$ suitably scaled and truncated to its integer part. Also, we used the incremental reverse-reachable index (IRR) in KB-TIM.

data	*# nodes*	*# links*	*avg in-deg.*	*avg path len.*	*clust. coeff.*	*assorta-tivity*
FriendFeed	493,019	19,153,367	38.85	3.82	0.029	-0.128
GooglePlus	107,612	13,673,251	127.06	3.32	0.154	-0.074
Instagram-LCC	17,521	617,560	35.25	4.24	0.089	-0.012

TABLE 5.1: Summary of the evaluation network datasets

5.5.3 Data

We used FriendFeed [28], GooglePlus [117], and Instagram [181][1] network datasets. Note that, for the sake of significance of the information diffusion process in latter network, we selected the induced subgraph corresponding to the maximal strongly connected component of the original network graph, hereinafter referred to as `Instagram-LCC` (LCC stands for largest connected component). As major motivations underlying our data selection, we wanted to maintain continuity with our previous studies [180, 181] and use publicly available datasets. Table 5.1 summarizes main structural characteristics of the evaluation network datasets.

5.6 Results

We present results of the evaluation of our proposed DTIM algorithms according to three main objectives: analysis of the identified seed nodes (Section 5.6.1), analysis of the activated target nodes (Section 5.6.2) and efficiency analysis (Section 5.6.3).[2]

5.6.1 Evaluation of identified seed nodes

5.6.1.1 Seed set overlap

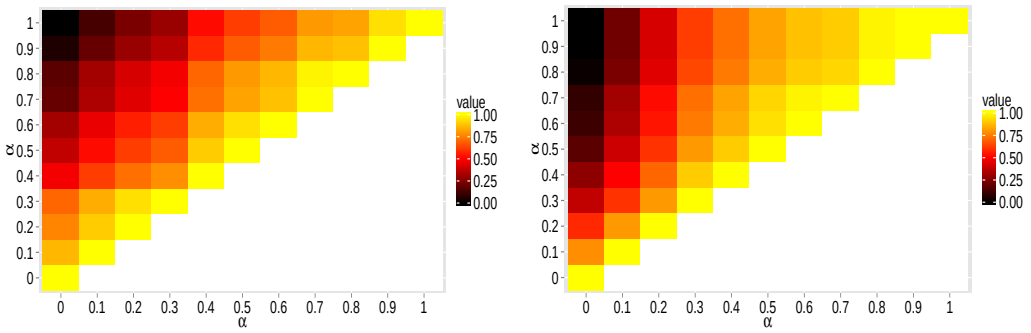
In order to investigate the impact of taking into account diversity on the seed identification process, we initially analyzed the matching among seed sets produced by the two DTIM methods with varying α.

This analysis of seed sets was twofold: (i) pair-wise evaluation of the overlaps between seed sets produced by a particular DTIM method by varying α, and (ii) pair-wise evaluation of the overlaps between seed sets produced by G-DTIM and L-DTIM for particular values of α. Unless otherwise specified, results correspond to the largest sizes of target set and seed set we considered (i.e., $L\text{-}perc = 25\%$ and $k = 50$), and express the *normalized overlap* of any two seed sets, i.e., their intersection divided by the seed set size.

Normalized seed set overlap. On `GooglePlus` (Figure 5.3), the normalized overlap values span over the full range $[0.0, 1.0]$, for both methods. In the heatmap corresponding to G-DTIM, an overlap above than 50% is observed for values of α in different subintervals, while variations in the seed set are generally more uniform for L-DTIM, whereby the normalized overlap increases for higher values of α. Also, for both methods there is no overlap when comparing the seed set obtained for $\alpha = 0$ (i.e., full contribution of diversity in the DTIM objective function) with the seed set obtained for any $\alpha > 0$. These remarks generally hold regardless of the target set size when

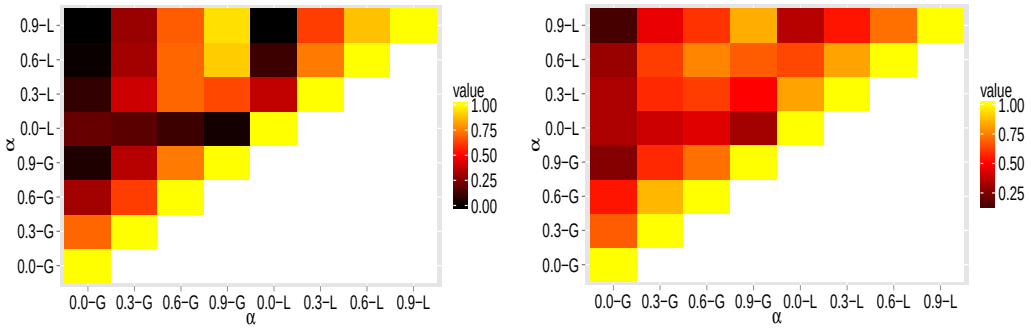
[1] Available at http://people.dimes.unical.it/andreatagarelli/data/.

[2] All experiments were carried out on an Intel Core i7-3960X CPU @3.30GHz, 64GB RAM machine. All algorithms were written in C++. All competing algorithms refer to the original source code provided by their authors.



(a) G-DTIM (b) L-DTIM

FIGURE 5.3: Heatmaps of normalized overlap of seed sets, for varying α, with *L-perc* $= 25\%$ and $k = 50$, on `GooglePlus`.



(a) `GooglePlus` (b) `FriendFeed`

FIGURE 5.4: Heatmaps of normalized overlap of seed sets between G-DTIM and L-DTIM, for $\alpha = \{0.0, 0.3, 0.6, 0.9\}$, *L-perc* $= 25\%$ and $k = 50$. (Suffix -L, resp. -G, denotes a particular setting of α that refers to L-DTIM, resp. G-DTIM.)

using L-DTIM, while the contingencies of null overlap are more likely to occur for lower *L-perc* when using G-DTIM. A large spectrum of normalized overlap values are observed on `FriendFeed` as well (results not shown), particularly at least 0.25 for G-DTIM and 0.4 for L-DTIM. Null overlap is mainly observed for low seed set size ($k = 5$ using L-DTIM, and $k \leq 15$ using G-DTIM). By contrast, `Instagram-LCC` generally shows a quite higher overlap than in the other networks (results not reported), which might be ascribed to the particular contingency of strong connectivity that characterizes `Instagram-LCC`.

Comparison between G-DTIM and L-DTIM seed sets. Figure 5.4 shows results on the comparison of seed sets identified by G-DTIM and L-DTIM, respectively, corresponding to $\alpha = \{0.0, 0.3, 0.6, 0.9\}$. On `GooglePlus` (Figure 5.4(a)), the seed sets appear to be significantly different from each other for higher contributions of diversity in the objective function ($\alpha < 0.3$), while values of normalized overlap in the range $[0.5, 1]$ are observed for higher values of α. Analogous observations can be drawn for `FriendFeed` (Figure 5.4(b)), yet with lower overlap values also for values of α in the range $[0.6, 0.9]$ (i.e., normalized overlap around 0.75).

Comparison with TIM+ and KB-TIM. We also analyzed the matching between seed sets produced by DTIM algorithms and competing ones (results not shown). Here we refer to the setting $\alpha = 1.0$ (i.e., no diversity contribution), since TIM+ and KB-TIM do not integrate any diversity notion in their formulations. The minimum overlap of seed sets produced by DTIM is reached against KB-TIM in all cases and on

all datasets; in particular, with the setting $k = 50$, $L\text{-}perc = 25\%$, 0.48 for `FriendFeed`, 0.46 for `GooglePlus`, 0.60 for `Instagram-LCC`. In general, for large k, the normalized overlap is within medium regimes, while it is close or equal to zero on FriendFeed. Only for $k = 5$, the normalized overlap corresponds to mid-high values on GooglePlus and Instagram-LCC. DTIM with $\alpha = 1$ can have relatively high overlap with TIM+ (about 0.75), especially for high $L\text{-}perc$, on all datasets. However, for lower $L\text{-}perc$, the overlap is low (for smaller k) to medium (for higher k).

Discussion. The seed set overlap analysis has revealed that accounting for diversity can yield significant differences in the behavior of the DTIM methods in terms of seed identification. Indeed, by varying α within its full regime of values leads to a wide spectrum of values of normalized seed set overlap. In particular, the changes in overlap are more evident when varying α at lower regimes, thus indicating that higher contribution of diversity w.r.t. capital leads to more significantly diversified seed sets. Remarkably, the overlap can be close to zero when comparing two seed sets respectively obtained with $\alpha = 0$ and with $\alpha = 1$, i.e., completely different seed nodes can be identified when accounting for either diversity or capital only in the target IM objective function.

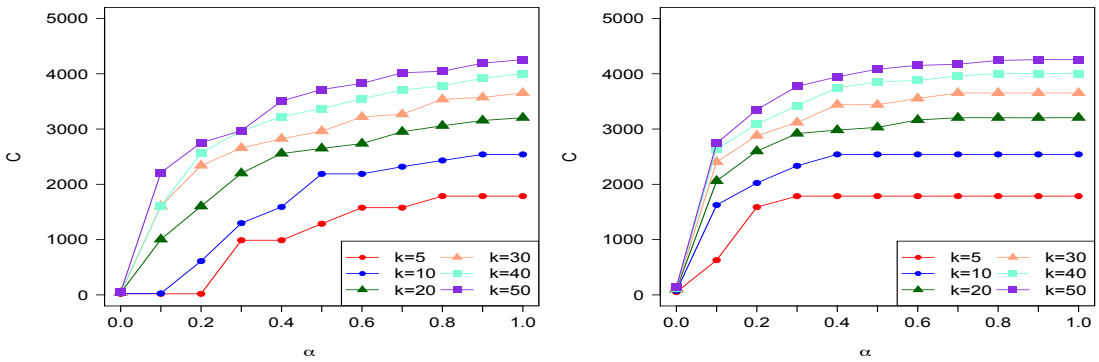
The two proposed notions of diversity turn out to be quite dissimilar to each other: indeed, the normalized overlap of seed sets yielded by L-DTIM and G-DTIM, respectively, is generally below 50%, which is further reduced for low values of α. The local diversity notion appears to be less sensitive to α than global diversity; however, for low α and size of target set, L-DTIM tends to produce more diverse seed sets than G-DTIM, for any particular setting of k.

Our DTIM methods with $\alpha = 1$ produce seed sets that have overlap with KB-TIM ones below 50% on `FriendFeed` and `GooglePlus`, and 60% on `Instagram` for $k = 50$, $L\text{-}perc = 25\%$; when compared to TIM+, the seed set overlap can be relatively higher.

5.6.1.2 Structural characteristics of seeds

We analyzed topological characteristics of the identified seeds, focusing on basic measures of node centrality, namely *outdegree*, *betweenness*, and *coreness*. We present here a summary of main findings, and refer the reader to the Appendix C for detailed results.

One major remark that stands out is that accounting for diversity in DTIM methods produces the effect of choosing seed nodes that can differ from those that would be obtained otherwise (i.e., using only capital term in the objective function) according to selected topological criteria. This result, coupled with analogous considerations previously drawn about diversification in terms of set overlap, hence strengthens the significance of accounting for diversity in the targeted IM process. Structural characteristics tend to be marginally affected by the setting of $L\text{-}perc$ when L-DTIM is used, while the behavior with G-DTIM is much more dependent on $L\text{-}perc$, especially for smaller size of target set ($L\text{-}perc = 5\%$). Also, each of the competing methods leads to the identification of seeds that are less different from each other than DTIM seeds being obtained for most of the settings of α, in terms of all the topological measures considered.



(a) G-DTIM (b) L-DTIM

FIGURE 5.5: Capital in function of α and k, with *L-perc* set to 25%, on `GooglePlus`.

5.6.2 Evaluation of activated target nodes

5.6.2.1 Capital

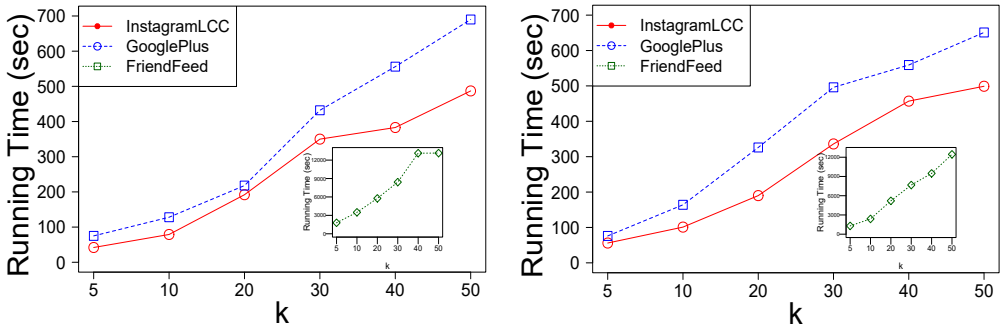
We discuss results on the expected capital of the target users activated by a given set of seed users. The estimation procedure is based on the results of I_{MC} Monte Carlo simulations of the LT diffusion process, with I_{MC} set to 10 000.[3] Note that while the identification of the seeds depends on the full DTIM objective function, here we focus on the value of the capital function C only.

Beyond the expected increase in capital with α (which means weighting less diversity than capital in the objective function), the impact of α on the behavior of DTIM algorithms is evident, especially for $k > 10$, with capital value that can vary up to three orders of magnitude. The generally upward trends of C are explained in function of both α and k, particularly they are more rapidly increasing for mid-low α and $k > 10$. Also on all datasets, L-DTIM yields a higher average capital value, for every k, than that observed with G-DTIM. Similar overall behaviors are shown by the DTIM algorithms for different sizes of target set.

More in detail, on `GooglePlus` (Figure 5.5), when using G-DTIM the capital value increases rapidly, reaching around 80% for $\alpha < 0.5$ and $k \geq 20$; for L-DTIM, we observe an even sharper increase in the value of C for small α (0.2), then the trends become nearly constant for higher α. Similar behaviors are shown on `FriendFeed`, though the increasing trends are less monotone for $k < 30$. On `Instagram-LCC`, the relatively small size and high connectivity of this network makes capital values subject to an average variation of about 15% over the full range of α.

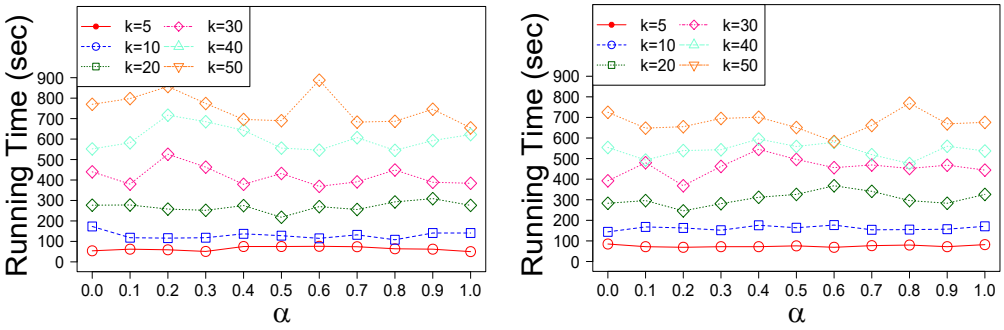
Comparison with TIM+ and KB-TIM. Capital obtained by DTIM methods is shown to be much higher than that of competing methods, on all networks and for various k and *L-perc*. The performance gain is more significant on `FriendFeed`, with average percentage of increment from 9.85% (for *L-perc* = 5%) to 3.49% (*L-perc* = 25%) w.r.t. TIM+, and even larger (from 35% to 59%) w.r.t. KB-TIM. On the two largest networks, as the size of target set increases, a general decreasing trend is observed in the gap between DTIM and TIM+ (resp. KB-TIM) capital values, which might be explained since a larger target set implies that a larger fraction of the entire node set could be reached.

[3] A pseudo-code of the Monte Carlo based algorithm for capital estimation can be found in the Appendix C



(a) G-DTIM (b) L-DTIM

FIGURE 5.6: Time performance (in seconds) for varying k, with $\alpha = 0.5$ and $L\text{-}perc = 25\%$.



(a) G-DTIM (b) L-DTIM

FIGURE 5.7: Time performance (in seconds) for varying k and α, with $L\text{-}perc = 25\%$, on `GooglePlus`.

5.6.2.2 Target activation probabilities

A further stage of evaluation was performed to understand how different settings of α and k impact on the activation probability of nodes targeted by DTIM methods. We regard the activation probability of a node as the number of times it has been activated divided by the number of runs of Monte Carlo simulation for the estimation of capital. We present here a summary of main findings concerning this evaluation, and refer the reader to the Appendix C for detailed results.

For both DTIM algorithms, the activation probability follows a non-decreasing trend as α increases. The likelihood of obtaining high activation probability grows with α, i.e., the amount of target nodes that have high probability of activation increases by increasing α. The analysis of density distributions also puts in evidence that the density peak corresponding to low activation probability is higher for lower values of α, whereas the density corresponding to high activation probability increases for higher values of α. Nevertheless, on the two largest datasets, we also observe that choosing a relatively large k leads a significant portion of target nodes to have mid-high activation probabilities already for $\alpha = 0.1$, thus suggesting that target nodes can be activated even with strongly unbalancing capital with diversity. By contrast, when choosing a small k, little changes in the value of α can significantly impact on the amount of more likely activated target nodes.

5.6.3 Efficiency analysis

Figure 5.6 reports on time performance of G-DTIM and L-DTIM on the various networks, for $5 \leq k \leq 50$ and $\alpha = 0.5$. The execution time of both methods shows a roughly linear increase with k, on all networks. (Note that the `FriendFeed` time series are shown in the figure insets, as they correspond to orders of magnitude higher than for the other networks, due to the larger size of `FriendFeed`). Also, G-DTIM turns out to be slightly faster than L-DTIM, which might be ascribed to the fewer computations of node diversity needed by G-DTIM w.r.t. L-DTIM.

As shown in Figure 5.7 for `GooglePlus` in particular (though similar behaviors also characterize the other networks), varying α with fixed k does not significantly impact on the time performance of both DTIM methods. This would indicate that, for a given seed set size, the methods' effort in computing the global/local diversity as well as the capital contributions in the objective function is not greatly affected by the value of α. Analogous remarks are also drawn for the other settings of *L-perc*.

As regards TIM+ and KB-TIM (results not shown), it comes without surprise that both outperform DTIM methods. For instance, on our largest network (i.e., `FriendFeed`), the execution times of TIM+ (with $\epsilon = 0.1$) are between 6.3 ($k = 50$) and 11.9 ($k = 5$) seconds — note that the increase in runtime by decreasing k is in line with the theoretical and experimental results shown in [189]; yet, KB-TIM execution times are always below 0.7 seconds regardless of *L-perc*, which might also depend on the extremely low number of queries and keywords used by KB-TIM in our setting.

5.7 RIS-based formulation of DTIM

The gap in efficiency shown by our DTIM algorithms w.r.t. the competing RIS-based ones, prompted us to investigate how to adapt RIS-based approximations to our diversity-sensitive, targeted IM problem.

5.7.1 Revisiting RIS theory for the DTIM problem

The reverse influence sampling (RIS) [20] relies on the concept of *reverse reachable* (RR) set. Intuitively, the random RR set generated from G for a randomly selected user u (i.e., the *root* of the RR set) contains the users who could influence u. By generating many random RR sets on different random users, if a user has high potential to influence other users, then s/he will likey appear in those random RR sets. Thus, if a seed set covers most of the RR sets, it will likely maximize the expected spread. Upon this principle, Corollary 1 in [189] states that $\mathbb{E}[F(S)/\theta] = \mathbb{E}[\mu(S)]/n$, where $F(S)$ denotes the number of RR sets covered by the node set S, $\mu(S)$ is the spread of S, θ is the number of RR-sets, and $n = |V|$.[4]

In our setting, every node $v \in V$ is selected as root of an RR-set with probability proportional to its status as target node, i.e., $p(v) = \frac{\ell'(v)}{L_{TS}}$, where $\ell'(v) = \ell(v)$ if $v \in TS$, zero otherwise, and $L_{TS} = \sum_{v \in TS} \ell'(v)$. In the following, we state that for any set of nodes S, the expected value of the fraction of RR sets covered by S is equal to the normalized expected value of the capital associated with the activation of target nodes due to S as seed set.

[4]For the sake of simplicity of notation, we omit to declare random variable symbols when using the expected value operator $\mathbb{E}[\cdot]$.

Proposition 4.

$$\mathbb{E}\left[\frac{F(S)}{\theta}\right] = \frac{\mathbb{E}[C(\mu(S))]}{L_{TS}} \tag{5.18}$$

Proof (sketch). Following notations used in [128], let $p(S \to v)$ denote the probability that v is activated by seed set S. Thus, the expected capital associated with S can be expressed as:

$$\mathbb{E}[C(\mu(S))] = \sum_{v \in V} p(S \to v)\ell'(v) \tag{5.19}$$

By Lemma 2 in [189], the probability that a set S overlaps with an RR set R_v rooted in a node v is equal to the probability that S, when used as a seed set, can activate v, i.e.,

$$p(S \to v) = \Pr[S \cap R_v \neq \emptyset]. \tag{5.20}$$

Therefore, it holds that

$$\begin{aligned} \mathbb{E}[F(S)/\theta] &= \sum_{v \in V} p(v) \Pr[S \cap R_v \neq \emptyset] \\ &= \sum_{v \in V} \frac{\ell'(v)}{L_{TS}} p(S \to v) \\ &= \frac{\mathbb{E}[C(\mu(S))]}{L_{TS}} \end{aligned} \tag{5.21}$$

□

Estimation of the number of RR sets. In [189], the objective is to find a number θ of RR sets such that $\theta \geq \lambda/OPT$, where OPT denotes the maximum expected spread of any size-k seed set, and λ is determined as a function of the size of the graph, k and the approximation factor ϵ. Since OPT is unknown, a lower bound for it must be computed.

Following from Lemma 4 in [189], the expected spread of a randomly sampled node can be expressed in terms of the expected value EPT of the number of edges pointing to nodes in an RR set (*width*), such that $EPT \leq \frac{m}{n} OPT$ holds, with $m = |E|$. We revise this result to state that the expected value of the width of an RR set can be an accurate estimator of the capital associated with any node when randomly selected as a seed.

Proposition 5.

$$(L_{TS}/m)\ EPT = \mathbb{E}[C(\{v\})]\ \leq\ OPT \tag{5.22}$$

Proof (sketch). Let $w(R_u)$ denote the width of an RR set rooted in node u, and $R_u \sim \mathcal{R}$ denote an RR set rooted in node u sampled from the distribution of all RR sets. We have thapp:tdiversity:at:

$$
\begin{aligned}
EPT &= \sum_{u \in V} \frac{\ell'(u)}{L_{TS}} \mathbb{E}_{R_u \sim \mathcal{R}}[w(R_u)] \\
&= \frac{1}{L_{TS}} \sum_{u \in V} \ell'(u) \sum_{R_u \sim \mathcal{R}} \Pr[R_u] \sum_{v \in V} \Pr[v \to u | R_u] \\
&= \frac{1}{L_{TS}} \sum_{R_u \sim \mathcal{R}} \Pr[R_u] \sum_{(v,u) \in E} \ell'(u) \Pr[v \to u | R_u] \qquad (5.23) \\
&= \frac{1}{L_{TS}} \sum_{(v,u) \in E} \mathbb{E}[C(\mu(\{v\}))] \\
&= \frac{m}{L_{TS}} \mathbb{E}[C(\mu(\{v\}))]
\end{aligned}
$$

□

To avoid unnecessarily large values of θ, it is desired to find a lower error bound in terms of the mean of the expected spread of a set S (over the randomness in S and the influence propagation process), denoted as KPT, such that $(n/m)EPT \leq KPT \leq OPT$ holds. To this aim, Lemma 5 in [189] estimates KPT as $KPT = n\mathbb{E}_{R \sim \mathcal{R}}[\kappa(R)]$, taking the average over a set of random RR sets R from the possible world $\mathcal{R}$, where $\kappa(R) = 1 - (1 - \frac{w(R)}{m})^k$ and $w(R)$ is the width of R. Again, we revise this result in our setting:

Proposition 6. *Given a random RR set R, and denoted with TS_R the set of target nodes in R, it holds that*

$$
\widehat{\kappa}(R) = \left[1 - \left(1 - \frac{|TS_R|}{m}\right)^k\right] \frac{\sum_{v \in R} \ell'(v)}{|TS_R|}. \qquad (5.24)
$$

Therefore,

$$
KPT = n\mathbb{E}_{R \sim \mathcal{R}}[\widehat{\kappa}(R)]. \qquad (5.25)
$$

Proof (sketch). Given an RR set R, let us denote with A the event of selecting an edge in G that points to a target node, and with B the event of selecting an edge in G that points to a node in R. The probability of these events are $\Pr[A] = |TS|/m$ and $\Pr[B] = w(R)/m$. The conditional probability of A given B is equal to $\Pr[A|B] = |TS_R|/w(R)$, where symbol TS_R is used to denote the set of target nodes in R. Thus, the probability of selecting an edge pointing to a target node contained in R is $\Pr[A \cap B] = \Pr[A|B] \Pr[B] = \frac{|TS_R|}{w(R)} \cdot \frac{w(R)}{m} = \frac{|TS_R|}{m}$. Given k randomly selected edges, the probability that at least one of these points to a target node in R is $\widehat{\kappa}(R) = 1 - \left(1 - \frac{|TS_R|}{m}\right)^k$. This quantity is finally smoothed by $\frac{\sum_{v \in R} \ell'(v)}{|TS_R|}$, i.e., the average ℓ' value over the target nodes belonging to R. □

5.7.2 Developing RIS-based DTIM algorithms

We sketch here a reformulation of DTIM based on the RIS approach. To this purpose, we start from TIM+ and adapt it to our DTIM problem. This requires four key modifications:

- **M1**: Revise the sampling over the nodes in G.
- **M2**: Modify the KPT estimation procedure (i.e., TIM+'s Algorithm 2).
- **M3**: Modify the refinement of KPT to obtain a potentially tighter lower-bound of OPT (i.e., TIM+'s Algorithm 3).

- **M4**: Modify the node selection procedure (i.e., TIM+'s Algorithm 1) for determining a size-k seed set.

In the following, we elaborate on each of the above points, which overall constitute a 4-stage workflow for the development of RIS-based DTIM methods.

Sampling (M1). As previously discussed, we define a probability distribution over the nodes in G such that the probability mass for each node v is non-zero and proportional to the value of $\ell(v)$ if $v \in TS$, and zero otherwise.

Parameter estimation (M2). The RR sets must be generated in such a way that the roots are sampled from the above defined probability distribution (i.e., the root of any RR set is a target node). Moreover, the original function κ is replaced with Equation (5.24).

Parameter refinement (M3). Starting from the set $\mathcal{R}'$ of all RR sets produced to estimate KPT, the size-k seed set S' is generated by selecting those nodes that, while covering RR sets in $\mathcal{R}'$, maximize the capital w.r.t. $\mathcal{R}'$. More specifically, each RR set in $\mathcal{R}'$ is associated with a score equal to the value of ℓ of its root node, and every node is associated with a score equal to the sum of RR-set-scores the node belongs to. In the main loop, at each of the k iterations, the node v with maximum score is identified and added to S', all RR sets covered by v are removed from $\mathcal{R}'$, and the node scores are recomputed.

Once computed S', a new set $\mathcal{R}''$ of RR sets is generated and used to derive $\bar{\mathcal{F}}$, which contains the root nodes of all RR sets in $\mathcal{R}''$, and $\mathcal{F}$, which is the subset of root nodes of RR sets that have non-empty overlap with S'. Next, we compute the fraction of capital associated with $\mathcal{F}$, i.e., $f = \sum_{v \in \mathcal{F}} \ell'(v) / \sum_{v \in \bar{\mathcal{F}}} \ell'(v)$. Quantity f is finally exploited to derive the new lower-bound analogously to the last two instructions in TIM+'s Algorithm 3.

Node selection (M4). Let us first consider the case in which the diversity function is discarded from the DTIM objective function. The node selection procedure turns out to be analogous to the first step described in **M3**, where the number θ of RR sets to generate is computed based on the refined KPT. In the general case, the node selection procedure needs to also include the global/local diversity values when scoring the nodes w.r.t. the RR sets they cover. We provide here an informal description of the essential steps to perform.

Let $\mathcal{R}_v$ denote the set of RR sets rooted in v. Upon this, we build a tree index $\Lambda(v)$, with root v, by aggregating all live-edge paths reaching v. Note that the tree is constructed in a backward fashion; also, every node other than v has at most one incoming edge, and it could appear in many paths and at different distance from v.

Let us first consider the global diversity of a node in $\mathcal{R}_v$. The boundary set of $\Lambda(v)$ is the multiset of all leaf nodes in the tree. The *RR-global-diversity* of a node u in $\Lambda(v)$ is determined as the mean of its global diversity values by possibly considering the multiple occurrences of u as leaf. By averaging the RR- global-diversity values over all trees in which node u appears, we compute the *total RR-global-diversity* of u. To compute the *RR-local-diversity*, we need to consider each *level* of $\Lambda(v)$ at a time, and hence the boundary set of each subtree resulting from truncating $\Lambda(v)$ at a given distance from v. We then average the scores of a node u over all trees in which u appears to have the *total RR-local-diversity* of u.

Finally, the total RR-diversity of a node is linearly combined with the corresponding capital score, in order to drive the search for the node with maximum DIC to be identified at the k-th iteration of the node selection procedure.

5.8 Chapter notes

We presented a novel targeted IM problem in which the objective function is defined in terms of spreading capability and topology-based diversity w.r.t. the target users. We proved that the proposed objective function is monotone and submodular, and developed two alternative algorithms, L-DTIM and G-DTIM, to solve the problem under consideration. Significance and effectiveness of our algorithms have been assessed, also in comparison with baselines and state-of-the-art IM methods, using publicly available, real-world network graphs. We have also provided theoretical foundations to develop RIS-based DTIM methods.

As future research, it would be interesting to investigate diversity notions based on boundary spanning principles that might rely on community detection solutions; other opportunities in this regard would certainly come from the integration of side information representing user profiles. We also plan to evaluate the RIS-DTIM method, which promises to overcome the efficiency issues of the current DTIM methods. Finally, it is worth noting that our proposed approach is versatile, as it can easily be generalized not only to other cases of user engagement (for example, introducing newcomers to a community), but also to any other application of targeted IM in which accounting for diversity of users based on their relationships/interactions with other users, is beneficial to the enrichment of influence propagation outcome with effects of varied social capital. In this respect, we can envisage further developments from various perspectives, including human-computer interaction, marketing, and psychology.

Chapter 6

Attribute-based Diversity-sensitive Targeted Influence Maximization

In the previous Chapter 5 we have highlighted the importance and significance of accounting for diversity measures in the context of influence maximization.

As it is also emerged from several other studies, our analysis confirmed that embedding diversity into knowledge discovery activities often leads to the identification of more novel, more meaningful and broader patterns.

Following a common thread with the previous chapter, here we address the same targeted influence maximization problem, while we consider diversity from an orthogonal perspective. That is, instead of measuring diversity with respect to the network topology, here we investigate the opportunity of defining diversity as a function of some side-information available on the user level.

More formally, in this chapter we propose the integration of a categorical-based notion of seed diversity into the objective function of a targeted influence maximization problem.

In this respect, we assume that the users of a social network are associated with a categorical dataset where each tuple expresses the profile of a user according to a predefined schema of categorical attributes.

Upon this assumption, we design a class of monotone submodular functions specifically conceived for determining the diversity of the subset of categorical tuples associated with the seed users to be discovered. This allows us to develop an efficient approximate method, with a constant-factor guarantee of optimality. More precisely, we formulate the *attribute-based diversity-sensitive targeted influence maximization* problem under the state-of-the-art reverse influence sampling framework, and we develop a method, dubbed ADITUM, that ensures a $(1 - 1/e - \epsilon)$-approximate solution under the general triggering diffusion model.

Extensive experimental evaluation based on real-world networks as well as synthetically generated data has shown the meaningfulness and uniqueness of our proposed class of set diversity functions and of the ADITUM algorithm, also i n comparison with methods that exploit numerical-attribute-based diversity and topology-driven diversity in influence maximization.

It should be also noted that the ADITUM algorithm represents a major advance over the approach described in Chapter 5, as it provides a stronger flexibility – it can easily incorporate different notions of diversity – and it delivers better performance both in terms of efficacy and efficiency.

6.1 Introduction

Online social networks (OSNs) are a suitable environment for propagating influence between connected individuals, so that they have become the most profitable channel for a variety of purposes related to viral marketing, advertisement campaigns, news propagation, and many others. In this regard, a classic optimization problem is *influence maximization* (IM), which is to discover a set of seeds, i.e., initial influencers or early-adopters, that can maximize the spread of information through the network (e.g., advertising of a product) [33, 97]. The basic principle is that, by finding the most effective users to endorse an idea/product/information and to influence other users in the network, a chain reaction of influence can be activated and driven by a "word-of-mouth" effect, in such a way that with a very small marketing cost (i.e., the number of initial influencers) a very large portion of the network will be reached. The extent of this portion can conveniently be limited to a selection of users depending on predetermined constraints, such as based on strategic location or interest in contents that are being diffused; in fact, in many practical scenarios, companies want to tailor their advertisement strategies in order to address only selected OSN-users as potential customers. This is the perspective adopted in the context of *targeted IM* which is also a focus of this work.

Maximizing the spread of information is directly related to an a-priori specified budget as the number of seeds. In a more complex "budgeted" scenario of profit maximization, each of these seeds could be associated with a different cost to engage it as early-adopter, which would imply to account for these costs as constraints in a (targeted) IM problem.

Moreover, we have the opportunity to make the seed-selection step in IM more sensitive to user features.

In particular, we believe that the "influence potential" of the seeds being selected can be well-explained in terms of *diversity* that may characterize the seeds. Intuitively, influencers that are diverse to each other according to certain features (e.g., age, gender, socio-cultural aspects, preferences) might have more opinions, experiences, and perspectives to bear on the influence propagation process. As a consequence, identifying a set of seed users that have as more different characteristics as possible from each other, will be helpful to enhance the marketing or information-propagation campaign strategies to engage the target users. Indeed, before taking any decision for active involvement in a given propagation scenario, every user in the network would like to acquire enough information, possibly from different perspectives. Therefore, by identifying the most diverse seed users, the triggering stimuli will also be diversified, and since diverse individuals tend to connect to many different types of members, the likelihood of influencing the targets would be higher.

Accounting for diversity in influence propagation has important implications, also from an ethical viewpoint. In fact, favoring diversity in selecting the early-adopters as well as in targeting the users to reach is strictly related to being exposed to diverse opinions: as previously argued in [6], the latter can significantly contribute to disrupt information bubbles or echo chambers — where pre-existing opinions are maintained and reinforced — thus raising the level of democratic debate.

Despite the importance of leveraging diversity for improved solutions to IM problems, it comes to our surprise that relatively few studies have considered diversity in such a context. Some work has focused on understanding relations between diversity, or fairness, and effectiveness/efficiency in the spreading ability [10, 60, 88, 177]. Node diversity into the IM task was first introduced by Tang et al. [185], where numerical

attributes reflecting user preferences on some predefined categories (e.g., movie genres) are considered to address a generic IM task. In [26], we originally defined an IM problem that is both targeted and diversity-sensitive for the seed selection, however, it only considers specific notions of diversity that are driven by the topology of the information diffusion graph. Also, [6] studied diversity of exposure, which relies on an item-aware propagation setting.

Contributions. In this work, we aim to advance research on IM by formulating a novel targeted IM problem that accounts for *categorical attribute-based diversity of the seeds* to be identified. Our contributions are summarized as follows.

- We propose the **A***ttribute-based* **DI***versity-sensitive* **T***argeted Infl***U***ence* **M***aximization* problem, dubbed ADITUM.[1] A key aspect is that the set of nodes in the network is associated with a categorical dataset, which would represent the node profiles according to a schema of categorical attributes and corresponding values.

- We provide conceptually different notions of diversity that are able to reflect the variety in the categorical attributes and their values that characterize the seeds being discovered. Remarkably, we design a class of nondecreasing monotone and submodular functions for categorical diversity, each of which also has the nice property of enabling incremental computation of a node's marginal gain when added to the current seed set. To the best of our knowledge, we are the *first to propose a formal systematization of approaches and functions for determining submodular set diversity in influence propagation* and related problems in information networks.

- We design our solution to the ADITUM problem under the Reverse Influence Sampling (RIS) paradigm [20, 189], which is widely recognized as the state-of-the-art approach for IM problems. One challenge that we address is revisiting the RIS framework to deal with both the targeted nature and the diversity-awareness of the ADITUM problem.

- We develop the ADITUM algorithm, which returns a size-k seed set ensuring an approximation ratio of $(1 - 1/e - \epsilon)$, with high probability (at least $1 - |V|^{-1}$), in $O(\epsilon^{-2}k(|E|+|V|)\log|V|)$ time (with ϵ sampling error) on a diffusion graph with $|V|$ nodes and $|E|$ edges, under the Triggering model, which is a general diffusion model adopted by most existing work in monotone submodular IM.

- We thoroughly analyzed our proposed diversity functions on synthetically generated datasets, and

  we experimentally evaluated ADITUM on publicly available network datasets, three of which were used in a user engagement context, one in community interaction, and the other one in recommendation.

  We make this choice so we can compare ADITUM against the methods in [185] and [26].

Plan of the chapter. The remainder of this chapter is organized as follows. Section 6.2 discusses related work, with emphasis on targeted IM and diversity-aware IM. Section 6.3 formalizes the information diffusion context model, the objective function, and the optimization problem under consideration. Section 6.4 presents our study on

[1] Latin term for *access*, *admission*, *audience*.

monotone and submodular diversity functions for categorical data modeling the profiles of nodes in a network. Section 6.5 describes our proposed approach and algorithm for the ADITUM problem. Sections 6.6 and 6.7 contain our experimental evaluation methodology and results, respectively. In Section 6.8, we provide our conclusions and pointers for future research.

6.2 Related work

Given a weighted directed graph, an information diffusion model, and a positive integer k, the problem of IM is to find a seed set S of size k that maximizes the expected number of active nodes at the end of the diffusion process started from S. The foundations of IM as an optimization problem were initially posed by Kempe et al. in their seminal work [97], and rely on two main findings. The first one is the intractability of the problem in its two sources of complexity, i.e., to discover a k-sized seed set that maximizes the expected spread, and to estimate the expected spread of the final active node-set. The second result is the possibility of designing an approximate greedy solution with theoretical guarantee. More precisely, despite IM red being proven to be NP-hard under most stochastic diffusion models, such as Independent Cascade (IC) and Linear Threshold (LT) models, a greedy framework can be developed to achieve $(1 - 1/e)$ approximation guarantee, provided that the influence function is *nondecreasing monotone* and *submodular* [152], like in the cases of IC and LT models. Intuitively, monotonicity means that adding more nodes to a seed set does not reduce its influence spread, while submodularity can be understood as diminishing marginal gains of the influence spread. However, since the expected spread cannot efficiently be evaluated exactly, the solution proposed in [97] resorts to a Monte Carlo based estimation, which however is a bottleneck preventing the IM method to scale on very large graphs. For this reason, many efforts have been devoted to address the scalability issue in the Monte Carlo based greedy algorithm, mostly by reducing the number of Monte Carlo estimations [114]. Alternatively, proxy-based methods have been developed to avoid running Monte Carlo simulations, by estimating the influence spread of the seed set through a reduced diffusion context; however, without ensuring theoretical approximation guarantee. Example methods following a proxy-based approach are MIA/PMIA [39], LDAG [41], and SimPath [72].

A breakthrough study that overcomes the efficiency bottleneck of the simulation based methods, while preserving the theoretical approximation guarantee, is proposed by Borgs et al. [20], which introduces the Reverse Influence Sampling (RIS) framework for IM. The key idea is that the expected spread can be estimated by taking into account a number of pre-computed sketches, i.e., realizations drawn from the distribution induced by both the diffusion model and the influence graph.

This breakthrough result paved the way for a variety of sketch-based algorithms. Tang et al. in [189] are the first to design upon the findings in [20] a practically efficient solution, TIM/TIM+, whose main improvement over RIS consists in the ability of keeping the same theoretical bound as [20] with significantly fewer sketches, bounded by the influence of the unknown optimal set (OPT). TIM/TIM+ can perform orders of magnitude faster than the greedy algorithm, overcoming the bottleneck in the computation of the expected spread by exploiting the RIS technique. Since then, other methods have followed, such as IMM [188], SSA [154], BCT [153], and TipTop [123], which share the common motif of estimating OPT with a fewer number of sketches. Also, [138] generalizes the theoretical results in [20, 189] to any diffusion

TABLE 6.1: Categorization of IM related works discussed in this article.

problem	approach	methods/references
Influence Maximization	Simulation-based	Greedy [97], CELF [114]
	Proxy-based	MIA/PMIA [39], LDAG [41], SimPath [72]
	Sketch-based	RIS [20], TIM/TIM+ [189], [138], IMM [188], SSA [154], [98], BCT [153], TipTop [123]
Targeted Influence Maximization	Single target	[76, 77, 204]
	Topic-dependent target	[12, 32, 111, 127]
	Topic-dependent diffusion	[50, 92, 109, 121, 126, 158, 212]
	Benefit-aware	[78, 140, 153, 187]
	Network-structure-aware	[21, 27, 42, 124, 165, 173, 175, 209, 210]
	Mixed	[86, 121], [9, 155]
Diversity in Influence Propagation	Fairness in spreading	[177]
	Diversity in diffusion models	[10, 88]
	Diversity-aware IM	[6, 26, 185]

model with an equivalent live-edge model of the diffusion graph. Furthermore, the sketch-based approach has been extended in [98] to deal with IM in dynamic graphs.

In the following, we will focus our discussion on variants of IM problems and approaches that consider notions of target users and criteria for seed selection, followed by the current status of research on diversity in IM contexts;

Table 6.1 provides a summary of the IM related works discussed in this article.

For broader and more complete views on the IM topic, the interested reader can refer to recent surveys, such as [125, 161, 178].

Targeted influence maximization. Research on targeted IM has gained attention in recent years. Early studies have focused on the special case of a single selected target-node [76, 77, 204]. By contrast, more general targeted IM methods, like ours, aim at maximizing the probability of activating a target set of arbitrary size by discovering a seed set which is neither fixed and singleton nor has constraints related to the topological closeness to a fixed initiator.

Targeted IM methods typically account for information on contents and/or users' characteristics; depending on the type of such characteristics, targeted IM methods can be divided into *topic-aware*, *benefit-aware*, and *network-structure-aware* methods.

The first category is motivated by the fact that a user is likely to be influenced by messages (e.g., advertisements) being diffused that are relevant to information that match the user's interests or preferences. Therefore, a user whose interests match with the query topics or keywords might be regarded as a target node, and the goal is to maximize the spread among such target users [12, 32, 111, 127]. Also, content information can be incorporated into the diffusion process or the influence probability estimation. For instance, in [109], a family of probabilistic diffusion models is proposed to exploit vectors of features representing the content of information to be diffused and the profile of users. In [212], the IC model is adapted to accommodate user preferences, which are learned from a set of users' documents labeled with topics. User activity, sensitivity, and affinity are considered in [50] to define node features, which are then used to adjust the influence between any two users. In the conformity-aware cascade model [121], the influence probability from node u to node v is computed based on a sentiment analysis approach and proportionally to the product of u's influence and v's conformity, where the latter refers to the inclination of a node to be influenced by others. A group-based influence maximization method is proposed in [126] to solve the

IM problem over the conformity-aware diffusion model which utilizes different types of conformity behaviors (where people conform to the opinions and actions of others by submitting to perceived group pressure) based on user profiles and group profiling. In [92], the targeted IM problem is studied in the context of user engagement, whereby a node is regarded as target on the basis of its social capital. In [158], under a non-submodular framework, the goal is to find a k-sized seed set to initiate the diffusion that will maximize the spread among the nodes with the target-label, while keeping the diffusion among the nodes with the non-target label within a pre-specified threshold.

The second category of targeted IM methods refers to *profit maximization* problems, where the influence spread is seen as the *benefit* gained by viral marketing and the *cost* for seed selection is the amount to pay for viral marketing. In this respect, the users in a social network are likely to bring different amounts of benefit if activated, and have different costs for seed selection. For instance, [140] proposed pricing strategies to optimize the profit return of viral marketing, through hill-climbing heuristics. More recently, other works focused on approximation solutions for profit maximization with theoretical guarantee [153, 187]. Also, a perspective under deterministic linear threshold model is taken in [78].

Leveraging the mesoscale structure of the information-diffusion network allows for the development of several strategies to drive the seed selection and targeted IM. Most existing methods of this category exploit the availability of a community structure or graph partitioning solution (e.g., [21, 27, 42, 124, 165, 173, 175, 210]. In some cases, the network structure is also exploited in combination with (community-level) topic interests, conformity-aware (e.g., [86, 121]), or cost constraints (e.g., [9, 155]). Coreness is considered in [209] for estimating nodes' influence and developing a simulated annealing based algorithm for IM.

Recently, in [25], the authors extensively explore where the set of influential nodes extracted by state-of-the-art IM methods are located in a network w.r.t. different notions of graph decomposition.

Note that the large majority of the aforementioned works are concerned with the development of heuristics for IM (even under a non-submodular framework, in some cases), while we are interested in designing a targeted IM solution with approximation guarantee. Moreover, all of these methods discard a major aspect in our work, that is, an explicit notion of categorical set diversity for the seed nodes.

Diversity in influence maximization. Diversity has been recognized as a key-enabling dimension in several tasks that are relevant in information management and filtering, such as web searching, ranking, and recommendation (e.g., [52, 70, 170, 202]). It also relates to novelty and serendipity, for instance to improve user satisfaction in content recommendation. Moreover, diversity has attracted attention in machine learning as concerns the development of fair strategies to control the bias and its effect on the outcomes of supervised learning tasks (e.g., [207, 208]).

Understanding how to design IM methods in a fair way is addressed in [177]. Fairness in seed selection (resp. in outreach) is meant as choosing (resp. reaching) nodes that are uniformly distributed w.r.t. the available communities in the network. Upon an analysis based on the IC model in a network that follows preferential-attachment with a two-community homophily model, the study in [177] argues that a strategic, feature-aware heuristic is fairer and more efficient than feature-blind methods.

However, a relatively little amount of work has been devoted to integrating diversity in the objective function of IM problems. One of the earliest attempts is provided in [10], which extends the IC model to account for the structural diversity of nodes' neighborhood, however without addressing an optimization problem. Other

works have studied relations between diversity and spreading ability, but focusing on a single node in a network [88].

Tang et al. [185] proposed the first study on diversity-aware IM, where a linear combination of the expected spread function and a numerical-attribute-based diversity is maximized by means of heuristic search strategies, defined upon classic centrality measures. In [26], we formulated the topology-driven diversity-sensitive targeted IM problem, dubbed DTIM, with an emphasis on maximizing the social engagement of a given network. The provided solution, built upon the SimPath method [72], relies on the LT model. It should be noted that, although the optimization problem presented in this work is similar to the one in [26], here we provide different formulation and algorithmic solution since, unlike DTIM, (i) our proposed ADITUM builds on state-of-the-art approximation methods for IM, and (ii) it is designed to handle different notions of attribute-based diversity. In Sections 6.6–6.7, we present a comparative evaluation with the methods in [185] and [26].

Another related work is that proposed in [6], where it is assumed that the diverse viewpoints (to which users may be exposed) are represented by a number of message items propagating through the network; moreover, users are supposed to have an individual predisposition towards an item, which impacts on the probability of further disseminating the item. Under this setting, the goal is to find a small number of seeds and items so that the overall diversity of exposure of all users is maximized. Again, this problem is however different from ours, both in terms of information propagation setting (i.e., we do not rely on messages, neither on concepts of item leanings and user leanings) and objective function (i.e., we want to maximize the overall diversity within the seed set rather than the sum of diversity exposure levels of all nodes resulting from a predetermined itemset assignment).

6.3 Problem statement

Representation model Given a social network graph $G_0 = \langle V, E \rangle$, with set of nodes V and set of edges E, let $G = G_0(b, t) = \langle V, E, b, t \rangle$ be a directed weighted graph representing the *information diffusion* context associated with G_0, with $b : E \to (0, 1]$ edge weighing function, and $t : V \to (0, 1]$ node weighing function.

Function t determines the status of each node as *target*, i.e., a node toward which the information diffusion process is directed. Given a user-specified threshold $\tau_{TS} \in [0, 1]$, we define the *target set* TS for G as:

$$TS = \{v \in V | t(v) \geq \tau_{TS}\}.$$

Function b corresponds to the parameter of the *Triggering* model [97] which, in line with several existing studies on IM, is also adopted here as information diffusion model. Under this model, each node chooses a random subset of its neighbors as *triggers*, where the choice of triggers for a given node is independent of the choice for all other nodes. If a node u is inactive at a given time and a node in its trigger set becomes active, then u becomes active at the subsequent time. Notably, Triggering has an equivalent interpretation as "reachability via live-edge paths", such that an edge (u, v) is designated as live when v chooses u to be in its trigger set. Therefore, $b(u, v)$ represents the probability that edge (u, v) is live. Linear Threshold and Independent Cascade [97] are special cases of Triggering with particular distributions of trigger sets.

Note also that function b and t are usually defined as data-driven. We will discuss possible instances of both functions in Section 6.6.2.

Objective function The objective function of our targeted IM problem is comprised of two functions. The first one, denoted as $C(\cdot)$, is determined as the cumulative amount of the scores associated with the target nodes that are activated by a given seed set. Following the terminology in [26], we call this function social capital, or simply *capital*, which is defined as

$$C(\mu(S)) = \sum_{v \in \mu(S) \cap TS} t(v) \tag{6.1}$$

where $\mu(S)$ denotes the set of nodes that are active at the end of the diffusion starting from S.

The second term in our objective function, denoted as $div(\cdot)$, is introduced to determine the *diversity* of the nodes in any subset of V. As previously mentioned, our approach is to measure diversity in terms of a-priori knowledge provided in the form of symbolic values corresponding to a predetermined set of *categorical attributes*. In Section 6.4, we provide a class of diversity functions for categorical datasets.

We now formally define our proposed problem of targeted IM, **A***ttribute-based* **DI***versity-sensitive* **T***argeted Infl***U***ence* **M***aximization* (ADITUM).

Definition 7. (Attribute-based Diversity-sensitive Targeted Influence Maximization) *Given a diffusion graph $G = \langle V, E, b, t \rangle$, a budget k, and a threshold τ_{TS}, find a set $S \subseteq V$ with $|S| \leq k$ of seed-nodes such that*

$$S = \underset{S' \subseteq \mathcal{V} \ s.t. \ |S'| \leq k}{\operatorname{argmax}} \alpha \times C(\mu(S')) + (1 - \alpha) \times div(S') \tag{6.2}$$

where $\alpha \in [0, 1]$ is a smoothing parameter that controls the weight of capital $C(\cdot)$ w.r.t. diversity $div(\cdot)$. □

The problem in Definition 7 preserves the NP-hard complexity of the IM problem. However, as for the classic IM problem, if we are able to design an objective function that is monotone and submodular, then the output of a greedy solution provides a $(1 - 1/e)$-approximation guarantee w.r.t. the optimal solution [152]. To this aim, we need to ensure that Equation 6.2 is a linear combination of two monotone and submodular functions. Recall that, given a finite set Ω, any function $f : 2^{\Omega} \mapsto \mathbb{R}$ is said to be *monotone* iff $f(S) \leq f(S')$ for any $S \subseteq S' \subseteq \Omega$, and *submodular* iff $f(S \cup \{x\}) - f(S) \geq f(S' \cup \{x\}) - f(S')$ for any $S \subseteq S' \subseteq \Omega$ and $x \in \Omega \setminus S'$.

Monotonicity and submodularity of the capital function $C(\cdot)$ was previously demonstrated in [26]. In the next section, we provide our definitions of $div(\cdot)$.

6.4 Monotone and submodular diversity functions for a set of categorical tuples

We assume that the nodes in the social network graph $G_0 = \langle V, E \rangle$ are associated with side-information in the form of symbolic values that are valid for a predetermined set of categorical attributes, or *schema*, $\mathcal{A} = \{A_1, \ldots, A_m\}$. For each $A \in \mathcal{A}$, we denote with dom_A its domain, i.e., the set of admissible values known for A, and with dom the union of attribute domains. Moreover, we define $val_A : V \mapsto dom_A$ as a function that associates a node with a value of A. For any $S \subseteq V$, we will also use symbols

$dom_A(S)$ and $dom(S)$ to denote the subset of values in dom_A, resp. dom, that are associated with nodes in S.

Given the schema $\mathcal{A}$, we will refer to the categorical tuple associated to any $v \in V$ as the *profile* of node v, and to the categorical dataset for all nodes in V as the *profile set* of V. We will use symbol $\mathcal{A}[v]$ to denote the profile of v and symbol $\mathcal{D}_S$ to denote the profile set of nodes in $S \subseteq V$. Note that $\mathcal{D}_S$ is a multiset such that $\mathcal{D}_S = \bigcup_{v \in S} \mathcal{A}[v]$, and any $\mathcal{A}[v]$ is generally regarded as a sparse vector, as it could contain *missing values* for some attributes; i.e., if we denote with $\perp$ a missing attribute value, $\mathcal{A}[v] = \langle val_{A_1}(v) \vee \perp, \ldots, val_{A_m}(v) \vee \perp \rangle$. Moreover, we will use symbol $|\mathcal{A}[v]|$ to denote the actual length of $\mathcal{A}[v]$ as the number of attribute values contained in the profile.

General requirements Given our setting of an information diffusion graph $G = G_0(b,t) = \langle V, E, b, t \rangle$ associated with G_0, here we define a class of functions div that, for any $S \subseteq V$ with associated $\mathcal{D}_S$, satisfy the following requirements:

- $div(S)$ defines a notion of diversity of nodes in S w.r.t. their categorical representation given in $\mathcal{D}_S$;
- $div(S)$ must be *nondecreasing monotone and submodular*; hereinafter, we will use the more simple term "monotone and submodular";
- for any $v \in V \setminus S$, the marginal gain $div(S \cup \{v\}) - div(S)$ should be computed efficiently;
- $div(S)$ should be *meaningful*, in terms of ability in capturing the subtleties underlying the variety of node profiles according to their categorical attributes and values.

6.4.1 Challenges in defining set diversity functions

Before providing our definitions of diversity functions in Sects. 6.4.2–6.4.5, here we mention some of the negative outcomes that were drawn by an attempt of devising apparently simple and intuitive approaches based on *attribute-wise* functions as well as based on *profile-wise* functions. Eventually, this demonstrates their unsuitability as diversity functions for the task at hand, as they do not satisfy one or more of the above listed general requirements.

Let us begin with attribute-wise functions. Given $A \in \mathcal{A}$ and $S \subseteq V$, one simple approach would be to compute the *number of unique values* admissible for A that occur in $\mathcal{D}_S$, normalized by the size of S; however, this coarse-grain function is not only unable to characterize the variety of nodes in terms of repetitions of the different values of the attribute under consideration, but also it is not nondecreasing monotone since it decreases by adding nodes with identical values of the attribute. The desired properties of monotonicity and submodularity could be satisfied by just counting the number of unique values of attribute in $\mathcal{D}_S$, however at the cost of a further worsening in meaningfulness, thus obtaining a useless notion of diversity.

An alternative approach would be to aggregate *pairwise distances* of the node profiles w.r.t. a given attribute. For instance, we could count the (normalized) number of mismatchings over each pair of nodes in a set; however, it is easy to prove that the derived function will not be submodular in general.

Let us now extend to calculating pairwise distances of the node profiles over the entire schema. In this regard, we could consider a widely-applied measure for computing the distance between two sequences of symbols, namely *Hamming distance*.

However, for different varying set-size-based normalization schemes, this might result in a function that is not submodular or even not monotone. Alternatively, we could consider a standard statistic for dissimilarity of finite sample sets, namely *Jaccard distance*. This is defined as the complement of Jaccard similarity, that is, for any two sets, substracting from 1 the ratio between the size of the intersection and the size of the union of the sets. (In our context, a sample set corresponds to a categorical tuple, i.e., a node profile.) Again, the resulting function will not ensure submodularity. The interested reader can refer to the ***Appendix*** for analytical details of the aforementioned functions and relating examples that show their unsuitability as monotone submodular diversity functions.

6.4.2 Attribute-wise diversity

In this section, we discuss the first of our proposed diversity functions, which is *attribute-wise*. We consider a notion of set diversity that builds on the variety in the amount and type of categorical values that characterize the nodes in a selected set. In particular, we consider a linear combination of the contributions the various attributes provide to the diversity of nodes in a set.

Definition 8. *Given a set of categorical attributes $\mathcal{A} = \{A_1, \ldots, A_m\}$ and associated profile set $\mathcal{D}$ for the nodes in a graph $G_0 = \langle V, E \rangle$, we define the* attribute-wise diversity *of any set $S \subseteq V$ as:*

$$div(S) = \sum_{j=1..m} \omega_j \; div_{A_j}(S) \tag{6.3}$$

where $div_{A_j}(S)$ evaluates the diversity of nodes in S w.r.t. attribute A_j, and ω's are real-valued coefficients in $[0, 1]$, which sum up to 1 over $j = 1..m$. □

To meet the monotonicity, submodularity, meaningfulness and efficiency requirements, we provide the following attribute-specific set diversity function.

Definition 9. *Given a categorical attribute A, with domain of values dom_A, and node set $S \subseteq \mathcal{V}$, we define the attribute-specific set diversity for S as:*

$$div_A(S) = \sum_{a \in dom_A(S)} \sum_{i=1}^{n_a} \frac{1}{i^\lambda} \tag{6.4}$$

where n_a is the number of nodes in S that have value a for A, and $\lambda \geq 1$. □

One nice property of the function in Equation 6.4 is that the contribution of a node to the set diversity, i.e., *the node's marginal gain* can be determined at constant time, thus without recomputing the set diversity from scratch. This holds based on the following fact.

Fact 1. *The marginal gain of adding a node v to S is equal to*

$$\sum_{j=1..m} \omega_j \sum_{a \in dom(A_j) \; \wedge \; a \in \mathcal{A}[v]} (n_a + 1)^{-\lambda},$$

where n_a is the number of nodes in S that have value a for A, and $\lambda \geq 1$.

Proposition 7. *The attribute-wise diversity function defined in Equation 6.3 is monotone and submodular.*

Proof. Function $div(S)$ in Equation. 6.3 is monotone and submodular provided that $div_A(S)$ in Equation 6.4 is such as well, for any choice of $A \in \mathcal{A}$ and setting of coefficients ω, since $div(S)$ is a linear combination of functions $div_A(S)$ with nonnegative weights. Monotonicity of Equation 6.4 is trivially satisfied. As concerns submodularity, let us assume $\lambda = 1$ without loss of generality. Note that the inclusion of a node u into S is $1/k_1$, with k_1 equal to the size of $S' \subseteq S$ such that, for each $v \in S'$, it holds that $val_A(v) \equiv val_A(u)$; moreover, the inclusion of node u into T ($S \subseteq T$) is $1/k_2$, with k_2 equal to the size of $T' \subseteq T$ such that for each $v \in T'$, $val_A(v) \equiv val_A(u)$. Since $S \subseteq T$, it holds that $k_2 \geq k_1$, or $1/k_1 \geq 1/k_2$, which concludes the proof. □

Lemma 2. *Given a set S and a categorical attribute A, let n_a denote the number of nodes in S whose profile contains the value $a \in dom(A)$. For any*

$$S = \underset{S' \subseteq \mathcal{V} \ s.t. \ |S'| \leq k}{\operatorname{argmax}} div_A(S')$$

it holds that, for every pair of categorical values $a, a' \in dom(A)$, if $n_a - n_{a'} > 1$, then there cannot be a node v in $V \setminus S$ such that $a' \in \mathcal{A}[v]$.

Proof. Assume by contradiction that there exists a set S that maximizes div_A (for any $A \in \mathcal{A}$) such that $M_A - m_A > 1$. Without loss of generality, assume $M_A = m_A + 2$ and $\lambda = 1$. Let $a^{(M)}$ and $a^{(m)}$ denote the categorical values corresponding to M_A and m_A, respectively. It is easy to note that if we remove a node with profile containing $a^{(M)}$, resp. $a^{(m)}$, then div_A will decrease by a $\delta^- = 1/(M_A)$ factor, resp. increase by a $\delta^+ = 1/(m_A + 1)$ factor. Since $\delta^- < \delta^+$, the diversity value is increased, therefore S cannot be the optimal solution, which proves our statement. □

We also observed that the *theoretical maximum* value reached by Equation 6.3 depends only on the budget k, as provided by the following result.

Proposition 8. *Given the set of categorical attributes $\mathcal{A} = \{A_1, \ldots, A_m\}$, m-real valued coefficients $\omega_j \in [0, 1]$ ($j = [1..m]$), and a budget k, the theoretical maximum value for Equation 6.3 is function of k and determined as ($d_j \triangleq |dom_{A_j}|$):*

$$div^*[k] = \sum_{j=1}^{m} \omega_j \left(d_j \sum_{i=1}^{k/d_j} \frac{1}{i^\lambda} + \frac{k \bmod d_j}{(1 + \frac{k}{d_j})^\lambda} \right) \tag{6.5}$$

*Proof (sketch).*Equation 6.5 can be derived based on the observation that the maximum possible value achievable w.r.t. a budget k is obtained when the categorical values are equally distributed among the k nodes. Without loss of generality, let us consider the case with one categorical attribute A. If we need to select k nodes, one at a time, the best choice corresponds to select the node with value $a^* = \operatorname{argmin}_{a \in dom_A(S)} n_a$, as it yields the maximum marginal gain. It straightforwardly follows that, by adopting this strategy, a set S can be produced to satisfy the requirement stated in Lemma 2 for the maximization of Equation 6.3. □

6.4.3 Distance-based diversity

In Section 6.4.1, we showed that an aggregation by sum of the profile-wise Hamming distances does not generally ensure submodularity or even monotonicity. Given the

profiles of two nodes u, v, the *Hamming distance* is defined as:

$$dist_H(u,v) = \sum_{j=1}^{m} \mathbb{1}[val_{A_j}(u) \neq val_{A_j}(v)], \tag{6.6}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function.[2]

To design a set function that satisfies both the properties of monotonicity and submodularity, we borrow the notion of *Hamming ball* introduced in [164], i.e., a set of objects each having a Hamming distance from a selected object-center at most equal to a predefined threshold, or *radius*. Our definition of Hamming ball for a given node in the network takes also into account the *influence range* of the node, i.e., all the nodes reachable starting from the node at the center of the "ball". Formally, given $v \in V$ and a positive integer ξ, we define the Hamming ball as:

$$\mathcal{B}_v^{\xi} = \{u \mid u \in \mathrm{IR}(v) \wedge dist_H(u,v) \leq \xi\}, \tag{6.7}$$

where $\mathrm{IR}(v) \subseteq V$ denotes the set of nodes u for which there exists a path connecting v to u. Restricting the Hamming balls to the center's influence range is beneficial in terms of efficiency, and also licit since only the Hamming balls that are meaningful in an influence spread scenario might be considered.

Definition 10. *Given a set of categorical attributes $\mathcal{A} = \{A_1, \ldots, A_m\}$ and associated profile set $\mathcal{D}$ for the nodes in a graph $G_0 = \langle V, E \rangle$ and a radius ξ, we define the* Hamming-based diversity *of any $S \subseteq V$ as:*

$$div(S) = |\bigcup_{v \in S} B_v^{\xi}| \tag{6.8}$$

□

Intuitively, since similar nodes have overlapping Hamming balls, by taking the union in Equation 6.8 we implicitly force the selection of seeds so that nodes are as different as possible from each other. Moreover, one nice effect of accounting for the influence reachable set in computing the Hamming balls, is that we inherently favor the selection of nodes with higher connectivity, hence large influence range, which is a particularly valuable aspect for our problem.

The above defined function has the property of allowing an incremental computation of the marginal gain of any node.

Fact 2. *The marginal gain of adding a node u to S, with u having Hamming ball B_u^{ξ}, is equal to $| B_u^{\xi} \setminus B_S^{\xi} |$, where $B_S^{\xi} = \cup_{v \in S} B_v^{\xi}$.*

Proposition 9. *The Hamming-based diversity function defined in Equation 6.8 is monotone and submodular.*

*Proof (sketch).*Monotonicity of Equation 6.8 is trivial. In fact, since the equation takes into account the union of the Hamming balls associated with any node in the set, greater sets can only lead to greater Hamming balls, thus Equation 6.8 is only allowed to increase.

As concerns the submodularity, it should be noted that for any $S \subseteq T \subseteq V$, it holds that $B_S^{\xi} \subseteq B_T^{\xi}$. In light of Fact 2, we can write the inequality between the

[2]For any two nodes u and v, we assume that if either u's or v's profile does not contain a value in the domain of A_j (i.e., missing value for A_j), with $j = 1..m$, then the indicator function will be evaluated as 1.

marginal gain of any node v with respect to S and T as:

$$\cancel{div(S)} + \mid B_v^\xi \setminus B_S^\xi \mid - \cancel{div(S)} \quad \geq \quad \cancel{div(T)} + \mid B_v^\xi \setminus B_T^\xi \mid - \cancel{div(T)}$$

In order to prove the submodularity, we can proceed by contradiction. Suppose there exists a node v such that the following inequality is strictly satisfied:

$$\begin{array}{lcl} \cancel{\mid B_v^\xi \mid} - \mid B_v^\xi \cap B_S^\xi \mid & < & \cancel{\mid B_v^\xi \mid} - \mid B_v^\xi \cap B_T^\xi \mid \\ \mid B_v^\xi \cap B_S^\xi \mid & > & \mid B_v^\xi \cap B_T^\xi \mid \end{array}$$

It is easy to verify that the above inequality is a contradiction, in fact since $B_S^\xi \subseteq B_T^\xi$, there cannot exist any node u belonging to the intersection in the leftmost side of the equation that does not belong to the intersection in the rightmost side. □

6.4.4 Entropy-based diversity

Diversity for categorical data can naturally be associated with notions of heterogeneity, or variability, for discrete random variables, such as entropy and Gini-index. Unfortunately, it is easy to note that such measures cannot be used to define a monotone submodular function of diversity as long as they are evaluated on any discrete random variable whose sample space (i.e., set of admissible values) corresponds to the categorical content of $\mathcal{D}_S$, for any $S \subseteq V$. For instance, if we describe each node-profile, resp. each attribute-value, in $\mathcal{D}_S$ by means of a vector whose generic entry represents the frequency of that profile, resp. attribute-value, then the entropy for the correponding probability mass function does not even preserve monotonicity for any $T \supseteq S$.

Nonetheless, it is known that entropy is monotone and submodular if defined for a *set of discrete random variables* [61]. Given a collection $\mathcal{X} = \{X_i\}_{i=1}^n$ of discrete random variables, with images (countable sets) here denoted as F_{X_i} $(i = 1, \ldots, n)$, the entropy function $H : 2^{\mathcal{X}} \mapsto [0, +\infty)$ is defined as:

$$\begin{aligned} H(X_1, \ldots, X_n) &= \mathbb{E}[-\log P(X_1, \ldots, X_n)] \\ &= -\sum_{x_1 \in F_{X_1}} \cdots \sum_{x_n \in F_{X_n}} P(X_1 = x_1, \ldots, X_n = x_n) \log P(X_1 = x_1, \ldots, X_n = x_n). \end{aligned}$$

where $x_1, \ldots, x_n$ are particular values of $X_1, \ldots, X_n$, respectively, and $P(X_1 = x_1, \ldots, X_n = x_n)$ denotes the joint probability that the values of the variables X_i are, respectively, equal to x_i $(i = 1, \ldots, n)$.

As previously mentioned, the entropy function defined above satisfies the properties of monotonicity and submodularity [61]. In fact, it holds that $H(\mathcal{X}_S) \leq H(\mathcal{X}_T)$ and that $H(\mathcal{X}_S, X) - H(\mathcal{X}_S) \geq H(\mathcal{X}_T, X) - H(\mathcal{X}_T)$, with $\mathcal{X}_S \subseteq \mathcal{X}_T \subseteq \mathcal{X}$ and $X \in \mathcal{X}, X \notin \mathcal{X}_T$. Hence, one question here becomes how to suitably define the variables over $\mathcal{D}_S$, for any $S \subseteq V$. We next provide an intuitive definition that is valid in our context.

Definition 11. *Given any $S \subseteq V$, we define a set $\mathcal{X}_S = \{X_i\}_{i=1..|S|}$ of discrete random variables associated with the profiles of nodes in S, where for each $v_i \in S$, $X_i : dom \mapsto \{0, 1\}$, such that dom is equipped with a probability function that assigns each $a \in dom$ with its relative frequency in $\mathcal{D}$, and X_i takes the value 1 if a is contained in $\mathcal{A}[v_i]$, 0 otherwise.* □

It is known that the set entropy $H(X_1, \ldots, X_n)$ can be rewritten in terms of conditional entropy through a *chain rule* for discrete random variables [47]:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}, \ldots, X_1).$$

That is, the entropy of a collection of random variables is the sum of the conditional entropies. In particular, given three variables, it holds that:

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \\ &= H(X_1, X_2) + H(X_3|X_2, X_1). \end{aligned}$$

It should also be noted that a sequence of random variables can be considered as a single vector-valued random variable, therefore the joint probability distribution $p(\mathcal{X})$ can also be seen as the probability distribution $p(\mathbf{X})$ of the random vector $\mathbf{X} = [X_1, \ldots, X_n]$. This naturally reflects as well on the computation of the conditional entropy of a variable given a sequence of random variables.

Definition 12. *Given a set of categorical attributes $\mathcal{A} = \{A_1, \ldots, A_m\}$ and associated profile set $\mathcal{D}$ for the nodes in a graph $G_0 = \langle V, E \rangle$, we define the* entropy-based diversity *of any $S \subseteq V$ as:*

$$div(S) = H(X_1, \ldots, X_{|S|}) = \sum_{i=1}^{|S|} H(X_i|\mathbf{X}^{<i}), \tag{6.9}$$

where $\mathcal{X}_S = \{X_i\}_{i=1..|S|}$ is the set of discrete random variables corresponding to nodes in S, $\mathbf{X}^{<i}$ denotes the vector of variables $X_1, \ldots, X_{i-1}$, and

$$\begin{aligned} H(X_i|\mathbf{X}^{<i}) = &- \sum_{x \in \{0,1\}^{i-1}} p(\mathbf{X}^{<\mathbf{i}} = x) \\ &\times \sum_{x_i \in \{0,1\}} p(x_i|\mathbf{X}^{<i} = x) \log p(x_i|\mathbf{X}^{<i} = x) \\ = &- \sum_{x \in \{0,1\}^{i-1}} p(\mathbf{X}^{<i} = x) \times H(X_i|\mathbf{X}^{<i} = x). \end{aligned}$$

□

In the above equation, note that the enumeration of 0-1 tuples of length i is only limited to the joint variable combinations corresponding to the attribute-values occurring in $\mathcal{D}$, whereas for all other attribute-values a' not in $\mathcal{D}$, the same tuple of all zeros is associated with the sum of probabilities of a' in $\mathcal{D}$.

The following fact states that the entropy-based diversity function allows for an incremental computation of a node's marginal gain.

Fact 3. *The marginal gain of adding a node v to S is equal to the conditional entropy $H(X_{|S|+1} \mid \mathbf{X}^{<|S|+1})$.*

Proposition 10. *The entropy-based diversity function defined in Equation 6.9 is monotone and submodular.*

*Proof (sketch).*Monotonicity and submodularity are ensured given the strict relation between the joint entropy function and a polymatroid [61]. Moreover, as concerns submodularity in particular, note that in the inequality $H(\mathcal{X}_S, X) - H(\mathcal{X}_S) \geq$

$H(\mathcal{X}_T, X) - H(\mathcal{X}_T)$ (with $\mathcal{X}_S \subset \mathcal{X}_T \subseteq \mathcal{X}$ and $X \in \mathcal{X}, X \notin \mathcal{X}_T$), each of the two terms is just the conditional entropy of variable X given $\mathcal{X}_S$ and $\mathcal{X}_T$, respectively. Therefore, $H(X|\mathcal{X}_S) \geq H(X|\mathcal{X}_T)$ holds since conditioning cannot increase entropy. □

6.4.5 Class-based diversity

We now introduce a subclass of diversity functions which differs from the ones previously described in that it exploits a-priori knowledge on a grouping of the node profiles. This might be particularly relevant in scenarios where we are interested in distinguishing the nodes based on a coarser grain than their individual profiles. An available organization of the profiles into categorically-cohesive groups could reflect some predetermined equivalence classes of the profiles w.r.t. a given schema of attributes $\mathcal{A}$. (This in principle also includes the opportunity of defining profile groups based on the availability of a *community structure* over the set of nodes in the network.)

A simple yet efficient approach to measure diversity based on the exploitation of profile groups is to cumulate the *selection rewards* for choosing nodes with a profile that belongs to any given class.

Definition 13. *Given a set of categorical attributes $\mathcal{A} = \{A_1, \ldots, A_m\}$ and associated profile set $\mathcal{D}$ for the nodes in a graph $G_0 = \langle V, E \rangle$, we define the* class-based diversity *of any $S \subseteq V$ as:*

$$div(S) = \sum_{l=1..h} \mathrm{f}\Big(\sum_{v_j \in C_l \cap S} r_j \Big) \tag{6.10}$$

where $\mathcal{C} = \{C_1, \ldots, C_h\}$ is a partition of $\mathcal{D}$ (i.e., $\bigcup_{l=1}^{h} C_l = \mathcal{D}$, and $C \cap C' = \emptyset$, for each $C, C' \in \mathcal{C}$, with $C \neq C'$), $\mathrm{f} : \mathbb{R} \mapsto \mathbb{R}$ is any non-decreasing concave function, and $r_j > 0$ is the selection reward for $v_j \in V$. □

The effect of f is that repeatedly selecting nodes of the same class yields increased diminishing gains for the previously selected nodes. In fact, since f is nonnegative concave and $\mathrm{f}(0) \geq 0$, f is also *subadditive* on $\mathbb{R}^+$, i.e., $\sum_{x_i=0}^{+\infty} \mathrm{f}(x_i) \geq \mathrm{f}(\sum_{x_i=0}^{+\infty} x_i)$. Therefore, adding (to the set S being discovered) a node from a different class is preferable in terms of marginal gain than adding a node from an already covered class. Example instances of $\mathrm{f}(x)$ are $\sqrt{x}$ and $\log(1+x)$, but any other non-decreasing concave function can in principle be adopted. We now provide the lower bound and upper bound of Equation 6.10 when the logarithmic function is adopted.

Proposition 11. *Given a budget k and h classes, the function in Equation 6.10, equipped with $\mathrm{f}(x) = \log(1+x)$, with $r_j = 1, \forall v_j \in V$, achieves the minimum value of $\log(1+k)$ when all k nodes belong to the same class (i.e., 1 class covered), and the maximum value of k when all k nodes belong to different classes (i.e., k classes covered).*

*Proof (sketch).*The values of $\log(1+k)$ and k are immediately derived by evaluating Equation 6.10 for the cases $h = 1$ and $h = k$, respectively. The proof of k as upper bound is immediate. To prove that $\log(1+k)$ is the lower bound of Equation 6.10, consider w.l.o.g. a uniform class distribution, i.e., there are k/h (with $h < k$) nodes that belong to each class. In this case, it holds that $div(S) = h \log(1 + k/h)$, for any size-k S. It follows that the inequality $\log(1+k) \leq h \log(1 + k/h)$ must be verified (with equality iff $h = 1$). This is immediately derived by observing that, after algebraic manipulation, the above inequality holds iff $(1+k)h^h \leq (h+k)^h$, which is true since the terms on the left side are contained in the polynomial $(h+k)^h$. □

Again, the above defined function enables an incremental computation of the marginal gain of any node.

Fact 4. *The marginal gain of adding a node v to S, with v belonging to class C_l, is equal to $\log(1 + r/R_l)$, where r is the reward of adding v and R_l is one plus the sum of rewards of nodes in S that belong to class C_l.*

Proposition 12. *The partition-based diversity function defined in Equation 6.10 is monotone and submodular.*

*Proof (sketch).*Monotonicity and submodularity of the function in Equation 6.10 can directly be derived from the mixture property of submodular functions and the composition property of submodular with nondecreasing concave functions [136], respectively. In fact, the summation argument of f is a collection of modular functions with nonnegative weights (and hence is monotone), the application of f yields a submodular function, and finally summing up over the groups retains monotonicity and submodularity. □

6.5 A RIS-based framework for the ADITUM problem

We develop our framework for the ADITUM problem based on the **R**everse **I**nfluence **S**ampling (RIS) paradigm first introduced in [20] and recognized as the state-of-the-art approach for IM problems.

As discussed in Section 6.2, the RIS based approach overcomes the limitations of the Monte Carlo based greedy approach to IM. The RIS paradigm relies on the following two concepts. Given the diffusion graph G with node set V and edge set E, let G be an instance of G obtained by removing each edge $e \in E$ with probability $1 - p(e)$, where $p(e)$ denotes the probability on edge e in G. The *reverse reachable set* (RR-Set) rooted in v w.r.t. G contains all the nodes reachable from v in a backward fashion. A *random RR-Set* is any RR-Set generated on an instance G, for a node selected uniformly at random from G.

The key idea of the RIS framework is that the more a node u appears in a random RR-Set rooted in v, the higher the probability that u, if selected as seed node, will activate v. The design of the RIS framework follows a *two-phase schema* [20]: (1) Generate a certain number of random RR-Sets, and (2) Select as seeds the k nodes that cover the most RR-Sets. (The latter step can be solved by using any greedy algorithm for the Maximum Coverage problem.)

Based on RIS, Tang et al. [189] developed the TIM and TIM+ algorithms that achieve $(1 - 1/e - \epsilon)$-approximate solutions for the IM problem, with at least $1 - |V|^{-l}$ probability in time $O((k+l)(|E|+|V|)\log|V|/\epsilon^2)$, where $l = 1$ by default. TIM/TIM+ works in two major stages: *parameter estimation* and *seed selection*. The first stage aims at deriving a lower-bound for the maximum expected spread that can be achieved by a size-k seed set, from which depends the number θ of random RR-Sets that must be generated in the second stage; the latter essentially coincides with the second phase of the RIS method. Note that TIM+ is designed to improve upon TIM by adding an intermediate step between parameter estimation and node selection, which heuristically refines θ into a tighter lower bound of the maximum expected influence of any size-k node set. Also, it should be noted that further developments introduced to speed up TIM+, like IMM [188], still maintain the same computational complexity as TIM+.

The effectiveness of TIM+ is explained by Lemma 2 provided in [189], which states that, if θ is sufficiently large, the fraction of random RR-Sets covered by any seed set S is a good and unbiased estimator of the average node-activation probability.

6.5.1 Proposed approach

Our proposed RIS-based framework follows the above discussed two-phase schema, however it originally embeds both the targeted nature and the diversity-awareness in an influence maximization task. To accomplish this, we revise the two-phase schema as follows.

Parameter estimation We want to understand how much capital can be captured from a size-k seed set. Therefore, to compute the number θ of RR-Sets, we need to identify a lower-bound on the maximum capital score.

We select a node v as the root of an RR-Set with probability $p(v) \propto t(v)$. Since we are interested in the activation of the target nodes only, we set

$$p(v) = \frac{t'(v)}{\mathcal{T}_{TS}}, \quad \text{with} \quad t'(v) = \begin{cases} t(v), & \text{if } v \in TS \\ 0, & \text{otherwise} \end{cases}$$

and $\mathcal{T}_{TS} = \sum_{u \in TS} t(u)$. We leverage on the TIM+ procedures *KPTEstimation* and *RefineKPT*, in order to estimate a lower-bound for the average activation probability achieved by any optimal seed set of size k. At a high level, the *KPTEstimation* procedure starts by generating a relatively small number of RR sets upon which an initial approximation of the expected spread is computed. The number of RR sets is iteratively increased until the estimate matches a certain error bound. The *RefineKPT* procedure improves over the first lower-bound estimation. More specifically, it computes an initial seed set upon the random RR sets generated in the last iteration of *KPTEstimation*, and estimates the spread of this initial seed set w.r.t. a new selection of RR sets; the number of these new RR sets is kept reasonably high in order to ensure the accuracy of the last estimate. Finally, *RefineKPT* returns the maximum value between the first and the last approximation.

Note that we can rely upon the two TIM+ procedures since the capital achieved is contingent on the activation process, thus we still need to have an unbiased estimator for the spread function. In fact, any target node will contribute in terms of capital as long as it has been activated starting from the seed set. The lower-bound on the expected spread allows us to derive a lower-bound on the average activation probability, from which we compute the *expected capital* score of a seed set as

$$\mathbb{E}[C(S)] = \mathcal{T}_{TS}(\mathbb{E}[\mu(S)])/|V|. \tag{6.11}$$

Above, the rightmost term is the average fraction of total capital score $\mathcal{T}_{TS}$, the seed set S is able to capture. Moreover, since every random RR-Set is rooted in a target node, the aforementioned Lemma 2 [189] ensures that $\mathbb{E}[\mu(S)]/|V|$ is very close to the average activation probability of the target nodes.

Seed selection Once all θ RR-Sets are computed, the second phase is in charge of detecting the k seeds. To this end, we take into account a notion of set-diversity to choose the candidate seeds. The selection of best seeds is accomplished in a greedy fashion, one seed at a time. A node v is associated with a linear combination of (i) the *node's capital score*, obtained by summing the target scores of the roots of the

RR-Sets to which v belongs and that are not already covered by seeds, and (ii) the *node's diversity score*, which corresponds to the node's marginal gain for the diversity function w.r.t. the current seed set.

Remarks The objective function we seek to maximize is a linear combination of two main quantities: the expected capital and the diversity of the seeds. Note that there is a key difference between these two measures: the former is defined globally over the whole network, while the latter is limited to the seed nodes, namely the solution itself.

Our approach hence reflects this inherent interplay between capital and diversity. In fact, the sampling procedure in the first stage corresponding to the parameter estimation, is driven by only the capital score — there are no seeds upon which the diversity must be assessed — whereas the diversity aspect comes into play only during the process of seed set formation, thus it drives the discovery of the seeds.

6.5.2 The ADITUM algorithm

We describe here algorithmic details of our proposed approach to solve the Attribute-based Diversity-sensitive Targeted Influence Maximization problem we provided in Definition 7, which originally embeds a categorical-set diversity function into the IM optimization criterion.

Algorithm 5 shows the pseudocode of our implementation of ADITUM. The algorithm starts by identifying the target nodes (line 1), then it infers the number θ of RR-Sets to be computed, according to TIM+ subroutines of estimation and refinement of KPT, i.e., the mean of the expected spread of possible size-k seed sets (line 2). In lines 4-6, the θ RR-Sets are generated by invoking the *computeRandomRRSet* procedure (lines 4-6). The procedure *buildSeedSet* eventually returns the size-k seed set (lines 7-8). In the following, we provide details about the two procedures.

Procedure *computeRandomRRSet* starts by sampling node r as the root of R from a distribution of probability proportional to the target-node scores (line 11). Each RR-Set is associated with an integer identifier and the root node (line 12) — this information is needed since the capital associated with a set is given by the target score of its root. Finally, an instance of the influence graph $G \sim \mathcal{G}$ is computed according to the live-edge model related to $\mathcal{M}$, then all the nodes that can reach r in G are inserted in the RR-Set to be returned.

Procedure *buildSeedSet* exploits a priority queue q, which is initialized (line 16) to store triplets comprised of: value of the linear combination of capital and diversity, node and iteration which the value refers to. The triplets are ordered by decreasing values of capital-diversity combination. For each node v, its capital score is computed by summing the target score of all nodes that are roots of an RR-Set v belongs to (line 18). Moreover, the v's diversity score is computed as its marginal gain for the div function w.r.t. the current seed set (line 19), according to the particular diversity notion involved (cf. Facts 1–4, Section 6.4). Once all the scores are computed, the procedure starts to select the seeds, by getting at each iteration the best triplet from the queue (line 23): if the choice is done at iteration it equal to the number of nodes currently in the seed set (line 24), then v is inserted in S, and all sets covered by v are stored in CS; otherwise, all the scores are to be recomputed. By denoting with $\mathcal{R}(v)$ the set of random RR-Set containing v, the v's capital score is decreased by the target score of each node r that is root of an already covered RR-Set (i.e., a set in $\mathcal{R}(v) \cap CS$) (line 28), and this set is also removed from $\mathcal{R}(v)$ (line 29). The diversity score needs

Algorithm 5 **A**ttribute-based **DI**versity-sensitive **T**argeted Infl**U**ence **M**aximization (ADITUM)

Input: A diffusion graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, b, t \rangle$ based on triggering model $\mathcal{M}$, a budget k, a target selection threshold $\tau_{TS} \in [0, 1]$, a smoothing parameter $\alpha \in [0, 1]$.
Output: Seed set S of size k.

1: $TS \leftarrow \{v \mid t(v) \geq \tau_{TS}\}$ {Select the target nodes}
2: Compute θ by using TIM+ procedures *KPTEstimation* and *RefineKPT*
3: $\mathcal{R} \leftarrow \emptyset$
4: **for** $i \leftarrow 1$ **to** θ **do**
5: $R \leftarrow$ computeRandomRRSet$(TS, \mathcal{M}, i)$
6: $\mathcal{R} \leftarrow \mathcal{R} \cup \{R\}$
7: **end for**
8: $S \leftarrow$ buildSeedSet$(\mathcal{R}, k, \alpha)$ {Seed Selection stage}
9: **return** S

10: **procedure** computeRandomRRSet$(TS, \mathcal{M}, id)$
11: $R \leftarrow \emptyset$ Initialize the RR-Set
12: Select node $r \in TS$ as root, with probability $p(r) \propto t(r)$
13: $R.id \leftarrow id, R.root \leftarrow r$ Associate id and root to the RR-Set
14: Add to R the nodes that can reach r according to live-edge model of $\mathcal{M}$
15: **return** R

16: **procedure** buildSeedSet$(\mathcal{R}, k, \alpha)$
17: $q \leftarrow \emptyset$ {Priority queue for lazy-greedy optimization}
18: **for** $v \in V$ **do**
19: $v.pushedC \leftarrow \sum_{R \in \mathcal{R}(v)} c(root(R))$
20: $v.pushedD \leftarrow marginalGainInDiversity(v, \emptyset)$
21: $q.add(\langle(\alpha \times v.pushedC + (1 - \alpha) \times v.pushedD), v, 0\rangle)$
22: **end for**
23: $S \leftarrow \emptyset, CS \leftarrow \emptyset$
24: **repeat**
25: $\langle aDC\_val, v, it \rangle \leftarrow q.removeFirst()$
26: **if** $it = |S|$ **then**
27: $S \leftarrow S \cup \{v\}, CS \leftarrow CS \cup \mathcal{R}(v)$
28: **else**
29: **for** $R \in \mathcal{R}(v) \cap CS$ **do**
30: $v.pushedC \leftarrow v.pushedC - t(root(R))$
31: Remove R from $\mathcal{R}(v)$
32: **end for**
33: $v.pushedD \leftarrow marginalGainInDiversity(v, S)$
34: $q.add(\langle(\alpha \times v.pushedC + (1 - \alpha) \times v.pushedD), v, |S|\rangle)$
35: **end if**
36: **until** $|S| = k$
37: **return** S

also to be recomputed, finally the updated triplet is inserted into the priority queue (lines 30-31). The procedure loop ends when the desired size k is reached for the seed set (line 32).

Proposition 13. *ADITUM runs in* $O((k+l)(|E|+|V|) \log|V|/\epsilon^2)$ *time and returns a* $(1 - 1/e - \epsilon)$*-approximate solution with at least* $1 - |V|^{-l}$ *probability.*

Proof (sketch). ADITUM is developed under the RIS framework and follows the typical two-phase schema of TIM/ TIM+ methods, i.e., parameter estimation and (seed) node selection, for which the theoretical results in the Proposition hold. Due to the targeted nature of the problem under consideration, the expected capital must be computed in place of the expected spread; however, this only implies to choose

TABLE 6.2: Summary of real-world networks used in our experimental evaluation. Assortativity corresponds to the directed version of degree-based assortativity. Sinks and sources are nodes having zero out-degree and zero in-degree, respectively.

network	#nodes	#edges	avg. in-degree	avg. path length	clust. coeff.	assorta-tivity	#sources	#sinks
FriendFeed	493 019	19 153 367	38.85	3.82	0.029	-0.128	41 953	292 003
GooglePlus	107 612	13 673 251	127.06	3.32	0.154	-0.074	35 341	22
Instagram	17 521	617 560	35.25	4.24	0.089	-0.012	0	0
MovieLens	943	229 677	243.5	1.87	0.752	-0.323	1	1
Reddit	11 224	91 924	8.18	4.11	0.083	-0.072	0	0

a distribution over the roots of the RR-Sets, which depends on the target scores of the nodes in the network. Thus, the asymptotic complexity of TIM/TIM+ is not increased. Moreover, two major differences occur in the seeds selection phase of ADITUM w.r.t. TIM/TIM+, i.e., the lazy forward approach and the computation of the marginal gain w.r.t. the diversity function. However, both aspects do not affect the asymptotic complexity, since the former allows saving runtime only and the latter does not represent any overhead (computing a node's marginal gain is made in nearly constant time, for each of the diversity functions). Therefore, we can conclude that ADITUM has the same asymptotic complexity of TIM/TIM+. □

6.6 Evaluation methodology

6.6.1 Data

We used both synthetic and real-world data for our experimental evaluation. We selected real-world online social networks (OSNs) as input graphs for the influence maximization task, while for the specification of the categorical data, we adopted a twofold methodology: firstly, we developed a generator of synthetic categorical datasets as benchmark for an in-depth analysis of the different diversity functions; secondly, we exploited user profile data, when available, associated to the users in the evaluation OSNs.

Synthetic data. Our developed generator of synthetic categorical datasets requires the following parameters as inpuattribute:t: the number m of attributes of the schema $\mathcal{A}$, the number n of categorical tuples (i.e., vectors of categorical attribute values) to be generated according to $\mathcal{A}$, the number of symbols (admissible values) for each attribute (i.e., $|dom_A|$, for every $A \in \mathcal{A}$), and the type of distribution of the values of each attribute. All generated data tuples will have no missing attribute values. We set $n = 1000$, $m = |\mathcal{A}|$ from 5 to 50, $|dom_A| = \{5, 10, 15\}$ ($\forall A \in \mathcal{A}$), while the selected types of per-attribute distributions are *uniform* (with parameters 0 and 1), *standard normal*, and *exponential* (with rate $\lambda = 1$).

Real-world networks. We chose five real-world OSNs, namely:
FriendFeed [26], *GooglePlus* [26], *Instagram* [26], *MovieLens* [185], and *Reddit* [106]. Table 6.2 shows main statistics about the evaluation networks. Our choice of the datasets is motivated by the following reasonattribute:s:

- *reproducibility*, i.e., all of the networks are publicly available;
- *diversification* of the evaluation scenarios, which include user engagement and item recommendation;

- *continuity* w.r.t. previous studies;
- *fair comparative evaluation*, i.e., we based our choice also in relation of the competing methods included in our evaluation, so to enable a fair comparison between them and our ADITUM.

FriendFeed, GooglePlus, and Instagram network datasets refer to OSNs previously studied in a *user engagement* scenario, which has been recognized as an important case in point for demonstrating targeted IM tasks [26]. For each of these networks, the meaning of any directed edge (u, v) is that user v is "consuming" information received from u (e.g., v likes/comments/rates a u's media post). No side information is originally provided with such datasets, therefore we synthetically generated the user profiles as followattribute:s: Given m categorical attributes, each with n_i admissible values ($i = 1..m$), we associated each user with a set of values sampled from either uniform or exponential (with $\lambda = 1$) distribution. We set $m = n_i = 10$. We used these datasets also for comparison with DTIM.

Originally used for *movie recommendation*, MovieLens is associated with a (user, movie-genre) rating matrix storing the number of movies each user rated for each genre, at any given time over a predefined observation period. This dataset was previously included in the evaluation of our competitor Deg-D. To enable ADITUM to work on MovieLens, we mapped each genre to an attribute, with unique rating-values as corresponding attribute-values. The MovieLens network was built so to have users as nodes and any directed edge (u, v) is drawn if user u rated *first* at least 10 movies in common with v (timestamps are available in the original data).

Reddit network represents the directed connections between two subreddits, i.e., communities on the Reddit platform. Each connection refers to a post in the source community that links to a post in the target community. From the original network, we kept only the connections for which the source post is explicitly positive towards the target post, and finally extracted the largest strongly connected component to overcome sparsity issues. Reddit connections are also rich in terms of numerical attributes associated with each source post, which include both lexical and sentiment information. We selected 11 attributes which appeared to be the most informative for influence propagation reasons.[3] To generate the profile of each node (community), we grouped the posts by community and summed up the scores for each attribute; finally, the values of each attribute were discretized through a 10-quantile binning scheme.

6.6.2 Evaluation goals and settings

We devised different settings according to our evaluation goals, which are organized into three main stages of analysis. Hereinafter, we will use notations $div^{(AW)}$, $div^{(HB)}$, $div^{(E)}$, $div^{(C)}$ to refer to the attribute-wise, Hamming-, entropy-, and class-based definitions of diversity functions, respectively.

Stage 1 – Sensitivity of diversity functions The first stage of evaluation is focused on an analysis of the proposed diversity functions regardless of the context of the ADITUM problem. More specifically, we want to understand the relations between the size of a set of categorical tuples, the number of attributes (size of the schema), the number of attribute symbols and their distribution, and how these affect the behavior of each diversity function.

[3]We selected the POST_PROPERTIES attributes corresponding to the following identifierattribute:s: 19, 20, 21, 43, 44, 45, 46, 51, 52, 53, 66.

Using our synthetically generated datasets, and for every combination of seed-set size k and number $i \in [1..m]$ of attributes to select from the data schema, we considered the following experimental settingattribute:s:

(S1.1) We evaluated the expected diversity via Monte Carlo simulation with 10 000 runs. More specifically, in each run of the simulation, we first randomly selected a size-k set of categorical tuples projected over the first i attributes in $\mathcal{A}$, then we measured its diversity, according to each diversity function.

(S1.2) For each diversity function, we computed the size-k set of categorical tuples that maximizes its value.

Intuitively, both settings enable us to characterize and comparatively evaluate the diversity functions when varying properties of the input data. Moreover, the second setting is also concerned with finding the input conditions that allow for maximizing each diversity function.

Stage 2 – Evaluation of ADITUM For the second stage of evaluation, we considered ADITUM instantiations with each of the definitions of diversity proposed in Section 6.4. We experimentally varied the setting of ADITUM parameterattribute:s: the seed-set size k, within $[5..50]$, the smoothing parameter α, from 0 to 1 with step 0.1, and the target selection threshold τ_{TS}; the latter was controlled in terms of percentage of top-values from the target score distribution, thus we selected target sets corresponding to the top-$\{5\%, 10\%, 25\%\}$. We used the default $\epsilon = 0.1$ for the approximation-guarantee in the parameter estimation phase. Concerning the edge weighing function (b) and the node weighing function (t), we devised the following settingattribute:s:

(S2.1) The first setting refers to the basic, non-targeted setting adopted in [185], i.e., $b(u, v) = 1/n_v$, with n_v number of v's in-neighbors, and $t(u) = 1$, for all u, v in V. We used this setting for MovieLens evaluation.

(S2.2) The second setting refers to Reddit, for which the influence weights are set to be proportional to the amount of interactions between communitiattribute:es: for any two nodes u and v, $b(u, v) = P_{uv}/P_v$, where P_{uv} is the number of posts of u directed to v, and P_v is the total number of posts having v as target. The node weighing function is here simply defined as the in-degree function, in order to mimic a scenario of influence targeting as corresponding to communities that are highly popular in terms of post recipients.

(S2.3) The third setting refers to a user engagement scenario and applies to FriendFeed, GooglePlus and Instagram, which were previously used in that context [26]. User engagement is addressed as a topology-driven task for encouraging silent users, a.k.a. "lurkers", to return their acquired social capital, through a more active participation to the community life. Note that such users are effective members of an OSN, who are not actively involved in tangible content production and sharing with other users in the network, but rather they are information consumers. Given this premise, in [26] a specific instance of targeted IM is developed such that lurkers are regarded as the target of the diffusion process. Therefore, the user engagement task becomattribute:es: Given a budget k, to find a set of k nodes that are capable of maximizing the likelihood of "activating" (i.e., engaging) the target lurkers. In this context, the two weighing functions

rely on a pre-existing solution of a *lurker ranking* algorithm applied to the social graph. The intuition is as follows (the interested reader is referred to [26] for analytical details about the weighing functions): For any node v, the node weight $t(v)$ indicates the status of v as lurker, such as the higher the lurker ranking score of v the higher should be $t(v)$; for any edge (u, v), the weight $b(u, v)$ is computed to measure how much node u has contributed to the v's lurking score calculated by the lurker ranking algorithm, which resembles a measure of "influence" produced by u to v.

Stage 3 – Comparative evaluation with competing methods The closest methods to ADITUM are DTIM [26] and Deg-D [185]. As previously mentioned, DTIM addresses targeted IM, but it considers topology-driven notions of diversity only; conversely, Deg-D assumes a numerical representation of node attributes and the addressed problem is not targeted.

The objective function in DTIM [26] shares the capital term with ADITUM, which is however combined with a diversity term defined as $\sum_{s\in S}\sum_{v\in TS} div_v(s)$, i.e., the sum of diversity scores that each seed has in relation with each of the target nodes, where $div_v(\cdot)$ is either the global topology-driven or the local topology-driven diversity function [26]. To enable a comparison with DTIM, we integrated its global topology-driven diversity function into our RIS-based framework, following the guidelines provided in [26].

Deg-D [185] follows a simple greedy scheme to maximize the objective function $(1-\gamma)\sum_{u\in S} deg(u) + \gamma D(S)$, where $deg(u)$ denotes the out-degree of node u, while $D(S)$ represents the diversity of the set S, whose value is given by: $D(S) = \sum_{m=1}^{M} f(\sum_{u\in S} \omega_{um} \times g(u))$, where M denotes a given number of types of external information, γ is a smoothing parameter, $\omega_{um} \in [0, 1]$ is a real-valued coefficient expressing the preference of node u toward type m, f denotes any nondecreasing concave function (with default form set to $f(x) = \log(1 + x)$), whereas g is a function defined for each node u, either as $g(u) = 1$ or $g(u) = deg(u)$; the two different definitions of g lead to the variants named Deg-DU and Deg-DW, respectively. Note that, compared to α in ADITUM, γ in Deg-D has an opposite role, therefore we set $\gamma = 1 - \alpha$ in all the experiments. Moreover, Deg-D requires a numeric vector of size M to be associated with each node. Therefore, to account for the numerical representation of node attributes handled by this method, we devised two settingattribute:s:

- Comparison of the two methodattribute:s: ADITUM upon categorical representation derived from a numerical representation of nodes' attributes vs. Deg-DU and Deg-DW upon normalized numerical representation;
- Integration of the *uniform* and *weighted* functions, i.e., Deg-DU and Deg-DW, resp., into our RIS-based framework, upon numerical representation of nodes' attributes.

Remarks on the scope of evaluation It should be noted that, *although many methods exist for targeted IM, no one has the same diversity-aware goal or data features as ours*; one exception is represented by the two methods discussed above, for which nonetheless we had to devise suitable settings to allow for a fair comparison. As a consequence, it would be useless to compare the spread produced by our method to the spread achieved by any other method that does not incorporate diversity into the activation function; also, no comparison would make sense in terms of the seeds to be

identified, since any variability in the categorical attributes of the seed set discovered by a method that is not diversity-sensitive would be due to random chance.

Another notable point is that, for the sake of significance of experimental results, our evaluation framework requires that an IM method must be equipped with an activation function that is monotone submodular, in order to support the development of an approximate solution with theoretical guarantee. This advocates our choice of leaving *out of consideration any non-submodular diversity measure* adopted in contexts of search-result diversification (cf. Section 6.2).

6.7 Experimental results

6.7.1 Stage 1 - Sensitivity of diversity functions

To characterize the behavior of our diversity functions, we analyzed their sensitivity to the input categorical data, by varying the number of attributes, the number of attribute symbols, and their distribution. Our assessment is focused around the following two statistattribute:ics: (i) the relative change rate and (ii) the average Jensen-Shannon divergence.

The relative change rate of a diversity function is computed w.r.t. the change in the size of the set of categorical tuples upon which the diversity is evaluated. Formally, given any two set-size values, k_1 and k_2, with $k_2 > k_1$, we define the *relative change rate* of a diversity function f aattribute:s:

$$rcr(\mathrm{f}, k_1, k_2) = \frac{1}{k_2 - k_1} \frac{\mathrm{f}(k_2) - \mathrm{f}(k_1)}{\mathrm{f}(k_1)},$$

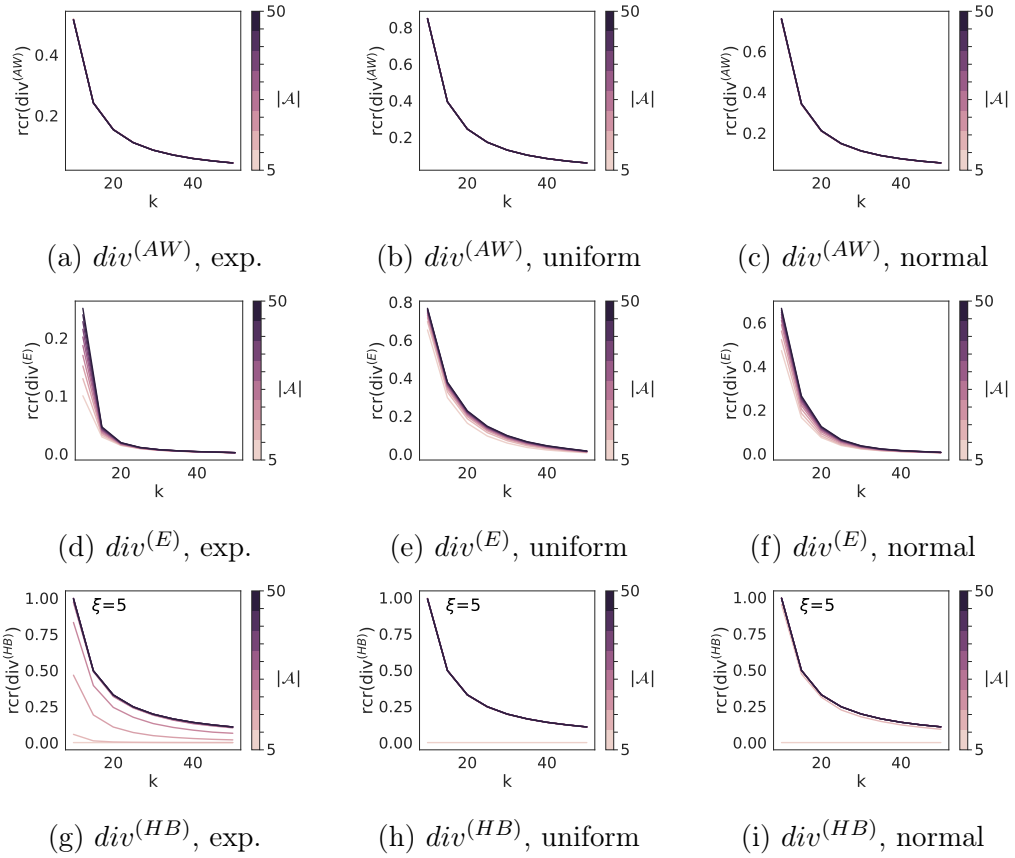
where $\mathrm{f}(k)$ denotes the evaluation of function f on sets of size k based on Monte Carlo simulations, as previously discussed in Section 6.6.2.

We also measured the Jensen-Shannon divergence to compare any two sets of categorical tuples according to the frequency of the values occurring in each particular attribute of their schema. More precisely, for any given attribute $A \in \mathcal{A}$ and set of categorical tuples S, we first computed a distribution of probabilities corresponding to the relative frequencies of the symbols in A over the tuples of set S. If we denote this distribution with $dist_{S,A}$, the *Jensen-Shannon divergence* between any two sets S_1 and S_2 w.r.t. A is defined aattribute:s:

$$JSDiv(dist_{S_1,A} \parallel dist_{S_2,A}) = \frac{1}{2} KL(dist_{S_1,A} \parallel mean) + \frac{1}{2} KL(dist_{S_2,A} \parallel mean)$$

where $mean = \frac{1}{2}(dist_{S_1,A} + dist_{S_2,A})$ and $KL(X \parallel Y)$ denotes the Kullback-Leibler divergence between any two probability distributions X, Y; we used base-2 logarithm so to bound the Jensen-Shannon divergence by 1, for any two probability distributions. Finally, for any two sets S_1 and S_2 sharing the same schema $\mathcal{A}$, we computed the overall Jensen-Shannon divergence between S_1 and S_2 as the average of their divergence values over all attributes in $\mathcal{A}$.

We organize the presentation of our main results as follows. We first investigate the sensitivity of each diversity function, in terms of its relative change rate w.r.t. the number of attributes, in Section 6.7.1.1, and the number of attribute symbols, in Section 6.7.1.2. Next, in Section 6.7.1.3, for each pair of diversity functions, we analyze the Jensen-Shannon divergence of the probability distributions for the categorical



(a) $div^{(AW)}$, exp. (b) $div^{(AW)}$, uniform (c) $div^{(AW)}$, normal

(d) $div^{(E)}$, exp. (e) $div^{(E)}$, uniform (f) $div^{(E)}$, normal

(g) $div^{(HB)}$, exp. (h) $div^{(HB)}$, uniform (i) $div^{(HB)}$, normal

FIGURE 6.1: Relative change rate of diversity functions by varying the number of attributes ($|\mathcal{A}|$), on different categorical datasets. Different colors correspond to different projections of the dataset: the darker the color, the higher the number i of attributes selected from the schema, where $i \in [5..50]$ with increments of 5. The number of attribute symbols is set to 15.

tuple sets that respectively maximize the corresponding diversity function (cf. second setting of Stage 1, Section 6.6.2).

6.7.1.1 Effect of the number of attributes

We are interested in understanding how the size of the schema impacts on the change rate of each diversity function. In fact, since increasing the number of attributes enables accounting for more information, it might be desirable that the relative change rate of a diversity function will not vary greatly; otherwise, when contextualized to the ADITUM setting, this would introduce an issue of selecting a proper size of the categorical data schema upon which the function needs to be maximized in order to provide its best perfomance.

Figure 6.1 shows the relative change rate of diversity functions by varying the number of attributes. Please note that the results shown in the figure discard the class-based diversity function, since our goal here is to analyze the explicit effect of the number of attributes on the behavior of a diversity function. This effect might not fully be understood for the class-based diversity, because it is actually evaluated on a single attribute (external to the schema $\mathcal{A}$) whose values depend on how the categorical tuples were originally grouped (cf. Section 6.4.5).

As a general remark, it can be noted a non-increasing trend of the relative-change-rate curve of each function, regardless of the data characteristics. This is obviously expected, since all functions are submodular: as k increases, the relative change rate decreases, which is explained since this can intuitively be seen as related to the average marginal gain due to the inclusion of new elements to the input set, which becomes smaller as the set-size increases.

Considering the relative change rate of the attribute-wise diversity (Figs. 6.1(a-c)), we observe differences in the range of values (wider for the uniform distribution, narrower for the exponential distribution), but no effects due to the number of attributes.

Conversely, as shown in Figs. 6.1(d-f), the entropy-based diversity turns out to be slightly more sensitive to the number of attributes in the dataset schema, with the relative-change-rate curves that tend to increase in amplitude and become more similar by increasing the number of attributes (this is particularly evident with the exponential distribution, i.e., Figure 6.1(d)). As concerns the Hamming-based diversity function (Figs. 6.1(g-i)), we observe that, when the number of attributes is small, there is almost no variation of the relative change rate of the function: in fact, when $\xi \simeq |\mathcal{A}|$, it is highly likely that only few tuples would have a Hamming ball comprising most of the profiles in the categorical dataset, and consequently the advantage of adding new tuples to the current set becomes negligible. Moreover, the Hamming-based diversity function shows to be more sensitive than the entropy-based one, especially when using an exponential distribution of the attribute values. In general, the relative change rate of the Hamming-based diversity function tends to be higher than those of the other two functions. (Additional results corresponding to other settings of ξ are reported in the ***Appendix***, Figure D.1.)

Overall, our analysis has unveiled a relative robustness of the diversity functions w.r.t. the size of the schema, which supports the hypothesis stated at the beginning of this section. This holds strongly for the attribute-wise diversity and, to a less extent, for the entropy-based diversity, whereas for the Hamming-based diversity, it holds for uniform and normal attribute-value distributions.

6.7.1.2 Effect of the number of attribute values

Here, we analyze the relative change rate of each diversity function from the perspective of the number of admissible values for each attribute.

As it can be observed from the results shown in Figure 6.2, the higher the number of attribute symbols, the higher the relative change rate of every function. This is clearly explained since, by extending the domain of the attributes in the schema associated with a given set of categorical tuples, it is likely an increase in the growth rate of diversification within the set.

In particular, a larger domain of attribute values improves the chance of picking a new tuple having symbols not already present in the current set's domain. As a consequence, the value of k at which each relative-change-rate curve starts to become nearly constant needs to be higher as the number of attribute symbols increases. In other words, the average marginal gain has a more rapidly decreasing trend when the number of attribute symbols is lower.

Looking at the plots in Figs. 6.2(a)-(d), it can be noted that when using an exponential distribution for the attribute values, the relative change rates do not significantly vary with the number of attribute symbols. This should be ascribed to the skewness of the distribution, which dilutes the impact of diversification within the dataset due in principle to an increase in the number of attribute values, especially if

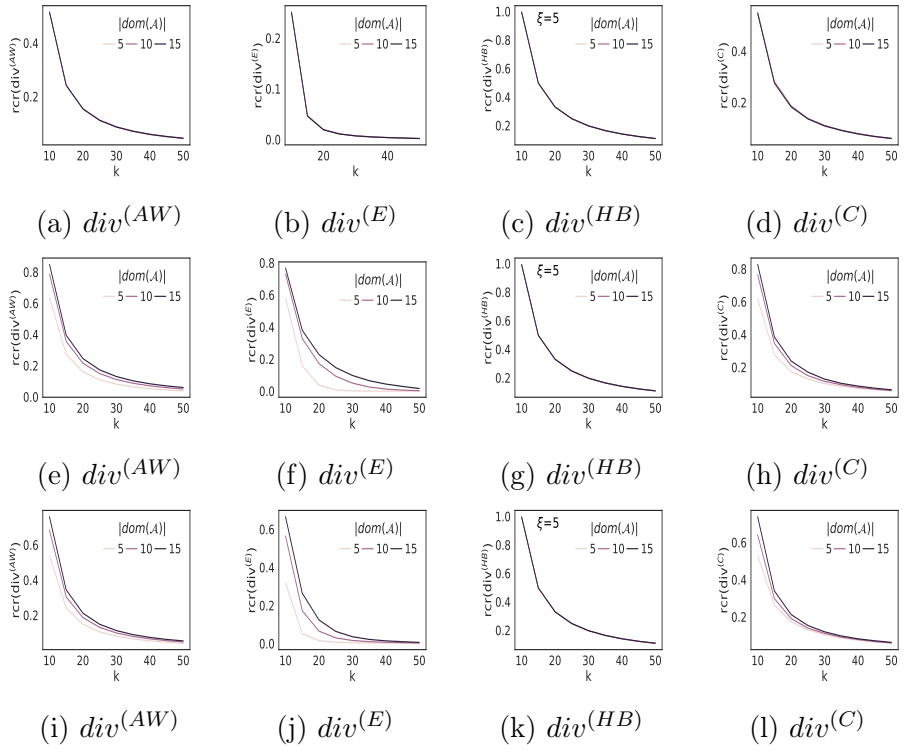


FIGURE 6.2: Relative change rate of diversity functions by varying the number of attribute symbols, on different categorical datasets (with $|\mathcal{A}|= 50$). Attribute values are distributed according to exponential (top row), uniform (mid row), and normal (bottom row) distributions. Different colors correspond to different number of values (darkest for 15).

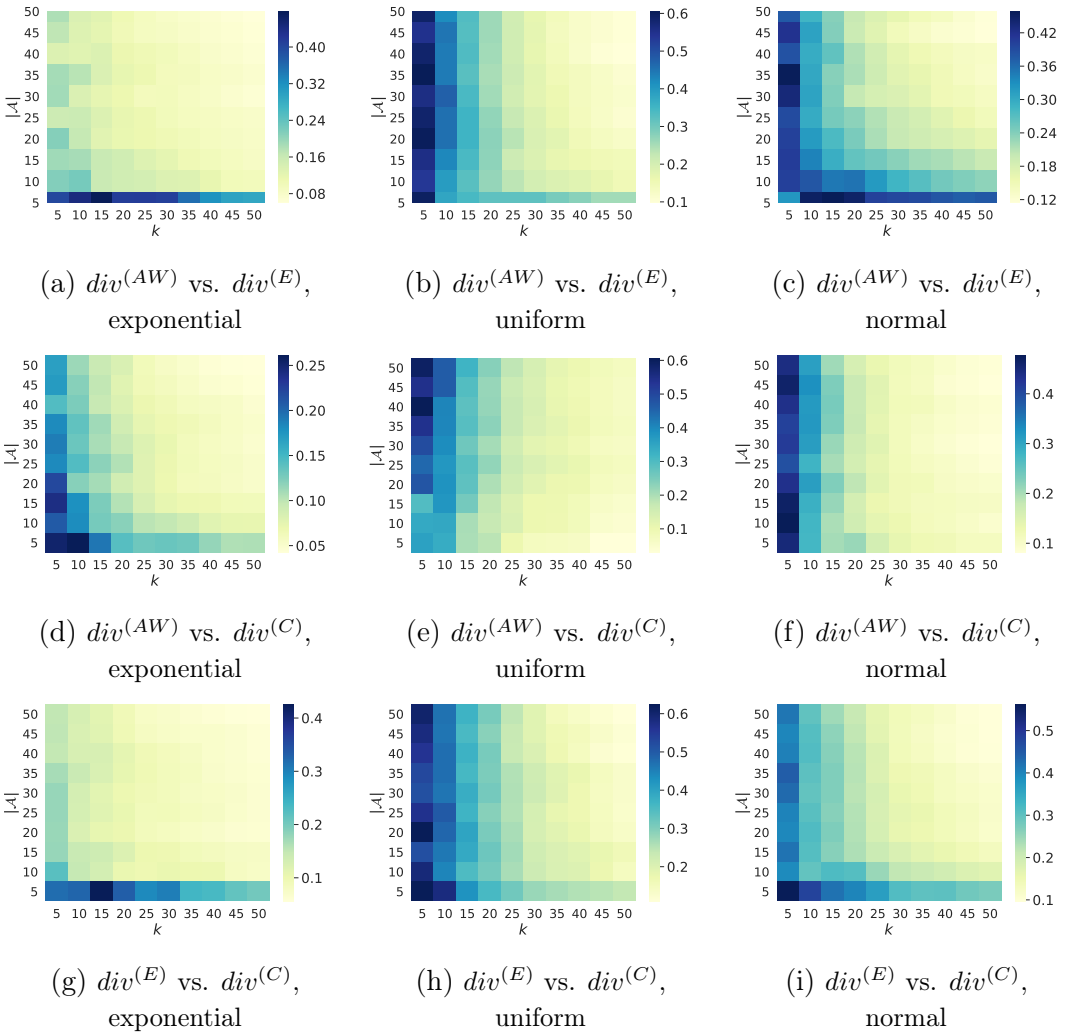
the data tuples are selected uniformly at random (like in this setting of evaluation), thus without an informed search strategy.

Analogously to what we observed in Figure 6.1, the trends corresponding to uniform distributions tend to be slightly smoother than the normal and exponential ones. The Hamming-based diversity function shows no significant variations with the size of attribute domains, which should however be ascribed to the fact that the size of the schema (i.e., $|\mathcal{A}|$ is set to 50) is significantly higher than the value of the Hamming ball radius (ξ).

Overall, our diversity functions reveal to be robust to variations in the attribute domains (especially for exponentially distributed attribute-values), which complements our remarks drawn from the previous evaluation.

6.7.1.3 Pairwise evaluation of diversity functions

Figures 6.3–6.4 show results on the average Jensen-Shannon divergences obtained by a pairwise comparison of our diversity functions, for various settings of the synthetically generated categorical data. Please note that for the setting of the radius in the Hamming-based diversity, we adaptively selected the radius in function of the size of the schema; besides for the sake of brevity of presentation of this comparative analysis of the diversity functions, this choice is also motivated by the outcomes of a sensitivity analysis that we conducted for the Hamming-based diversity, which is discussed later in this section.



(a) $div^{(AW)}$ vs. $div^{(E)}$, exponential

(b) $div^{(AW)}$ vs. $div^{(E)}$, uniform

(c) $div^{(AW)}$ vs. $div^{(E)}$, normal

(d) $div^{(AW)}$ vs. $div^{(C)}$, exponential

(e) $div^{(AW)}$ vs. $div^{(C)}$, uniform

(f) $div^{(AW)}$ vs. $div^{(C)}$, normal

(g) $div^{(E)}$ vs. $div^{(C)}$, exponential

(h) $div^{(E)}$ vs. $div^{(C)}$, uniform

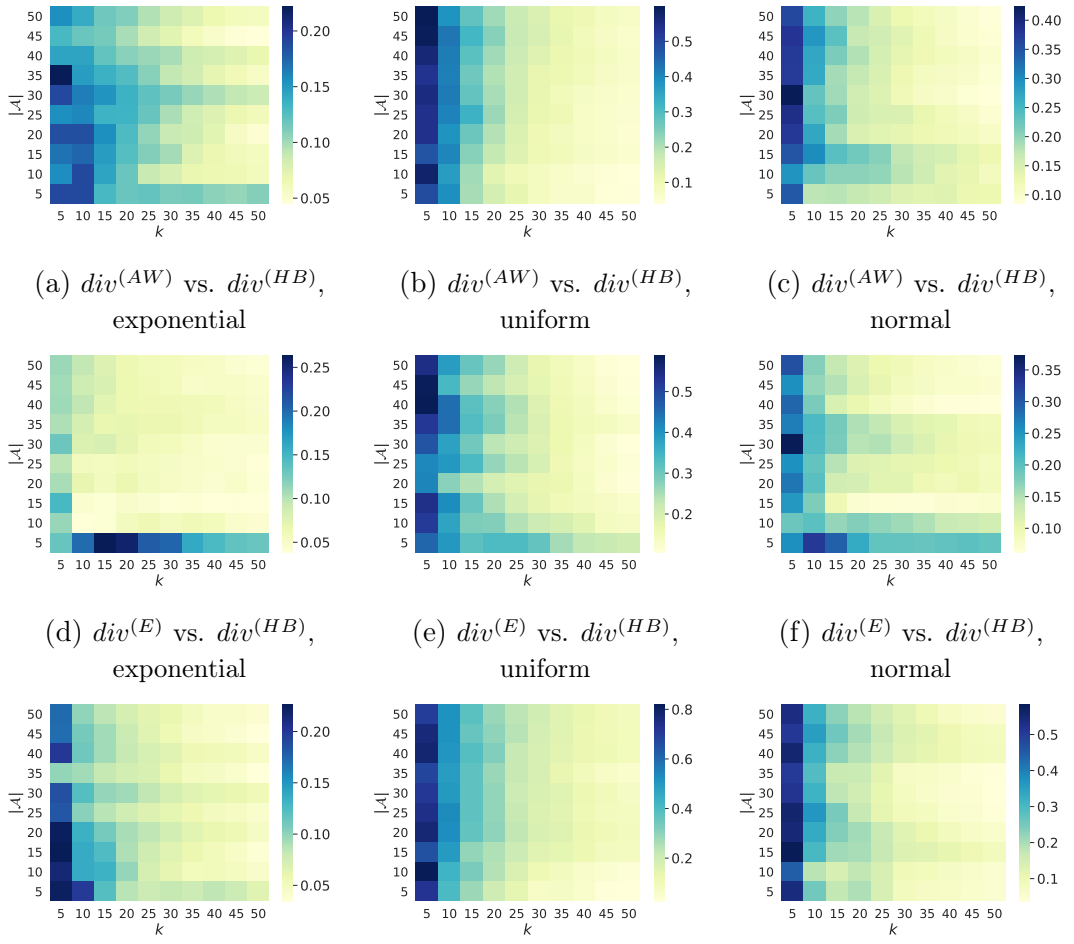
(i) $div^{(E)}$ vs. $div^{(C)}$, normal

FIGURE 6.3: Average Jensen-Shannon divergence of the probability distributions associated with the optimal k-sized sets for any two diversity functions, by varying k, size of the schema $\mathcal{A}$, and attribute-value distributions. The number of attribute symbols is set to 15.

At a first glance, it can be noted that the divergence generally varies within a relatively large interval — recall that the divergence is always non-negative and bounded by 1. This indicates that the set of categorical tuples that maximizes a particular diversity function can be very different to the optimal sets for the other diversity functions.

Looking at the scale of the values in each heatmap, the range of divergence values is generally wider for uniform attribute-value distributions (bounded by 0.6 in most cases), and narrower for exponential distributions, regardless of a particular pair of diversity functions being compared. Moreover, for exponential distributions, when the entropy-based diversity is involved in a comparison, we observe little variations and, generally, quite low (resp. high) values of divergence for number of attributes greater than (resp. equal to or lower than) five.

Another general remark is that the divergence tends to decrease as k increases and, to a less extent, as $|\mathcal{A}|$ increases. For low-mid regimes of k (i.e., up to 25-30), the divergence values lay on the corresponding mid-upper range, with little variations as different numbers of attributes are used. However, when the size of the schema is very



(a) $div^{(AW)}$ vs. $div^{(HB)}$, exponential

(b) $div^{(AW)}$ vs. $div^{(HB)}$, uniform

(c) $div^{(AW)}$ vs. $div^{(HB)}$, normal

(d) $div^{(E)}$ vs. $div^{(HB)}$, exponential

(e) $div^{(E)}$ vs. $div^{(HB)}$, uniform

(f) $div^{(E)}$ vs. $div^{(HB)}$, normal

(g) $div^{(C)}$ vs. $div^{(HB)}$, exponential

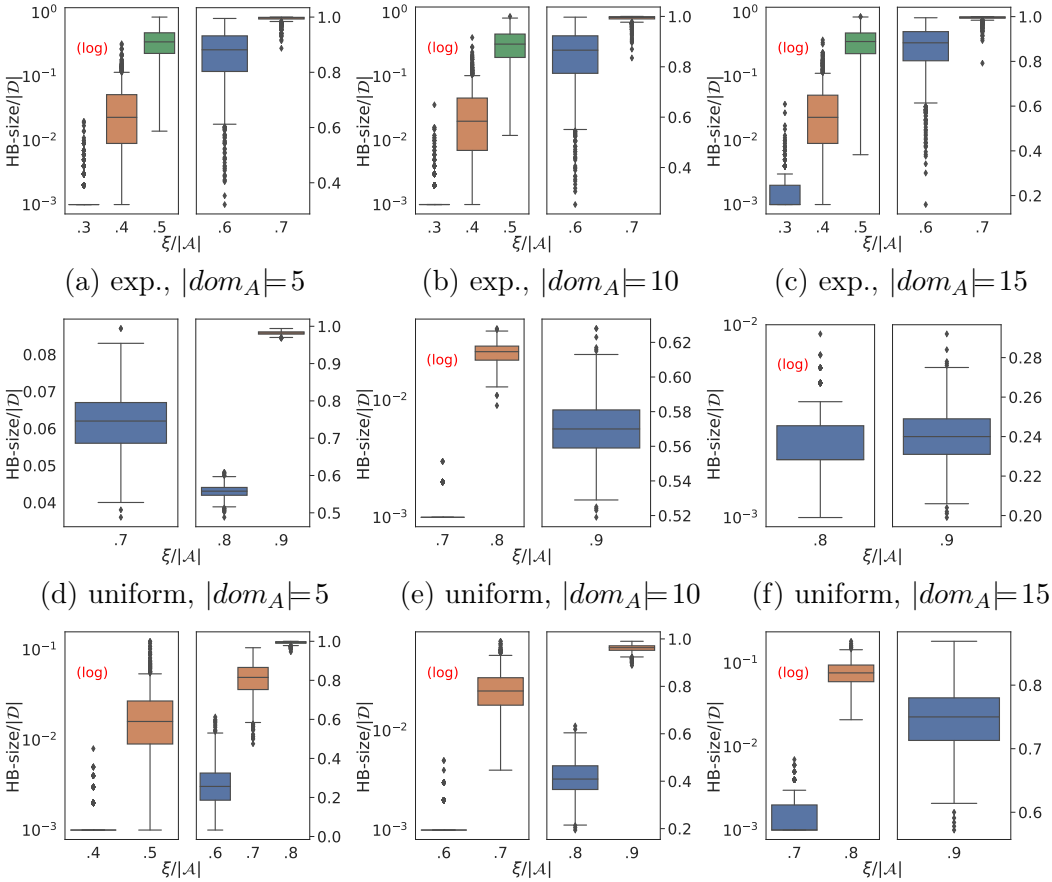
(h) $div^{(C)}$ vs. $div^{(HB)}$, uniform

(i) $div^{(C)}$ vs. $div^{(HB)}$, normal

FIGURE 6.4: Analogously to Figure 6.3, average Jensen-Shannon divergence of the probability distributions associated with the optimal k-sized sets for any two diversity functions. The radius of the Hamming-based diversity is set as a function of the number of attributattribute:es: $\xi = 0.4 \times |\mathcal{A}|$ for exponential distribution, and $\xi = 0.8 \times |\mathcal{A}|$ for normal and uniform distributions.

low (i.e., $|\mathcal{A}| \leq 5$) and regardless of the type of attribute-value distribution, it is likely to have very high divergence for any k. This hints at an interesting finding about the diversity functions, which are indeed capable of diversifying sets of categorical tuples in a different way, even for a small number of attributes.

Sensitivity of Hamming-based diversity to ξ As previously mentioned, we investigated about the effect of the radius on the behavior of $div^{(HB)}$. The setting of ξ in $div^{(HB)}$ is crucial, since too low values of the radius will lead to very small Hamming balls for most tuples in a given set, and hence to limited diversity of the set; by contrast, too high values of ξ will lead to Hamming balls that likely cover a large fraction of the tuple set. Instead, it might be preferable to set ξ in such a way to obtain as many relatively large and non-overlapping Hamming balls as possible, in order to better capture the diversity between the different tuples in the categorical dataset.



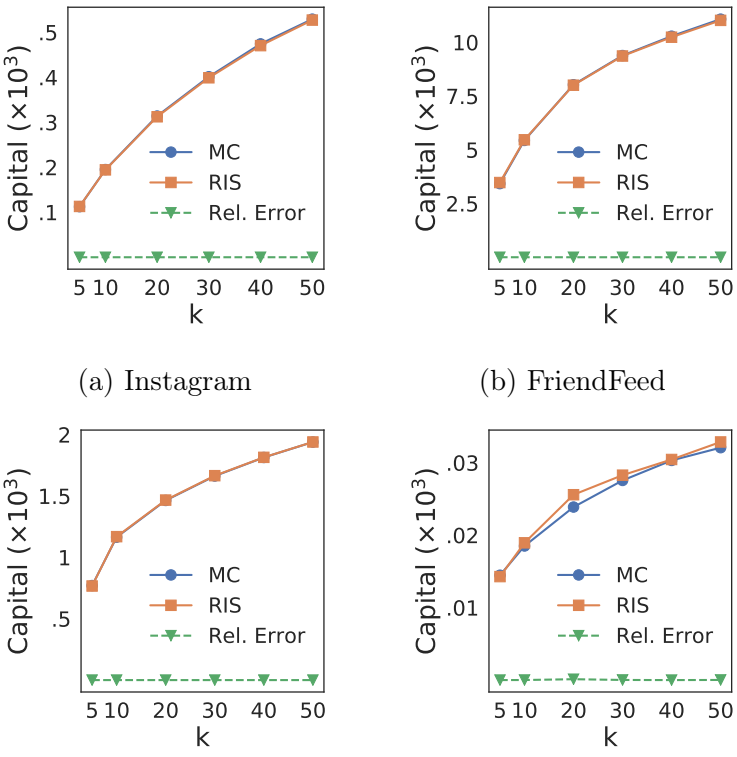
(a) exp., $|dom_A|=5$ (b) exp., $|dom_A|=10$ (c) exp., $|dom_A|=15$

(d) uniform, $|dom_A|=5$ (e) uniform, $|dom_A|=10$ (f) uniform, $|dom_A|=15$

(g) normal, $|dom_A|=5$ (h) normal, $|dom_A|=10$ (i) normal, $|dom_A|=15$

FIGURE 6.5: Distribution of the Hamming-ball sizes of the tuples in $\mathcal{D}$ (normalized by the number of tuples, i.e., $|\mathcal{D}|$) for selected values of the ratio between the radius ξ and the number of attributes $|\mathcal{A}|$, and for different attribute distributions and number of symbols (i.e., $|dom_A|$).

We computed the distribution of the individual Hamming-ball sizes associated with all tuples in an input dataset by setting the ratio $\xi/|\mathcal{A}|$ in $(0, 1]$ with increments of 0.1. Results are shown in Figure 6.5, where we present only the distribution boxplots corresponding to meaningful values of the ratio $\xi/|\mathcal{A}|$, i.e., we discarded boxplots corresponding to Hamming-ball sizes near to zero or the total number of tuples. Looking at the figure, we identify three main situations, which correspond to the type of attribute distributions. For the exponential case, we observe that $\xi/|\mathcal{A}| \geq 0.6$ yields too large Hamming balls, since the median size is above the 80% of the total number of tuples; by contrast, setting $\xi/|\mathcal{A}|$ within $[0.4, 0.5]$ leads to smaller, more preferable Hamming-ball sizes. When the attribute value distribution is uniform, high values of $\xi/|\mathcal{A}|$ lead to Hamming-ball sizes that can be very large or small, depending on the number of attributes; in particular, as the number of attribute symbols increases, the range of the boxplots decreases. Analogous remarks can be drawn for the normal distribution case, although it appears to be less sensitive to the number of attribute symbols if compared with the uniform distribution case (e.g., for the same number of attribute symbols (15), a ratio of 0.9 yields a boxplot whose median is above 0.7, whereas for the uniform distribution is around 0.24).



(a) Instagram (b) FriendFeed

(c) GooglePlus (d) Reddit

FIGURE 6.6: Capital estimation for seed sets obtained by ADITUM: RIS-based estimation by ADITUM vs. estimation by Monte Carlo simulations, with top-25% target selection.

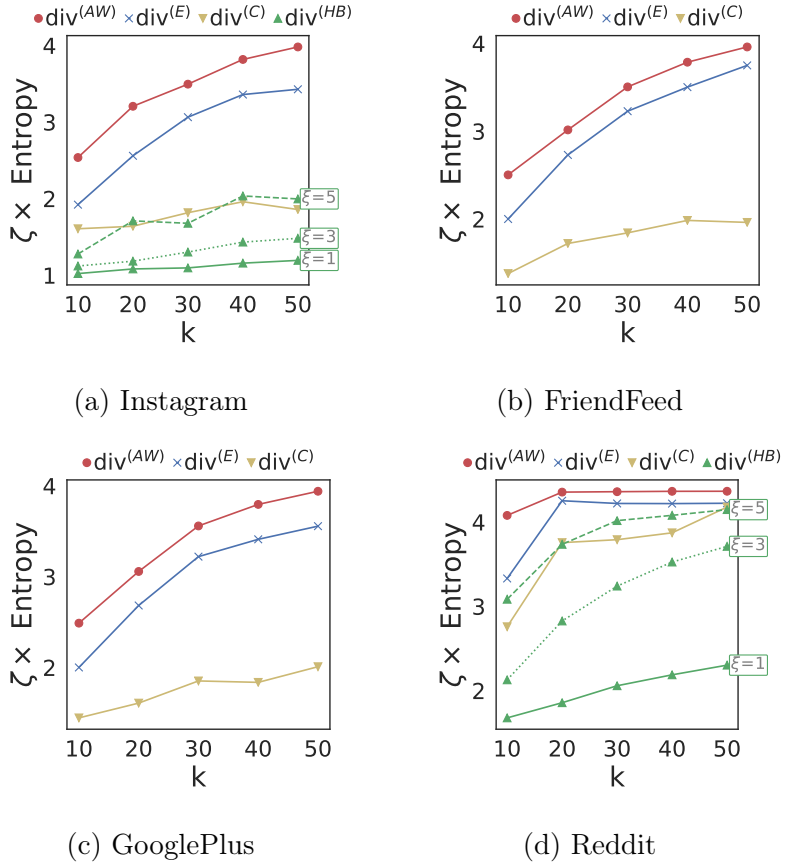
6.7.2 Stage 2 - Evaluation of ADITUM

We pursued three main evaluation goals, around which we organize the presentation of our results. First, we want to assess the significance of the estimation of capital produced by ADITUM (Section 6.7.2.1). Second, we want to understand the effect of each of the proposed definitions of diversity on the solutions provided by ADITUM (cf. Section 6.7.2.2). Third, we analyze the sensitivity of ADITUM w.r.t. the α parameter and the attribute distributions (Section 6.7.2.3).

6.7.2.1 Capital estimation

To begin with, we analyzed the *correctness* of the RIS-based estimation of the capital captured by the seeds discovered by ADITUM, which refers to Equation 6.11. By correctness, here we mean that the RIS-based estimation of capital in ADITUM should be close to the capital estimation provided by Monte Carlo simulation. To this purpose, we compared the ADITUM capital estimation (i.e., $\alpha = 1$) with the average capital score produced by a given seed set (provided by ADITUM) over 10 000 Monte Carlo runs.

As shown in Figure 6.6, for top-25% target selection and varying k, the two capital estimations are practically identical (i.e., relative error almost zero), even for higher k. The same holds for other settings of target selection. This confirms the correctness of the RIS-based estimation of capital in ADITUM.



(a) Instagram (b) FriendFeed

(c) GooglePlus (d) Reddit

FIGURE 6.7: Entropy of the seed sets obtained by ADITUM for various diversity functions, with top-25% target selection and $\alpha = 0$.

6.7.2.2 Effect of the diversity functions

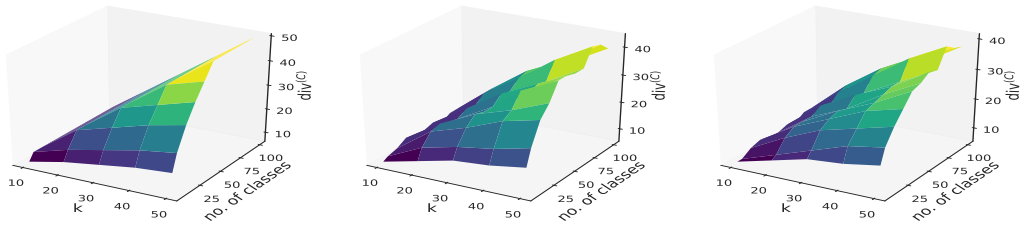
To understand the impact of the diversity notion on the ADITUM performance, we inspected the degree of diversification within the set of categorical tuples associated with the seed-set solution induced by each of the proposed functions. To this purpose, we first measured the entropy of the distribution of attribute-values in the profiles associated with a seed-set S, which is defined as

$$Entropy(S) = \sum_{a \in dom(S)} \frac{n_a}{\sum_{a' \in dom(S)} n_{a'}} \log \left(\frac{n_a}{\sum_{a' \in dom(S)} n_{a'}} \right).$$

Then, we multiplied the value of $Entropy(S)$ by a factor $\zeta = (1+\log(|dom|/\,|dom(S)|))^{-1}$ that penalizes more for smaller fraction of attribute-values covered by the profile set of S. Clearly, the higher the value of $\zeta \times Entropy(S)$, the better in terms of diversification the set S detected by ADITUM.

Results shown in Figure 6.7 indicate that $div^{(AW)}$ generally yields seed sets with higher entropy than the other diversity functions — in fact, to maximize $div^{(AW)}$, ADITUM tends to favor a uniform distribution of the attribute-values over the seed set. Also, $div^{(AW)}$ achieves higher coverage of the attribute domains (i.e., lower penalization factor ζ). The second best diversity function is $div^{(E)}$, which shows trends similar to $div^{(AW)}$ but with lower values of seed-set entropy for any k.

Conversely, $div^{(C)}$ and $div^{(HB)}$ lead to less diversified seed sets. This is actually



(a) $\alpha = 0.0$ (b) $\alpha = 0.5$ (c) $\alpha = 1.0$

FIGURE 6.8: *Class-based* diversity on Instagram by varying the number of classes, k, and α, with top-25% target selection.

not surprising since the class-based notion of diversity relies on the grouping of the profiles (i.e., coarser grain than at attribute-value level) and it is maximized when all profiles in S are chosen from different classes (i.e., $k \equiv h$, cf. Section 6.4.5), regardless of the distribution of their constituent attribute-values. In this regard, we further investigated how the combination of the budget k and the number of classes (into which the profile set is partitioned) affects the diversity value. Figure 6.8 shows that $div^{(C)}$ increases more rapidly with the increase in the number of classes w.r.t. k.

Also, the Hamming-based diversity $div^{(HB)}$ consistently behaves worse than $div^{(AW)}$ and $div^{(E)}$, while it is comparable to $div^{(C)}$ for radius set to five. As we have discussed in Section 6.7.1.3, $div^{(HB)}$ strongly depends on the setting of the radius ξ, and the diversity increases by increasing ξ since the union of the Hamming balls of the nodes in the seed set tends to grow.

It should be emphasized that the above results are complementary to the ones discussed in Section 6.7.1.3. In fact, while the latter focused on comparing the diversity functions in terms of the average Jensen-Shannon divergence between attribute-specific distributions, here we are interested in comparing the ability of seed-set diversification due to the functions in terms of entropy of the distributions over the whole domain of the attributes and also proportionally to the amount of covered attribute domain.

In the remainder of the result presentation we will refer to the attribute-wise diversity only. Our justification is that $div^{(AW)}$ (i) has shown effectiveness in the diversification of the seed set that is as good as or better than $div^{(E)}$, while outperforming $div^{(C)}$ and $div^{(HB)}$, (ii) it allows marginal gain computation that is clearly more efficient than the conditional entropy computation required in $div^{(E)}$, and (iii) it does not depend from additional a-priori knowledge like $div^{(C)}$ does, or parameters like $div^{(HB)}$ does.

6.7.2.3 Sensitivity to α

Parameter α allows for controlling the balance between the capital function and the diversity function in our problem (cf. Definition 7), i.e., the higher the value of α, the more ADITUM focuses on maximizing the capital rather than the diversity of the categorical data associated with the nodes in the current seed set. The setting of α can be experimentally provided to meet user-specified requirements. Nonetheless, in the following we discuss results of an analysis of sensitivity that we carried out to improve our understanding of how α impacts on the behavior of ADITUM.

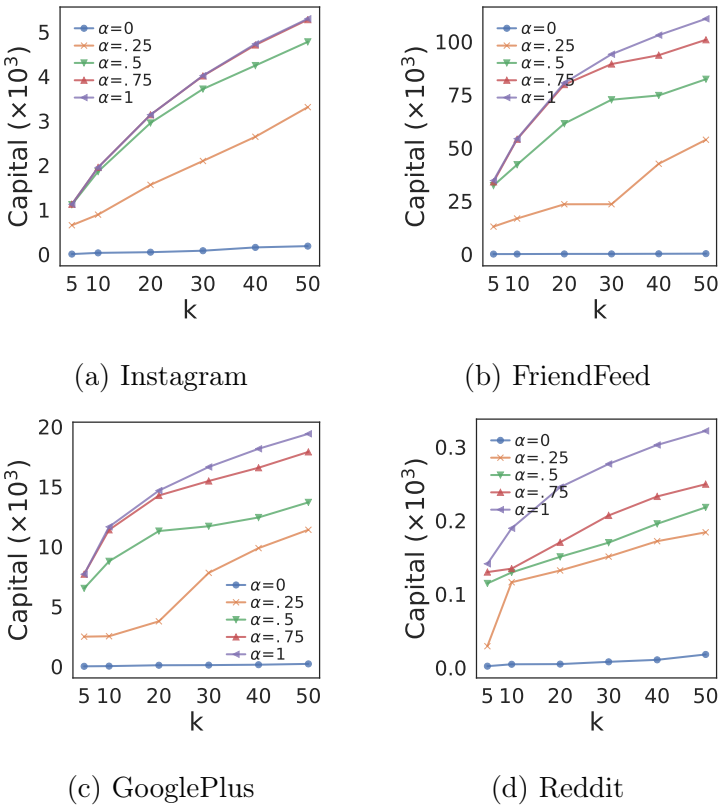
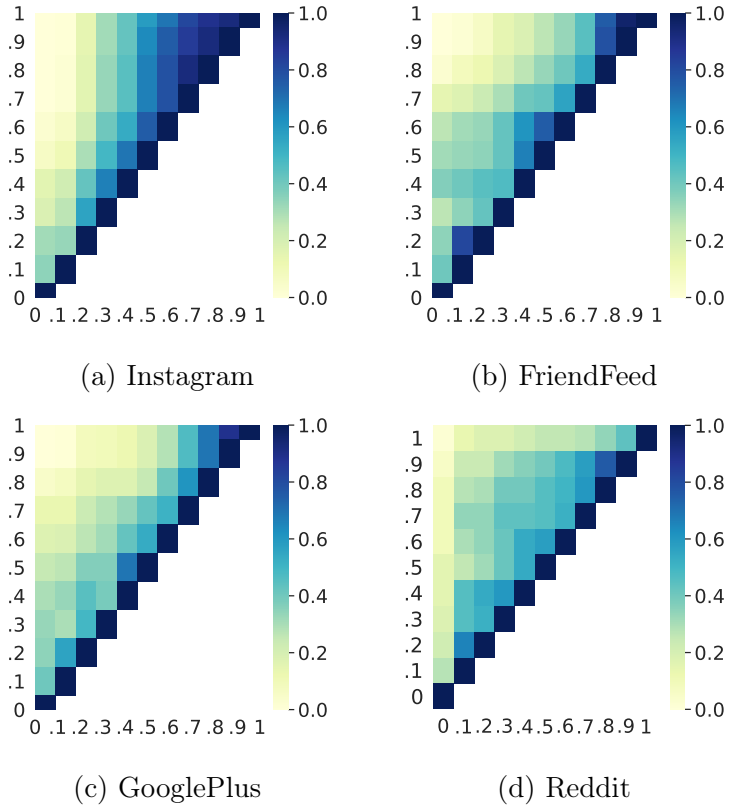


FIGURE 6.9: Expected capital, by varying $\alpha \in \{0, 0.25, 0.5, 1\}$, with $k \in [5, 50]$, top-25% target selection, and exponential distribution of attributes (except Reddit).

Effect on the expected capital We investigate the relation between α and the expected capital achieved by the identified seed set, for varying k. Results in Figure 6.9 show that, as expected for $\alpha = 0$, the capital remains very low and grows very slowly by increasing k. More interestingly, we observe that even mid-low values of α are sufficient to enable ADITUM to achieve a significant fraction of the capital that would be obtained with $\alpha = 1$ (i.e., without contribution of diversity); moreover, for $\alpha \simeq 0.75$, this gap tends to become quite low or even negligible. This is particularly evident in Instagram, whereby we observe no particular variations already with $\alpha = 0.5$, especially for low values of k. This fact might be ascribed to the relatively high connectivity of our Instagram network, which in fact corresponds to the maximal strongly connected component of the original graph [26]; consequently, there might be many solutions having high performance in terms of achieved capital. Similar considerations hold for FriendFeed and GooglePlus networks, where the capital gap becomes very small for $\alpha = 0.75$; compared to the situation observed in Instagram, we should consider that our FriendFeed and GooglePlus networks contain a certain amount of source and sink nodes, which inevitably hinder the spreading process upon which the capital estimation is based. Also, as the network connectivity becomes sparser, like in Reddit (which has an average in-degree significantly lower than the other networks), the capital gap due to a setting of α below one may remain high for any $k > 5$. Please note that results for top-5% and top-10% target selection thresholds, which are reported in the ***Appendix***-Figure D.2, confirm the trends observed in Figure 6.9.



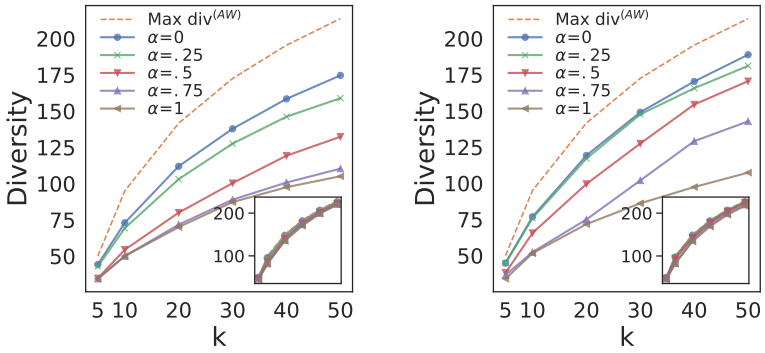
(a) Instagram (b) FriendFeed

(c) GooglePlus (d) Reddit

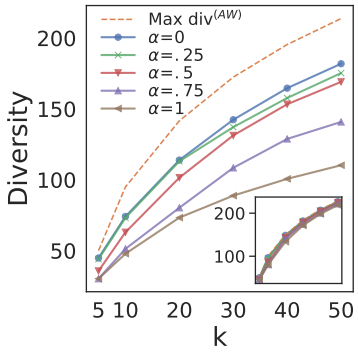
FIGURE 6.10: Normalized overlap of seed sets, by varying α within the range $[0, 1]$ (with increments of 0.1, on both x-axis and y-axis), and for $k = 50$, top-25% target selection, and exponential distribution of attributes (except for Reddit).

Evaluation of identified seed sets Heatmaps in Figure 6.10 show the pairwise overlaps of seed sets, normalized by k, for varying α. Focusing first on the overlaps between the seed set corresponding to $\alpha = 1$ (i.e., capital contribution only) and the ones corresponding to diversity at different degrees ($\alpha < 1$), the overlap decreases rapidly for lower α. This trend is less evident for Instagram because of its tighter connectivity than FriendFeed, GooglePlus and Reddit, as previously discussed. While overlaps always change for pairs of seed sets corresponding to different settings of α, it seems that the fading of overlaps becomes more gradual on networks with stronger small-world characteristics (i.e., GooglePlus). Please note that results corresponding to top-5% or top-10% target selection (shown in ***Appendix***, Figs. D.3), also confirm the variability in the seed set overlap, which is again more evident on the larger networks.

Effect of the attribute distribution The previous analysis refers to exponential distribution of the attributes. We observed however that the sensitivity of ADITUM to the setting of α becomes much lower for a uniform distribution of the attribute values. This prompted us to investigate the reasons underlying this behavior. To this end, we compared the diversity value associated to each seed set, by varying α and distributions, with the maximum possible value $div^*[k]$ (Equation 6.5; this is achieved when all the attribute values are equally distributed over the seeds. Not surprisingly, looking at the insets of Figure 6.11 that correspond to uniform distribution, we observe that the trends of seed-set diversity at varying α are all close to each other as well



(a) Instagram (b) FriendFeed



(c) GooglePlus

FIGURE 6.11: Exponential (main) vs. uniform (inset) distribution: attribute-wise of seed set for varying k and α, top-25% target selection, and comparison to maximum diversity value.

as to the maximum value. By contrast, using exponential distributions (main plots of Figure 6.11), it is evident that the slope of the diversity tends to decrease with higher α, thus increasing the gap with the maximum diversity curve. Moreover, different settings of the target selection threshold have no significant impact on the trends already observed for top-25% (results shown in ***Appendix***, Figs. D.6-D.5). In the following, results correspond to exponential distribution of the attributes, unless otherwise specified.

Usage recommendations for setting the value of α Here we aim to provide a summary of our major findings concerning the impact of α on the ADITUM behavior, in the form of practical guidelines for the setting of this parameter in real scenarios.

As a general remark, the user should take into account two main aspectattribute:s: the topological structure of the input network, and the distribution of the categorical attribute values. Indeed, networks showing high connectivity or with a large strongly connected component may favor the setting of α around mid values (i.e., $\alpha \simeq 0.5$) to achieve both capital and diversity that diverge the least from the respective values corresponding to the extremes of the range of α. In general, since real networks normally contain source and sink nodes, a good trade-off turns out to be setting $\alpha \simeq 0.75$. However, if the network structure is quite sparse, a significant gap in the capital should be expected as α moves away from one. Moreover, the above remarks are related to exponential distributions of the attribute values, while a uniform distribution would not significantly impact on the choice of α.

6.7.3 Stage 3 - Comparative evaluation with competing methods

In the last stage of evaluation, we comparatively evaluated ADITUM with the competing methods DTIM (Section 6.7.3.1) and Deg-D (Section 6.7.3.2).

6.7.3.1 Comparison with DTIM

We first evaluated the integration of the topology-driven diversity function of DTIM [26] into our RIS-based framework. We analyzed the normalized overlap of seed sets obtained by ADITUM and by the variant based on the topology-driven function, respectively. Figure 6.12 shows low-mid normalized overlap between the pairs of seed sets corresponding to most of the combinations of α. Remarkably, when discarding the contribution of capital in the respective objective functions (i.e., $\alpha = 0$ for both methods), the overlap is zero, or almost zero, which clearly confirms the expected, large difference between topology-driven and attribute-based notions of diversity. Moreover, the overlap tends to be lower for the largest networks, which are also sparser (and hence, more realistic) than our Instagram network.

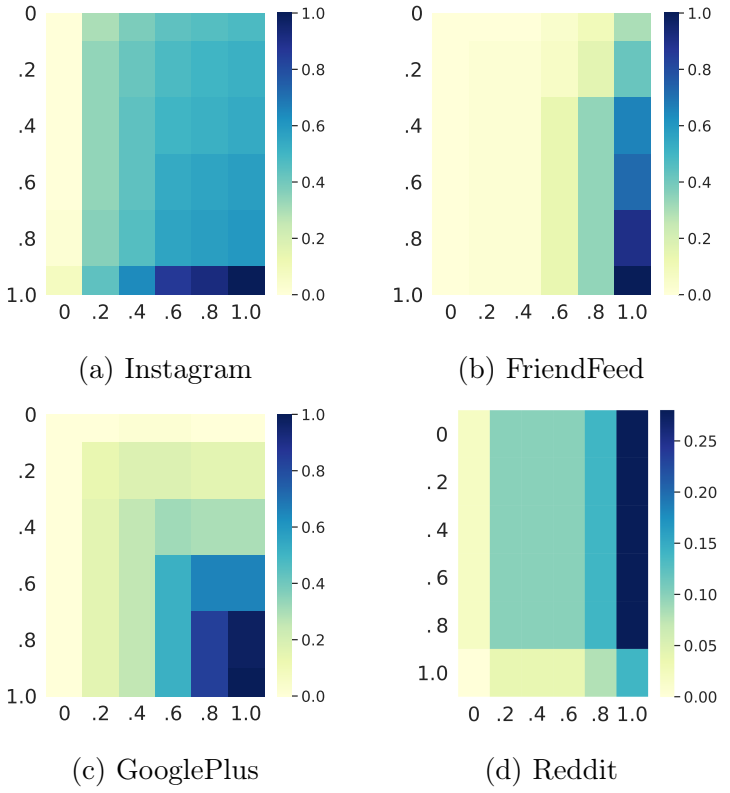
We also compared ADITUM and DTIM in terms of the expected capital and running time. In Figure 6.13, the insets show results of a Monte Carlo simulation (with 10 000 runs) for the estimation of the capital associated with the seed sets provided by each of the methods with $\alpha = 1$ (i.e., without the diversity contribution). It should be noted that, to ensure the highest estimation accuracy for DTIM without significantly worsening its efficiency, we set its path-pruning threshold η to 10^{-4}, which is the lowest value recommended in [26, 72]. We observe that ADITUM keeps a relatively small advantage over DTIM in terms of estimated capital. Nonetheless, as shown in the main plots of Figure 6.13, ADITUM outperforms DTIM in terms of running time, up to 3 orders of magnitude (e.g., in FriendFeed with $k \geq 10$), and this gap becomes even more evident as both k and the network size increase. The runtime advantage of ADITUM w.r.t. DTIM is actually not surprising, and can be clearly explained due to the differences in practical efficiency between the approach used by ADITUM and the approach used by DTIM, i.e., RIS-based TIM+ for ADITUM and SimPath for DTIM. In fact, it has been demonstrated in [189] that TIM+ consistently outperforms SimPath in terms of efficiency.

Note also that, while the running time of DTIM tends to increase linearly in k, for ADITUM it may even decrease with k: likewise TIM+, this is a result of the interplay of the main factors that determine the number of random RR-Sets.

6.7.3.2 Comparison with Deg-D diversity and attribute representation

As previously discussed in Section 6.6.2, we conducted a comparative evaluation with Deg-D to accomplish two goals. First, and more importantly, we want to understand how the seed sets produced by ADITUM differ from the ones produced by the variant of Deg-D with uniform function, i.e., Deg-DU, and the variant of Deg-D with weighted function, i.e., Deg-DW. The second aspect of evaluation does not involve ADITUM, rather it is concerned with an adaptation of our RIS framework to the numerical-attribute diversity used by Deg-D.

Figure 6.14 shows the normalized overlaps of seed sets obtained by ADITUM compared with either those obtained by Deg-DU or by Deg-DW. Results correspond to selected values of Deg-D parameter γ, which equals $1 - \alpha$, and refer to a generic, non-targeted IM scenario. Looking at the two heatmaps in the figure, there is evidence of the fact that the seed-set overlaps between ADITUM and Deg-D are always quite



(a) Instagram (b) FriendFeed

(c) GooglePlus (d) Reddit

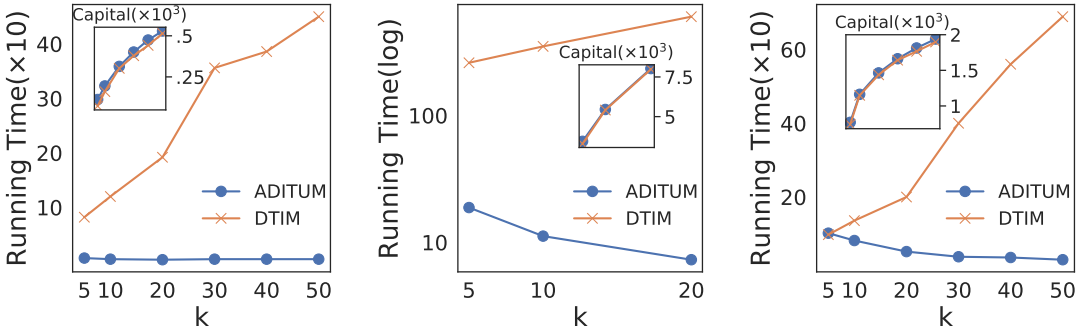
FIGURE 6.12: Topology-based vs. attribute-based diversity: Normalized overlap of seed sets, for selected values of α (on x-axis, corresponding to ADITUM, and on y-axis, corresponding to the ADITUM variant equipped with the global topology-driven diversity function of DTIM), $k = 50$, and top-25% target selection.

low, roughly in the range $0.28 \sim 0.43$. This holds for both variants of Deg-D and for any choice of γ, since Deg-D appears to have little sensitivity to the setting of γ (i.e., $1 - \alpha$) and the type of function.

Figure 6.15 shows results corresponding to numerical attribute representation and integration of Deg-DU and Deg-DW functions into our framework, here denoted as *RIS-U* and *RIS-W*. We set $\gamma = \alpha = 0.5$ to equally balance the contributions of diversity and spread in the methods' objective function. We observe that the seed-set diversity values are the same for the two methods in the uniform setting of the numerical-attribute diversity (i.e., Deg-DU and *RIS-U*). Conversely, in the weighted setting, the RIS-based diversity curve is only slightly below the Deg-DW curve. Also, the insets show very similar expected spread (on average over 10 000 Monte Carlo runs). Overall, this indicates flexibility of our RIS-based framework, which can also be properly adapted to integrate numerical-based diversity functions.

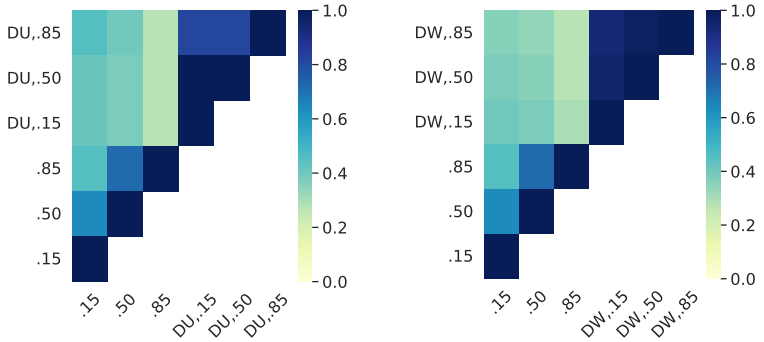
6.8 Chapter notes

We proposed a novel targeted influence maximization problem which accounts for the diversification of the seeds according to side-information available at node level in the general form of categorical attribute values. We defined a class of nondecreasing monotone and submodular functions to determine diversity of the categorical profiles associated to seed nodes. Our developed RIS-based ADITUM algorithm was compared to two IM methods, the one exploiting topology-driven diversity and the other one



(a) Instagram (b) FriendFeed (c) GooglePlus

FIGURE 6.13: ADITUM ($\epsilon = 0.1$) vs. DTIM ($\eta = 10^{-4}$): Running time in seconds (main plot) and expected capital (inset) for varying k, top-25% target selection and $\alpha = 1$.



(a) (b)

FIGURE 6.14: ADITUM vs. Deg-DU (left) and Deg-DW (right): Normalized overlap of seed sets, for $(1-\alpha) \equiv \gamma \in \{0.15, 0.5, 0.85\}$, $k = 50$, and top-100% target selection, on MovieLens.

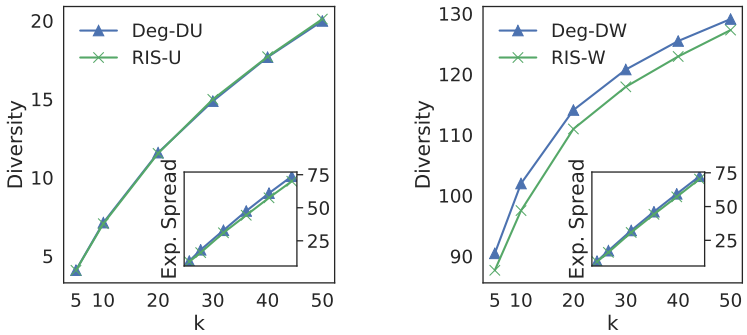


FIGURE 6.15: Deg-DU vs. RIS-U (left) and Deg-DW vs. RIS-W (right) on MovieLens numerical attribute representation: seed set diversity and, in the inset, expected spread by varying k, for $\gamma = 0.5$.

accounting for numerical-based diversity in IM. While showing different and more flexible behavior than the competitors, ADITUM takes the advantages of ensuring the RIS-typical theoretical-guarantee and computational complexity under a general,

categorical-based setting of node diversity. A further strength point of our diversity-sensitive framework lays on its versatility since ADITUM can easily be extended to incorporate other definitions of node diversity. In this regard, we plan to define diversity notions based on representation learning techniques, including network embedding methods.

Chapter 7

Conclusions

The research presented in this thesis has investigated the analysis of social influence and information diffusion in online social networks. We provided an extensive analysis on existing diffusion models to individuate the weaknesses that prevent them to capture the complexity of real-world propagation phenomena. Upon the recognition of such weaknesses, we designed a novel class of diffusion models, namely the F^2DLT (Friend-Foe Dynamic Linear Threshold Model) models. Even though they are inspired by the LT model, they show significant differences with this classic diffusion model. For instance,our diffusion models are focused around the notion of *trust*.

Also, our diffusion models provide a rich set of features, such as the *activation-threshold* and the *quiescence* functions. Together, these two features enable the possibility to represent phenomena as the *time-aware activation* and the *delayed response* of users with respect to the network's activation attempts. We also showed how, by tuning the above two aspects, we can configure different propagation environments, to which we referred as *biased* and *unbiased* scenario.

We believe that our models can pave the way to design more sophisticated methods to solve emerging and challenging problems in the domain of information diffusion. For instance, as regards problems concerned with the misinformation spread, our models can be a valid tool to account for the intricate patterns that drive the information consumption of polarizing information items, which might lead users to remain trapped into their information bubble.

We also devoted a lot of attention to one of the key-algorithmic problems in the context of information diffusion, namely the *influence maximization* problem. We assessed if, and to what extent, graph-decomposition algorithms can be used to support the identification of the most influential users in a social network. Surprisingly, in contrast with previous studies, we found out that the correlation between the spreading potential of a users and its position into the inner-most cores of a network weaker than expected. In fact, in our experiments, we provided evidence of the fact the state-of-the-art algorithms for IM do not necessarily pick their optimal spreaders within the inner most regions of a network. We also observed that with the adoption of more sophisticated methods for graph-decomposition, such as the *distance-generalized core decomposition* (DGC) algorithm, whose main feature is the ability to incorporate higher-order degree of information, we can easily detect the regions of the graph that are more densely populated with very effective spreaders.

Finally, another important contribution of this work is the definition of two optimization problems that can be regarded as variants of the classic IM problem. More specifically, we addressed a targeted influence maximization problem, where the objective function takes into account diversity of the selected seed set. In the first problem, i.e., the ***D**iversity-sensitive **T**argeted **I**nfluence **M**aximization* (DTIM), the diversity seed nodes is defined with respect to their topological properties. To this purpose, we

formulated two different measures to quantify diversity: the *local diversity* and *global diversity*. We also designed two algorithmic solutions, i.e., the L-DTIM and G-DTIM, to effectively solve the DTIM problem.

The second problem, i.e., **A***ttribute-based* **DI***versity-sensitive* **T***argeted Infl***U***ence* **M***aximization* (ADITUM), is similar to the previous one, but it considers a different setting. In fact, in contrast to the DTIM problem, the ADITUM problem assumes that each node is associated with a set of categorical attributes. These attributes define what we called the *profile* of a user, with respect to which we measure the seeds' diversification. To this purpose, we proposed a class of diversity functions with four separate definitions. Each one tackles diversity from a different perspective. Nonetheless, every proposed function has the convenient properties of monotonicity and submodularity. The above definitions are then embedded within our algorithmic solution, i.e., the ADITUM algorithm, which is formulated after the state-of-the-art RIS framework [20]. We also showed the superiority of our proposed solution over several competing methods on different axes, i.e., effectiveness, efficiency and flexibility. Especially, we recognize the versatility of our approach as one of its key feature. In fact, our algorithm can easily accomodate other diversity functions and, as long as they are monotone submodular, we can take advantage of the approximation guarantee ensured by the greedy framework our algorithm is designed upon.

A future direction of this research would be to explore the opportunity of extending the notion of diversity using other paradigms. For instance, based on representation learning techniques, such as network embedding methods. These methods are able to provide a low-dimensional, vector-based, representation of the graph by the means of state-of-the-art machine-learning methods. Therefore, each node is associated with a vector that inherently characterizes its topological properties. One can easily envisage an extension to our proposed diversity-sensitive influence maximization problems. In fact, nodes can be also diversified with respect to their vector-based representation, previously discretized as our framework requires, and then incorporated within the ADITUM algorithm.

Another direction would be to extend our set of diversity functions, renouncing on their submodularity property. This brings another challenging research question, since, as we have largely discussed in Chapter 2, a submdoular function enables the definition of effective solutions to the IM problem. Under this new setting, we need to completely redesign our algorithm, as we can no loner rely on the approximation bound ensured by the RIS framework. Therefore, we believe it would be interesting to reformulate our approach after different paradigms, such as the *sandwitch approximation* discussed in [139], which is specifically designed to optimize non-submodular functions.

Finally, as a general remark, we believe that studies on social influence should address novel social media platforms, since their are essential to understand how it is changing the way we interact with each other. In factg, the communication tools provided to their users by modern social media platforms like Instagram or TikTok, which are mostly based on visual contents (e.g., photos, short videos), are extremely different from the earliest social media platforms. It implies that social influence may manifest itself in different ways. Due to the impact that these platforms have on the real life of people, especially the youngsters, we encourage the computer, social, or data science community to embrace this challenge, so to gain a deeper knowledge on this complex phenomena.

Appendices

App. A

Complex Influence Propagation

A.1 Additional details on the properties of the models

Figure A.1 shows an example of serialization for a *spC-F²DLT* diffusion graph with time horizon set to 2. Dashed lines correspond to the edges in the original graph, whereas solid lines correspond to the edges in the resulting serialized graph. Each of the four nodes in the original graph is replicated as a triple on each of the two time-layers. Triples act as "connectors" between two consecutive time-layers.

Analogously to the reduction of *spC-F²DLT* to *H-CLT*, we can conveniently devise a notion of "connector" component between any two consecutive layers, shown in Figure A.2, which in the case of *npC-F²DLT* needs to account for node deactivations.

Example 6 shows a selection of possible configurations for the component utilized in competitive models shown in Figure 3.7, in order to prove the correctness of the set of constraints in Equation 3.6.

Example 6. *In the example of Figure A.3, we assume that node v in the original graph needs three consecutive time steps to reach the unit value for its threshold.*

On the right side of each subfigure, there are the additional node replicacomplex:s: $\langle v_t^{3,r1}, v_t^{3,r2}, v_t^{3,r3}\rangle$*, where* $v_t^{3,r3}$ *has the maximum value for the activation threshold.*

FigureA.3a represents the case when the node has already reached the maximum value of its threshold and its in-neighbors are only able to activate the first two replicas, which is not enough for making v change its activation campaign. We need thus to verify that:

$$\overbrace{w^{3,r1} + w^{3,r2}}^{red} < \overbrace{w_1^{13} + w_2^{13} + w_3^{13}}^{green} \tag{A.1}$$

$$w^{3,r1} + \cancel{w^{3,r2}} < w_1^{13} + \cancel{w_2^{13}} + w_3^{13} \tag{A.2}$$

$$w^{3,r1} - w_1^{13} < w_3^{13} \tag{A.3}$$

Note that the inequality in (A.3) holds (cf. Equation 3.6(e)).

Figure A.3b shows the case when v is active for just two consecutive time steps. Therefore, the configuration on the right side of the figure is enough to activate v in favor of the red-campaign. Therefore, the following inequality must hold:

$$\overbrace{w^{3,r1} + w^{3,r2}}^{red} > \overbrace{w_1^{13} + w_2^{13}}^{green} \tag{A.4}$$

$$w^{3,r1} + \cancel{w^{3,r2}} > w_1^{13} + \cancel{w_2^{13}} + w_3^{13} \tag{A.5}$$

$$y_1 > x_1 \tag{A.6}$$

Above, inequality in (A.6) holds as given in Equation 3.6(a).

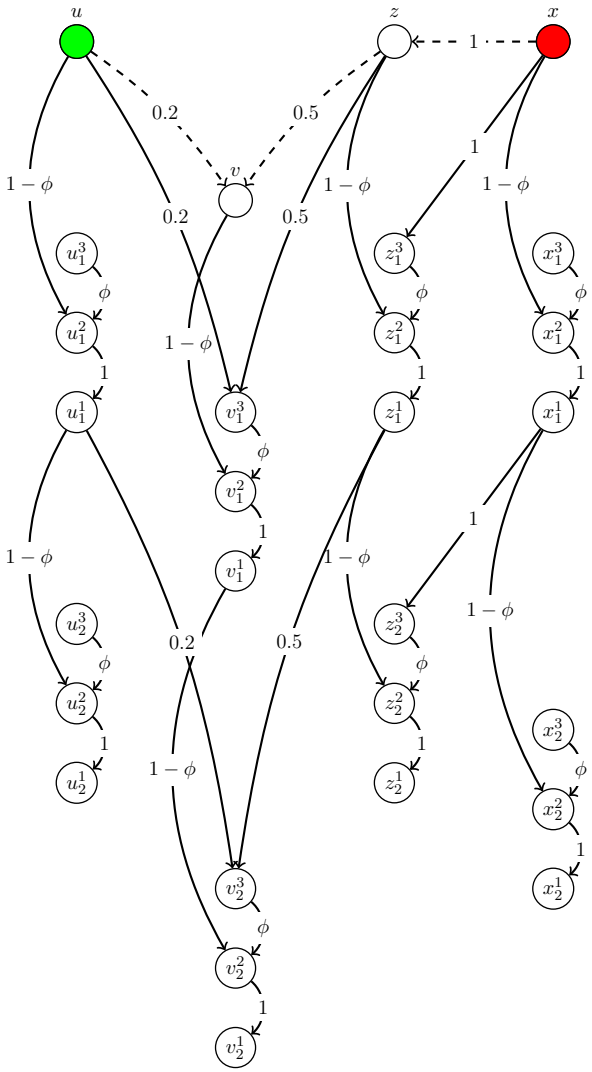


FIGURE A.1: Serialization of the diffusion subgraph involving nodes u, v, z, x,
under *spC-F²DLT*, with time horizon set to 2. Symbol ϕ denotes a value
chosen at random in (0.5,1].

Figure A.3c shows the case when the node v switched from one campaign to the opposite in the middle of the three consecutive time steps. In this case the activation of node $v_t^{3,r1}$ *must guarantee the change of activation campaign. So the following inequality must hold:*

$$\overbrace{w^{3,r1} + w_2^{13}}^{red} > \overbrace{w_1^{13} + w_3^{13}}^{green} \tag{A.7}$$

Indeed the validation is straightforward, because all the summations on the left side in A.7 are by definition greater than the one on the right side.

Figure A.3d shows the case when at time step t none of the in-neighbors of v is able to activate it. In this case, the node will keep its previous state, and it will do it only after the activation of the corresponding replica at the very previous time, this allow us to guarantee the sequentiality of the whole process. ■

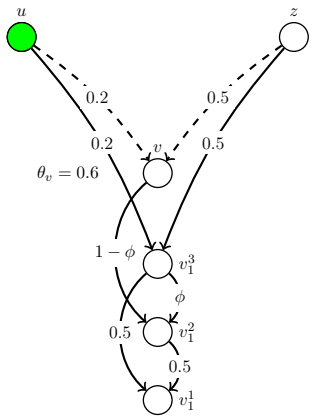


FIGURE A.2: Focus on a connector from Figure A.1 and its adaptation for the
npC-F²DLT model.

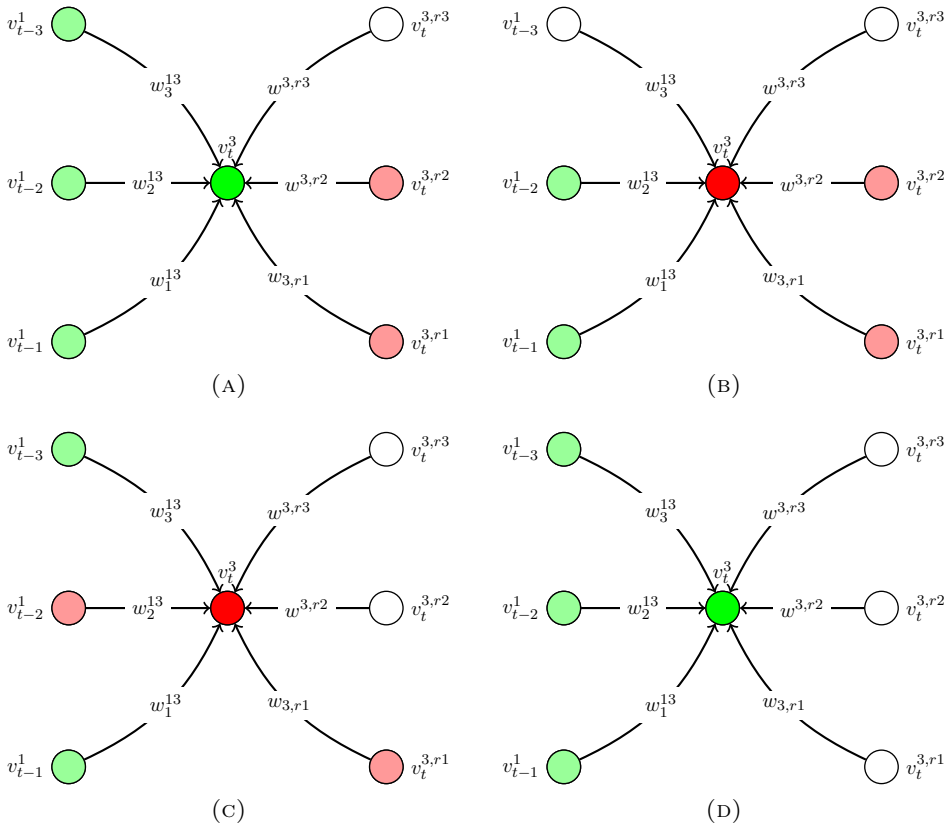
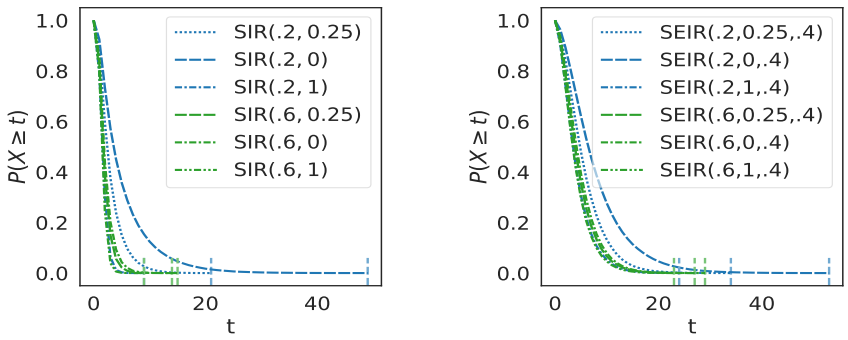
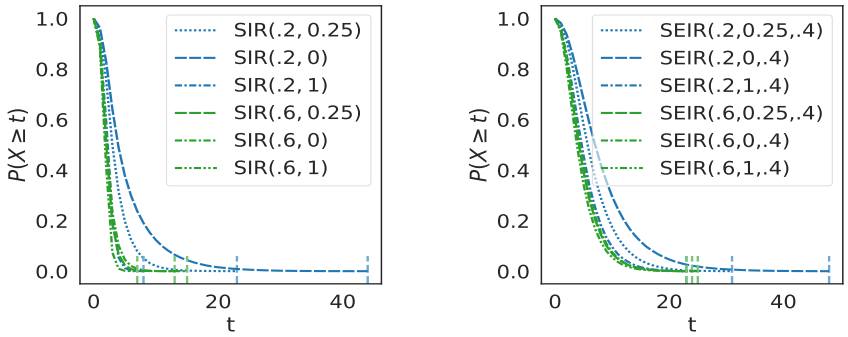


FIGURE A.3: Possible configurations for the connector in Figure 3.7.



(a) Epinions



(b) Slashdot

FIGURE A.4: Complementary cumulative distribution functions of node infections for SIR and SEIR with $\beta \in \{0.2, 0.6\}$, $\gamma \in \{0, 0.25, 1\}$), and $\sigma = 0.4$, using $k = 50$ and strategy I-Sources.

A.2 Additional details on epidemic models

Figure A.4 provides a focus on the behavior of SIR and SEIR models in terms of varying parameters (i.e., transmission rate β, recovery rate γ, and incubation rate σ) for the analysis discussed in Section 3.5.3.1.

We observe that, for both models, most of the infections tend to occur at the early time steps of the propagation as β increases. On the other hand, higher values of γ yield cascades that show a smoother decay over time, and consequently they last longer than those corresponding to smaller γ.

TABLE B.1: Maximum peak-number and number of different contours (first column) vs. maximum core-index and number of different cores (second column)

	k-peak	k-core
FF	160 / 30	160 / 160
Ig	48 / 14	48 / 47
DB	113 / 45	113 / 47
Ep	85 / 22	85 / 85
Net	31 / 15	31 / 13
Tw	23 / 14	24 / 24

App. B

Topological characterization of the most influential nodes

B.1 Seed selection order

To begin with, we analyzed the selection order of seeds discovered by each IM method in relation to the decomposition index values. decomposition algorithm.

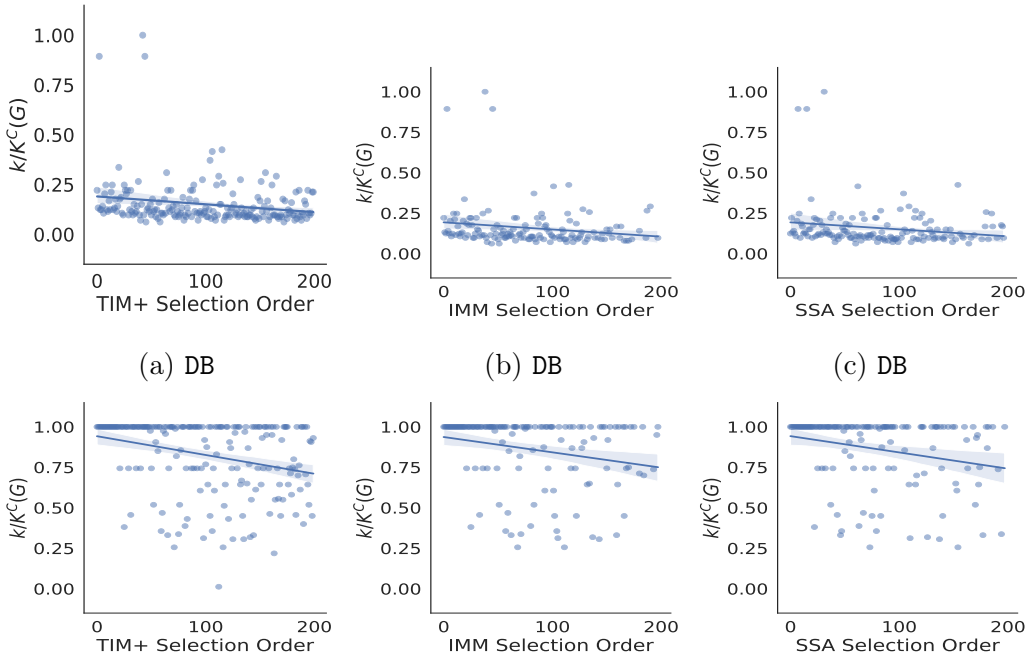
Considering first the k-core decomposition, Figure B.1 shows the core-index (divided the degeneracy of graph) for the first 200 seeds — computed by TIM+, IMM, and SSA, respectively — according to their selection order, i.e., the iteration corresponding to the insertion of a node into the seed set being computed. Results in B.1 refer to the LT model. It is clear that the diffusion model has no perceivable impact on the seed selection process It should be noted that the choice of the diffusion model has no perceivable impact on the seed selection process.

We observe that the way each algorithm locates the seed nodes through the cores of the network is invariant with respect to the propagation model.

B.2 Effect of graph decomposition

The analysis carried out in the previous section is extended to other decomposition methods. Results are reported in Figure B.2.

We can conclude that, the particular choice of the diffusion model does not play a crucial role in determining the seeds location within different regions of a graph. In fact, this consideration applies consistently with respect to every considered network and decomposition algorithm.



(a) DB (b) DB (c) DB

(d) FF (e) FF (f) FF

FIGURE B.1: Normalized core-index ($k/K^C(G)$) of the first 200 seeds computed by (a,d) TIM+, (b,e) IMM, and (c,f) SSA, with respect to the LT model.

B.2.0.1 Consistency between the algorithms

In Figure B.5-B.6 we show the same results as the ones shown in the previous section, but with respect to the LT model and the other IM algorithms considered in Chapter 4.

No particular different trend can be noted between the different algorithms. This result confirms, once again, that all the state-of-the-art IM algorithms share a similar behavior of seed selection.

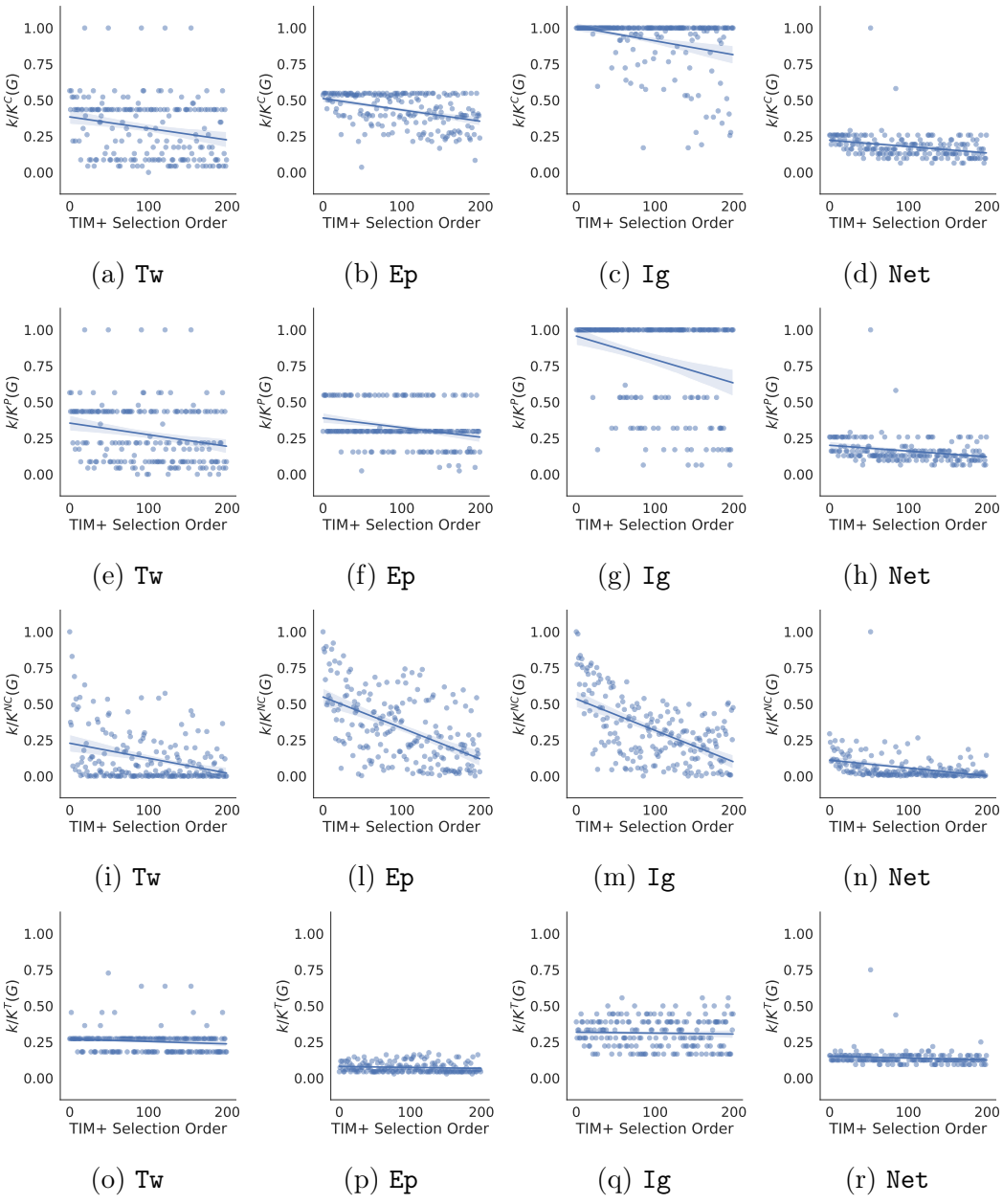
B.2.0.2 Characterization of the Cores/Contours

In this section we report the results related to the contour distributions, which complement the results on core distributions we presented in Section 4.5.2.

Table B.1 shows that the k-peak decomposition provides in general fewer distinct contours than distinct cores. Moreover, Figure B.3 shows that the k-Peak decomposition tends to induce skewer distributions than the one induced by the core decomposition.

Figure B.4 shows the results related to the classification of the edges into the **outward** and **inward** classes as defined in Section 4.5.2.

The same trends observed for the k-Core decomposition also apply in this context.



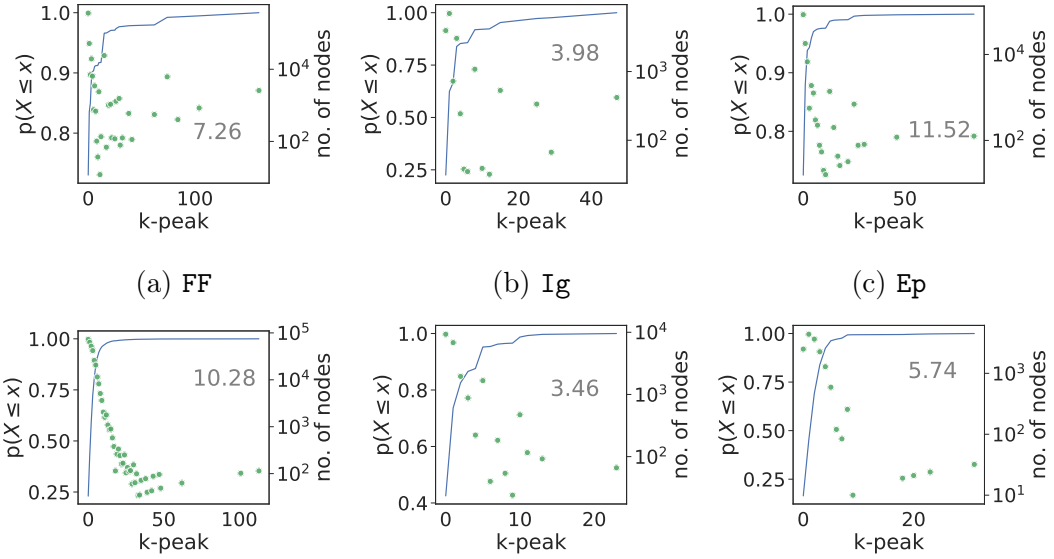
(a) Tw (b) Ep (c) Ig (d) Net

(e) Tw (f) Ep (g) Ig (h) Net

(i) Tw (l) Ep (m) Ig (n) Net

(o) Tw (p) Ep (q) Ig (r) Net

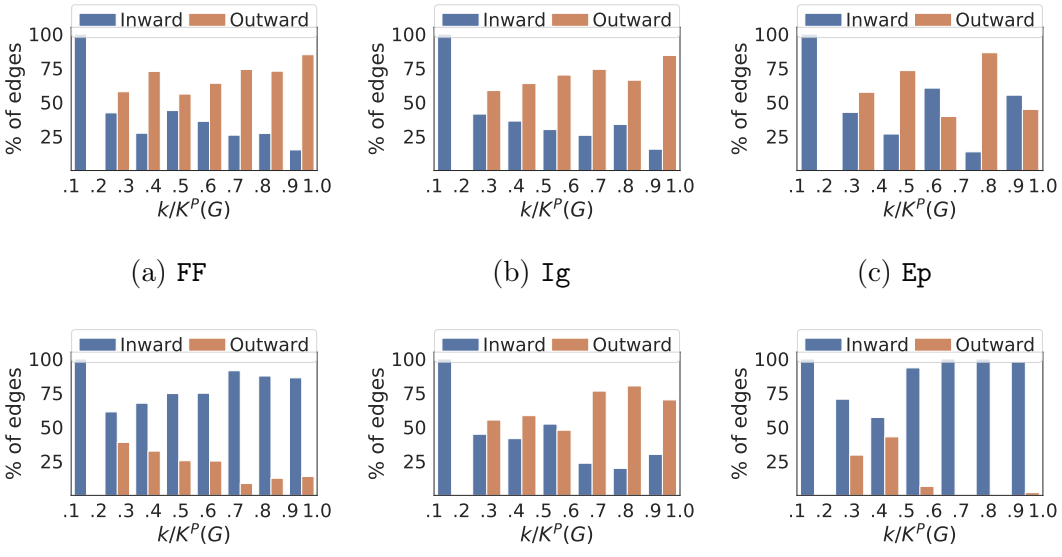
FIGURE B.2: From top to bottom, normalized core-index ($k/K^C(G)$), peak-number ($k/K^P(G)$), neighbor-coreness ($k/K^{NC}(G)$), and truss-index ($k/K^T(G)$) of the first 200 seeds computed by TIM+, under the LT model.



(a) FF (b) Ig (c) Ep

(d) DB (e) Tw (f) net

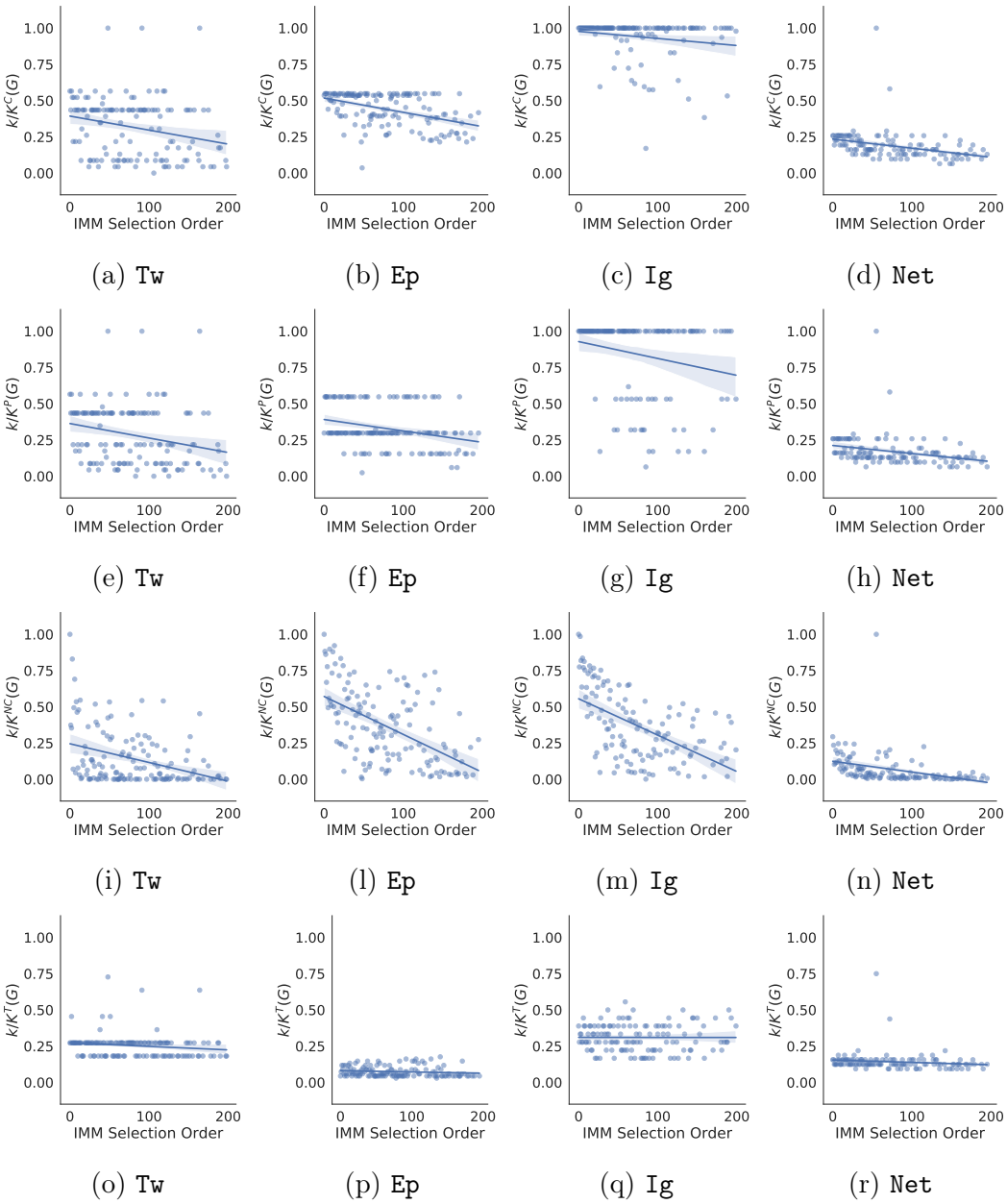
FIGURE B.3: Distribution of nodes over the peak-numbers of the network. Each plot shows, for every core-index k (x-axis), the number of nodes with peak-number at most k on the leftmost y-axis, and the cumulative distribution of core-index on the rightmost y-axis. Also, the skewness of the distribution is reported inside each plot.



(a) FF (b) Ig (c) Ep

(d) DB (e) Tw (f) Net

FIGURE B.4: Percentage of inward and outward edges vs. normalized peak-number $k/K^P(G)$. The i-th percentage bar ($i = 1..9$) corresponds to edges such that the source node has normalized core-index in $(x_i, x_{i+1}]$, upon a segmentation of the x-axis values into ten intervals $(x_1, x_2], \ldots, (x_9, x_{10}]$.



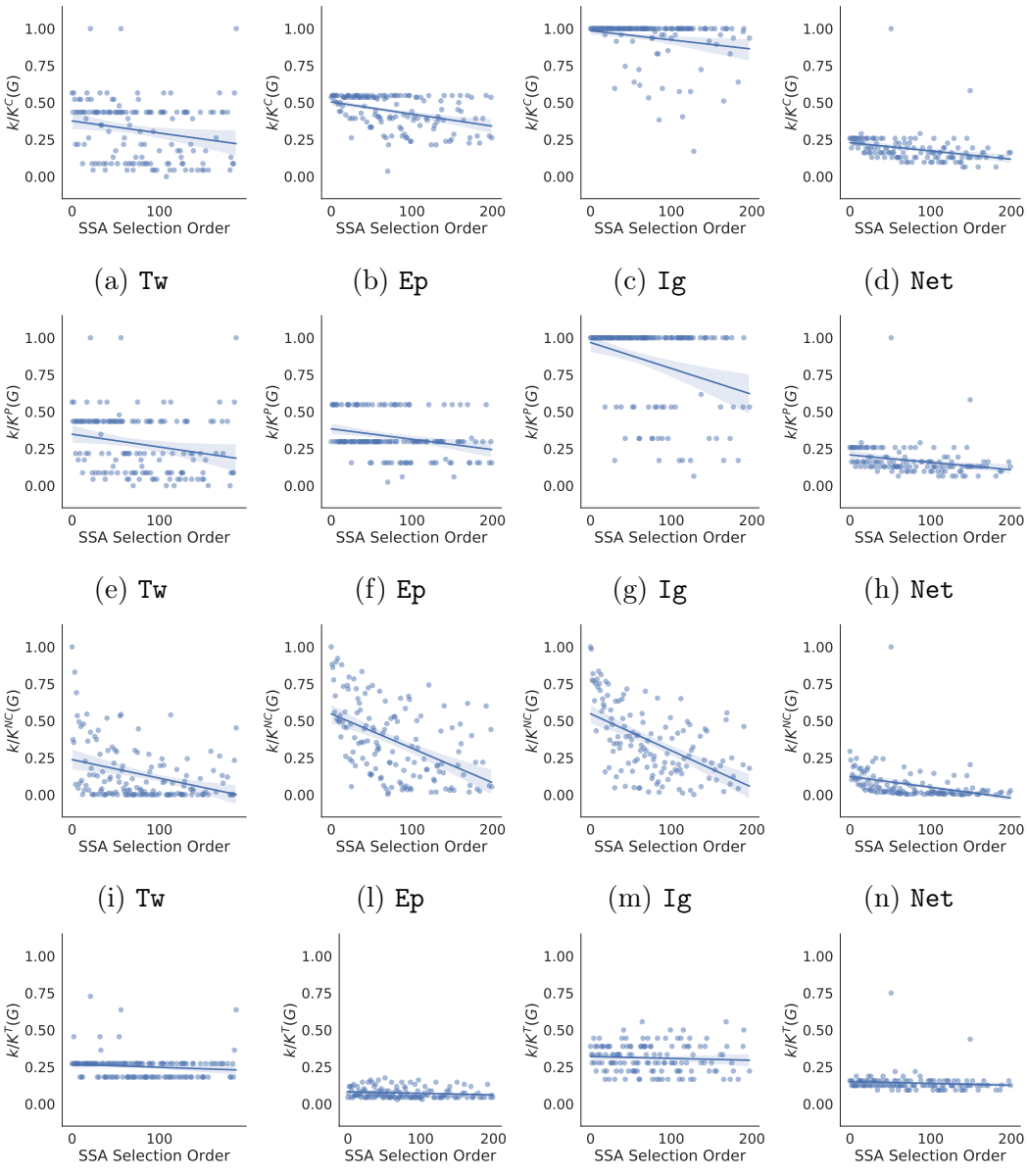
(a) Tw (b) Ep (c) Ig (d) Net

(e) Tw (f) Ep (g) Ig (h) Net

(i) Tw (l) Ep (m) Ig (n) Net

(o) Tw (p) Ep (q) Ig (r) Net

FIGURE B.5: From top to bottom, normalized core-index ($k/K^C(G)$), peak-number ($k/K^P(G)$), neighbor-coreness ($k/K^{NC}(G)$), and truss-index ($k/K^T(G)$) of the first 200 seeds computed by IMM, under the LT model.



(a) Tw (b) Ep (c) Ig (d) Net

(e) Tw (f) Ep (g) Ig (h) Net

(i) Tw (l) Ep (m) Ig (n) Net

(o) Tw (p) Ep (q) Ig (r) Net

FIGURE B.6: From top to bottom, normalized core-index ($k/K^C(G)$), peak-number ($k/K^P(G)$), neighbor-coreness ($k/K^{NC}(G)$), and truss-index ($k/K^T(G)$) of the first 200 seeds computed by SSA, under the LT model.

Algorithm 6 Monte Carlo Estimation of Capital

Input: A graph $G = (V, E, b, \ell)$, a target selection threshold $L \in [0, 1]$, seed set S, number of Monte Carlo iterations I_{MC}
Output: Capital $C(\mu(S))$

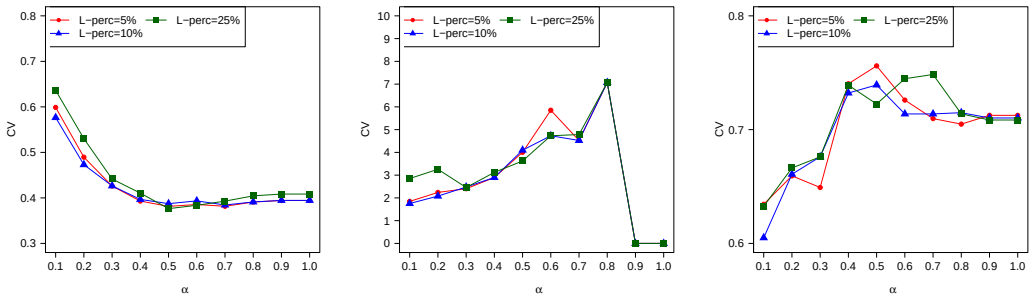
```
curr_C ← 0
for u ∈ S do
    u.isActive ← true
end for
for j = 1 to I_MC do
    for v ∈ V \ S do
        v.isActive ← false
        v.receivedInf ← 0
        ϑ_v ← −1
    end for
    temp ← S
    while temp ≠ ∅ do
        u ← temp.remove(0)
        for v ∈ N^out(u) ∧ v.isActive = false do
            v.receivedInf ← v.receivedInf + b(u, v)
            if ϑ_v = −1 then {node v has been reached for the first time during the current simulation}
                choose ϑ_v ∼ U[0, 1]
            if v.receivedInf ≥ ϑ_v then
                v.isActive ← true
                temp ← temp ∪ {v}
                if ℓ(u) ≥ L then
                    curr_C ← curr_C + ℓ(v)
        end for
    end while
end for
return curr_C/I_MC
```

App. C

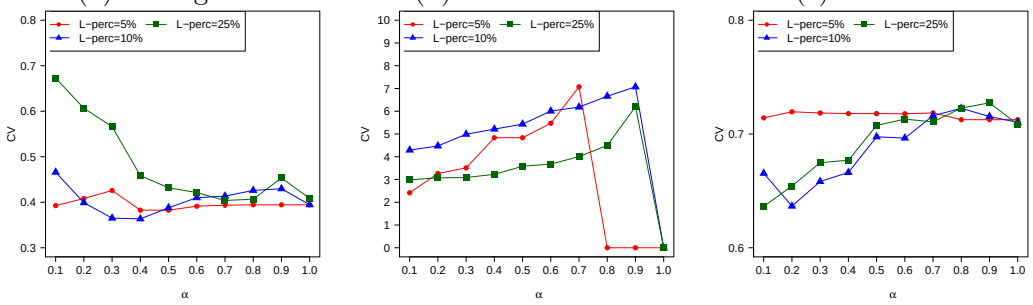
Topology-based Diversity-sensitive Targeted Influence Maximization

C.1 Monte carlo estimation of capital

Algorithm 6 sketches the Monte Carlo procedure of simulation of the LT diffusion process for estimating the capital associated with the target nodes that are finally activated by a given seed set.



(a) outdegree (b) betweenness (c) coreness



(d) outdegree (e) betweenness (f) coreness

FIGURE C.1: Coefficient of variation (CV) of topological properties of identified seed nodes, with $k = 50$, by varying α and *L-perc*, on `GooglePlus`: (a)–(c) L-DTIM, (d)–(f) G-DTIM.

C.2 Note on LurkerRank for targeted IM

LurkerRank does not require any information other than the network topology, in which node (user) relationships are asymmetric and indicate that one node receives information from another one. The actual meaning of "received information" can depend on the specific context of network evaluation; in general, it refers to either a social graph (i.e., $(u, v) \in E$ means that v *is follower* of u) or an interaction graph (e.g., v *likes* or *comments* u's posts); LurkerRank has been indeed evaluated on both scenarios [180, 181].

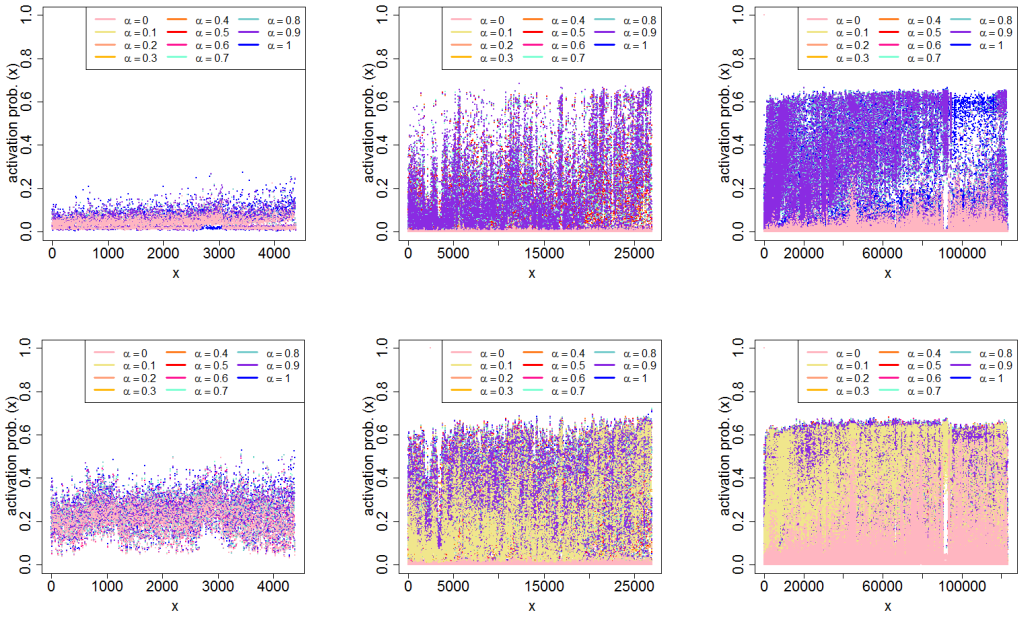
For purposes of targeted IM, both social and interaction relations can be seen as indicator of user influence. However, we note that influence is normally produced regardless of actual, visible interaction between two users. Yet, information on interaction data might be significantly sparse in real SNs, causing a flawed setting for an IM task. Without any loss of generality we have assumed that the graph G_0 (on which LurkerRank is applied) is a followship graph.

C.3 Additional results

C.3.1 Structural characteristics of seeds

In this section we report details concerning analysis of structural characteristics of the detected seeds (cf. Section 5.6.1.2)

Figure C.1 shows the *coefficient of variation* (hereinafter denoted as CV) of selected topological measures over the seed nodes, by varying α and target set size (*L-perc*). Looking at results on the outdegree, we observe decreasing trends for CV by increasing α up to 0.5, followed by roughly constant trends set around 0.4, for both DTIM methods. Consistently with the analysis on seed set overlap, L-DTIM seeds tend



(a) (b) (c)

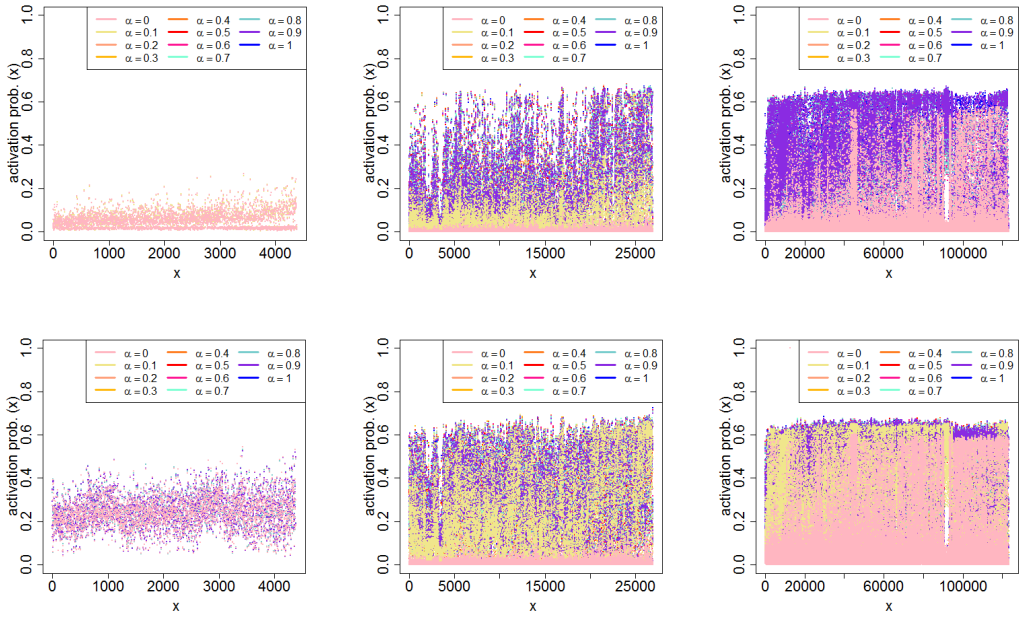
FIGURE C.2: Activation probabilities (y-axis) for each target node (x-axis), obtained by G-DTIM for varying α. Results correspond to $L\text{-}perc = 25\%$, k set to 5 (top) and 50 (bottom), on (a) `Instagram-LCC`, (b) `GooglePlus`, and (c) `FriendFeed`.

to have similar outdegree regardless of $L\text{-}perc$, while in the case of G-DTIM, relatively small variations occur for $L\text{-}perc = \{5\%, 10\%\}$ by varying α. As concerns betweenness, CV generally increases with α up to high values (0.7, 0.9), then drastically reduces to zero; this indicates that when diversity is discarded, seeds tend to correspond to source nodes in the graph. Analogously to the outdegree analysis, the trends for varying $L\text{-}perc$ are quite similar to each other in the L-DTIM case. Considering coreness, CV ranges within a much smaller interval than that corresponding to outdegree and betweenness, i.e., (0.6, 0.76) with L-DTIM, (0.64, 0.73) with G-DTIM. Again, the variability over the seeds computed by L-DTIM is much less affected by the setting of $L\text{-}perc$ than in the G-DTIM case, with a general increasing trend up to mid-high values of α.

As concerns the competing methods, KB-TIM identifies seed nodes having average CV that does not significantly change in terms of $L\text{-}perc$, specifically: (0.42, 0.40) for outdegree, (3.41, 3.52) for betweenness, and 0.61 for coreness. TIM+ identifies seed nodes that have on average 0.45 CV of outdegree, 0.0 CV of betweenness, and 0.70 CV of coreness.

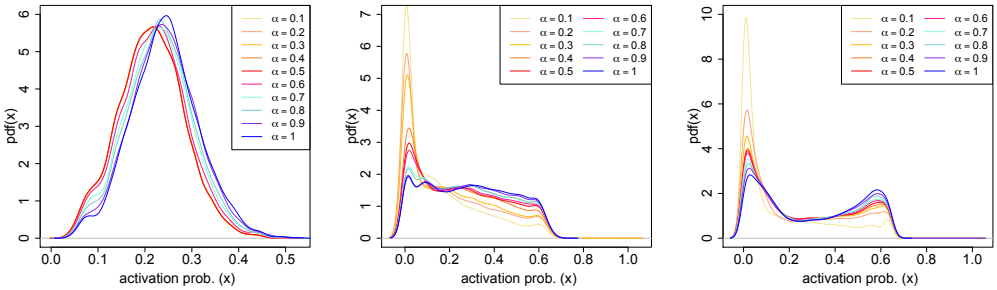
C.3.2 Target activation probabilities

In this section we report detailed results concerning the analysis of the target activation probabilities (cf. Section 5.6.2.2) with the aim of deepening our understanding of how different settings of α impact on the activation probability of nodes targeted by DTIM. We regard the activation probability of a node as the number of times the node has been activated divided by the number of Monte Carlo runs (I_{MC}, cf. Algorithm 6).



(a) (b) (c)

FIGURE C.3: Activation probabilities (y-axis) for each target node (x-axis), obtained by L-DTIM for varying α. Results correspond to *L-perc* = 25%, k set to 5 (top) and 50 (bottom), on (a) `Instagram-LCC`, (b) `GooglePlus`, and (c) `FriendFeed`.



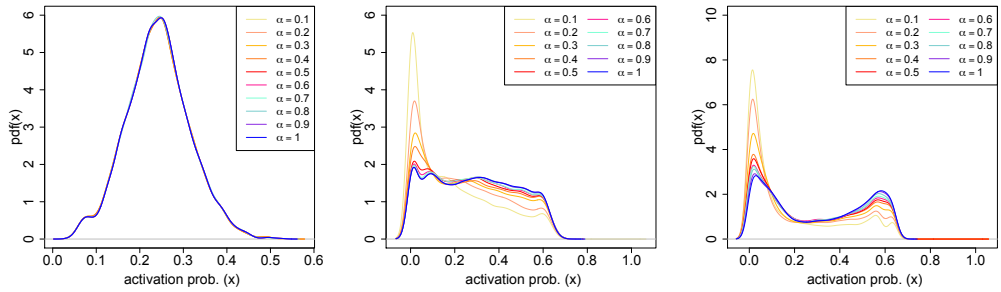
(A) (B) (C)

FIGURE C.4: Density distributions of activation probabilities obtained by G-DTIM, for varying α, with *L-perc* set to 25%, $k = 50$, on (a) `Instagram-LCC`, (b) `GooglePlus`, and (c) `FriendFeed`.

In order to analyze the above property of target nodes, we present first the activation probability values of the nodes in the final active set, shown in Figures C.2 and C.3. Next we discuss the density distributions $pdf(x)$ with variable x modeling the vector of activation probabilities associated with the nodes in the final active set, reported in Figures C.4 and C.5.

Plots of activation probability distributions. Figures C.2 and C.3 show the activation probabilities versus the target nodes, by varying the values of α and k, for G-DTIM and L-DTIM.

Considering first the performance of G-DTIM (Figure C.2), there is an evident gap between the activation probabilities obtained for low α (i.e., $\alpha \leq 0.4$), and higher



(A) (B) (C)

FIGURE C.5: Density distributions of activation probabilities obtained by L-DTIM, for varying α, with *L-perc* set to 25%, $k = 50$, on (a) `Instagram-LCC`, (b) `GooglePlus`, and (c) `FriendFeed`.

values of the parameter, with the maximum activation probability values (and maximum coverage of the target set) generally obtained for $\alpha = 0.9$ and $\alpha = 1.0$. On `Instagram-LCC` (Figure C.2(a)), given the generally low values of activation probabilities, and the high overlap among the seed sets obtained when varying α, the gap between minimum and maximum values is strongly reduced w.r.t. other datasets, with $\alpha = 1.0$ showing only small increase on the activation of targets w.r.t. $\alpha = 0.0$. Note also that for $k = 50$, there is a very small number of nodes showing activation probability within $[0.0, 0.1]$: this would hint that, when estimating the activation probabilities, the set of activated nodes remains almost unaffected in all the R Monte Carlo runs (while in other cases there are a bunch of nodes which are reached by the influence diffusion process only for a small number of runs, resulting in near-zero activation probabilities). More interesting behaviors are observed for `GooglePlus` (Figure C.2(b)). For $k = 5$ (upper plot), mid-high activation probabilities are reached for a small set of nodes starting from $\alpha = 0.5$, but the majority of target nodes is activated for $\alpha \geq 0.9$, with activation probabilities in the range $[0.0, 0.6]$. However, for $k = 50$ (lower plot), a significant set of target nodes shows mid-high activation probabilities already for $\alpha = 0.1$, indicating that, with a relatively large k, low values of α are sufficient to activate target nodes while taking into account diversity. As regards `FriendFeed` (Figure C.2(c)), activation probabilities obtained for $0.0 \leq \alpha \leq 0.6$ are generally higher than the ones obtained for the other two datasets. Nevertheless, for $k = 5$ (upper plot), a value of $\alpha = 0.7$ is needed to reach significant activation probabilities on a vast portion of the target set. Most target nodes are again reached for $\alpha = 0.9$, but it can be noted that there is a large band of target nodes (on the right side of the plot) which reaches mid-high probabilities only for $\alpha = 1.0$. This indicates that in large networks, when using low k, even small variations on the value of α can significantly impact on the effectiveness of the influence maximization process. Looking at the results obtained for $k = 50$ (lower plot), we observe that the set of target nodes obtaining a significant activation probability is relevant already for $\alpha = 0.0$, with a coverage on a large portion of the target set starting for $\alpha = 0.1$.

Quite similar qualitative remarks can be drawn about the performance of L-DTIM (Figure C.3). As regards `Instagram-LCC` (Figure C.3(a)), for $k = 5$ (upper plot) no visible improvement in the activation probabilities can be observed starting from $\alpha \geq 0.1$, while the results are similar to the ones discussed for G-DTIM for $k = 50$ (lower plot). On `GooglePlus` (Figure C.3(b)), a general improvement of the performance obtained for $\alpha = 0.1$ can be noted, while the results obtained for different α values are similar to the ones observed for G-DTIM. The improvement is more evident for $k = 5$

(upper plot), but remains significant also for $k = 50$ (lower plot). On `FriendFeed`, an increment in the activation probability values obtained for $0.0 \leq \alpha \leq 0.5$ can be noted for $k = 5$ (upper plot), w.r.t. the situation described for G-DTIM. With $k = 50$ (lower plot), higher probabilities than the ones observed for G-DTIM are observed for $\alpha = 0.0$.

Density distributions of activation probability. Figures C.4 and C.5 show density distributions of activation probability obtained for G-DTIM and L-DTIM, respectively.

Focusing first on `GooglePlus`, similar trends can be noted for both G-DTIM (Figure C.4(b)) and L-DTIM (Figure C.5(b)). A density peak corresponding to low activation probability values (close to 0.0) can be noted for low values of α (i.e., $\alpha \leq 0.6$ for G-DTIM and $\alpha \leq 0.4$ for L-DTIM). This peak slightly decreases for increasing values of α, yielding a relatively wide area of nearly constant density (e.g., around 2) which covers a range of activation probabilities from 0.0 up to about 0.6.

A roughly bi-modal distribution can be observed for `FriendFeed`, for both G-DTIM (Figure C.4(c)) and L-DTIM (Figure C.5(c)). It is easy to recognize a first peak corresponding to near-zero activation probability values, and a second one located around 0.6; hence, the first peak becomes lower and the second peak higher by increasing α.

Analogously to previous evaluation settings, situation on `Instagram-LCC` is drastically different from the other two datasets, which in this case corresponds to roughly Normal distributions for varying α. Using G-DTIM (Figure C.4(a)), the density distribution has a mean activation probability which spans from approximately 0.2 for low values of α to values close to 0.3 for higher values of α. Using L-DTIM (Figure C.4(b)), due to the high overlap of the seed sets obtained when varying α, all distributions are nearly identical, and centered on an average value of activation probability around 0.25.

It should be noted that the density distributions referring to the setting $\alpha = 0.0$ are omitted from Figures C.4 and C.5. The reason behind this choice is that, as discussed in the previous analysis, in some cases there is a large gap between the activation probabilities obtained with $\alpha = 0.0$ and $\alpha = 0.1$. Here the entity of such a gap causes the curve of density distribution for $\alpha = 0.0$ to have a peak corresponding to very high values of probability density function for near-zero values of activation probability (which, if showed, would force us to use a larger scale, making the other curves difficult to read). This contingency is observed on `GooglePlus` for both versions of DTIM, and `FriendFeed` for G-DTIM, while in other cases the density curve for $\alpha = 0.0$ can be relatively close (`FriendFeed` with L-DTIM) or nearly identical (`Instagram-LCC` for G-DTIM and L-DTIM) to the curve shown for $\alpha = 0.1$.

C.3.3 Correlation analysis between capital and diversity measurements

Tables C.1 and C.2 summarize results of correlation analysis between the sequence of capital values and the sequence of diversity values associated to the nodes at convergence of the diffusion process, for each of the DTIM methods and for selected settings of parameters.

TABLE C.1: Correlation analysis between capital and diversity measurements: G-DTIM

network	α	*L-perc* (%)	k	correlation
GooglePlus	0.1	10	5	-0.001
GooglePlus	0.5	10	5	-0.004
GooglePlus	0.9	10	5	-0.005
GooglePlus	0.1	25	5	0.006
GooglePlus	0.5	25	5	-0.001
GooglePlus	0.9	25	5	-0.006
FriendFeed	0.1	10	5	-4.4e-05
FriendFeed	0.5	10	5	-7.8e-05
FriendFeed	0.9	10	5	-8.1e-05
FriendFeed	0.1	25	5	0.004
FriendFeed	0.5	25	5	0.003
FriendFeed	0.9	25	5	0.001
GooglePlus	0.1	10	50	-0.008
GooglePlus	0.5	10	50	-0.008
GooglePlus	0.9	10	50	-0.007
GooglePlus	0.1	25	50	-0.008
GooglePlus	0.5	25	50	-0.006
GooglePlus	0.9	25	50	-0.011
FriendFeed	0.1	10	50	-1.6e-04
FriendFeed	0.5	10	50	-2.3e-04
FriendFeed	0.9	10	50	-2.7e-04
FriendFeed	0.1	25	50	5.5e-04
FriendFeed	0.5	25	50	3.0e-04
FriendFeed	0.9	25	50	3.3e-04

TABLE C.2: Correlation analysis between capital and diversity measurements: L-DTIM

network	α	*L-perc* (%)	k	correlation
GooglePlus	0.1	10	5	0.169
GooglePlus	0.5	10	5	0.059
GooglePlus	0.9	10	5	0.008
GooglePlus	0.1	25	5	0.148
GooglePlus	0.5	25	5	0.054
GooglePlus	0.9	25	5	0.004
FriendFeed	0.1	10	5	0.085
FriendFeed	0.5	10	5	0.046
FriendFeed	0.9	10	5	0.018
FriendFeed	0.1	25	5	0.076
FriendFeed	0.5	25	5	0.052
FriendFeed	0.9	25	5	0.020
GooglePlus	0.1	10	50	0.225
GooglePlus	0.5	10	50	0.088
GooglePlus	0.9	10	50	0.025
GooglePlus	0.1	25	50	0.229
GooglePlus	0.5	25	50	0.097
GooglePlus	0.9	25	50	0.020
FriendFeed	0.1	10	50	0.164
FriendFeed	0.5	10	50	0.126
FriendFeed	0.9	10	50	0.069
FriendFeed	0.1	25	50	0.180
FriendFeed	0.5	25	50	0.131
FriendFeed	0.9	25	50	0.064

App. D

Attribute-based Diversity-sensitive Targeted Influence Maximization

D.1 Example calculation of diversity functions

We provide a numerical example of application of the proposed diversity functions. Let us consider the following simple categorical dataset, with five tuples and four attributes:

	A_1	A_2	A_3	A_4
v_1	$\bullet$	$\square$	$\oplus$	\$
v_2	$\bullet$	$\triangledown$	$\odot$	\$
v_3	$\bullet$	$\square$	$\ominus$	€
v_4	$\star$	$\square$	$\oplus$	£
v_5	$*$	$\triangle$	$\oplus$	£

We want to compute the marginal gain of each node w.r.t. the set $S = \{v_1, v_2\}$, according to each diversity function.

Attribute-wise diversity. For the sake of simplicity, we assume each attribute is equally important, i.e., the coefficients ω_i are the same for all A_i with $i = 1..4$, therefore they will be ignored for the purpose of this example. By setting $\lambda = 1$, the attribute-wise diversity of S is $div^{(AW)}(S) = 7$. To compute the marginal gain, Fact 1 applies. For instance, the marginal gain of adding v_3 to S is given by $(n_\bullet + 1)^{-\lambda} + (n_\square + 1)^{-\lambda} + (n_\ominus + 1)^{-\lambda} + (n_€ + 1)^{-\lambda} = 1/3 + 1/2 + 1 + 1 = 2.83$. Analogously, the marginal gain of adding v_4 to S is 3.0, and the marginal gain of adding v_5 to S is 3.5.

Hamming-based diversity. Assuming a radius $\xi = 2$, and that the influence range of each node v_i ($i = 1..4$) in the graph contains all the other nodes, the Hamming-balls associated with each tuple are the following: $B^\xi_{v_1} = \{v_2, v_3, v_4\}$, $B^\xi_{v_2} = \{v_1\}$, $B^\xi_{v_3} = \{\}$, $B^\xi_{v_4} = \{v_1, v_5\}$ and $B^\xi_{v_5} = \{v_4\}$. Given the above Hamming-balls, the Hamming-based diversity of S is $div^{(HB)}(S) = |B^\xi_{v_1} \cup B^\xi_{v_2}| = 4$. Based on Fact 2, the marginal gain of adding v_3 to S is given by $|B^\xi_{v_3} \setminus B^\xi_S|$, which is equal to 0. Analogously, the marginal gain of adding v_4 to S is 1, and the marginal gain of adding v_5 to S is 0.

Entropy-based diversity. Each tuple is associated with a random variable; for instance, v_1 is associated with the following variable denoted as X_{v_1}:

$$X_{v_1} = \langle \overbrace{\underbrace{1}_{\bullet}, \underbrace{0}_{\star}, \underbrace{0}_{*}}^{A_1}, \overbrace{\underbrace{0}_{\triangledown}, \underbrace{0}_{\triangle}, \underbrace{1}_{\square}}^{A_2}, \overbrace{\underbrace{0}_{\odot}, \underbrace{1}_{\oplus}, \underbrace{0}_{\ominus}}^{A_3}, \overbrace{\underbrace{1}_{\$}, \underbrace{0}_{€}, \underbrace{0}_{£}}^{A_4} \rangle$$

Moreover, a probability distribution is defined over the attribute symbols, where each symbol is associated with its relative frequency through the entire dataset; for instance, we have the probabilities $p(\bullet) = 3/20$ and $p(\star) = p(*) = 1/20$.

In order to measure the diversity of S, we need to compute the joint probability distribution of variables X_{v_1} and X_{v_2}. Let us consider the following table, whose first row corresponds to the attribute probability distribution and the subsequent two rows correspond to the variables associated with the tuples in S, i.e., v_1 and v_2.

	$\bullet$	$\star$	$*$	$\triangledown$	$\triangle$	$\square$	$\odot$	$\oplus$	$\ominus$	\$	€	£
p	3/20	1/20	1/20	1/20	1/20	3/20	1/20	3/20	1/20	2/20	1/20	2/20
X_{v_1}	1	0	0	0	0	1	0	1	0	1	0	0
X_{v_2}	1	0	0	1	0	0	1	0	0	1	0	0

We can present the joint probability distribution of X_{v_1} and X_{v_2} as the following table:

X_{v_2} \ X_{v_1}	0	1	$P(X_{v_2})$
0	$7/20 = 0.35$	$6/20 = 0.3$	0.65
1	$2/20 = 0.1$	$5/20 = 0.25$	0.35
$P(X_{v_1})$	0.45	0.55	

Moreover, the outer bottom row and the outer rightmost column correspond to the marginal probability distribution for X_{v_1} and the marginal probability distribution for X_{v_2}, respectively.

The entropy-based diversity of S, assuming without loss of generality that v_2 was added after v_1, is as follows: $div^{(E)}(S) = H(X_{v_1}, X_{v_2}) = H(X_{v_1}) + H(X_{v_2}|X_{v_1}) = H([9/20, 11/20]) + (9/20)\, H([7/9, 2/9]) + (11/20)\, H([6/11, 5/11]) = 0.99 + 0.89 = 1.88$.

In order to compute the marginal gain of adding v_3 to S, following Fact3, we first need to derive its conditional distribution $P(X_{v_3}|X_{v_1}, X_{v_2})$, following the same procedure as before. Then, the marginal gain is given by the conditional entropy $H(X_{v_3}|X_{v_1}, X_{v_2})$, which is equal to 0.84. Analogously, the marginal gain of adding v_4 to S is 0.34, and the marginal gain of adding v_5 to S is 0.64.

Class-based diversity. Suppose to partition the dataset according to the first attribute of the schema, A_1, i.e., the symbols in A_1 are regarded as class labels. Also, let us set reward $r = 1$ for all tuples and the aggregation function $f(x) = \log(1 + x)$. The class-based diversity of the set S is $div^{(C)}(S) = 1.09$. Based upon Fact 4, the marginal gain of adding v_3 is given by $\log(1 + r/R_l) = 0.28$, where $C_l = \bullet$ (i.e., the class of v_3) and $R_\bullet = 3$. Analogously, the marginal gain of adding v_4 or v_5 is 0.69, as their respective classes are not covered by the tuples in S.

D.2 Inappropriate set-diversity functions

We report details about a number of functions that, despite their simplicity, were demonstrated to be unsuitable as diversity functions for our problem (cf. Section 6.4.1).

Concerning attribute-wise functions, we discussed that a simple approach would be to aggregate *pairwise distances* of the node profiles w.r.t. a given attribute A. We consider in particular the following definition based on pairwise attribute-value mismatchings:

$$f_1(S, A) = \frac{1}{|S|} \sum_{u,v \in S} \mathbb{1}[val_A(u) \neq val_A(v)],$$

where $\mathbb{1}[\cdot]$ denotes the indicator function.[1] It is easy to prove that this function is non-submodular; to give empirical evidence of this fact, consider the following example. We are given $S = \{u, v, x\}$ with $val_A(u) = val_A(v) = a_1$ and $val_A(x) = a_2$, and $T = \{u, v, x, y\}$ with $val_A(y) = a_1$. Suppose that node z, with $val_A(z) = a_2$, is inserted into S and T, then it holds that: $f_1(S, A) = \frac{2}{3}$, $f_1(S \cup \{z\}, A) = \frac{4}{4}$, $f_1(T, A) = \frac{3}{4}$, and $f_1(T \cup \{z\}, A) = \frac{6}{5}$. It follows that $f_1(S \cup \{z\}, A) - f_1(S, A) \not\geq f_1(T \cup \{z\}, A) - f_1(T, A)$. Note also that the property of submodularity still does not hold if the normalization term (i.e., $|S|$) is discarded in $f_1(\cdot)$.

Let us now extend to computing pairwise distances of the node profiles in their entirety, focusing on the *Hamming distance*, as defined in Equation (6.6). Upon this, let us define $f_2(S) = \sum_{u,v \in S, u \neq v} dist^H(u, v)$, and two normalized versions: $\widehat{f}_2(S) = (1/(2|S|))f_2(S)$ and $\widehat{\widehat{f}}_2(S) = (1/|S|(|S|-1))f_2(S)$. It is easy to check that none of such functions is appropriate. Let us consider the following example. We are given a schema with three attributes ($m = 3$) and sets $S = \{u, v\}$, such that $\mathcal{A}[u] = \langle a_1, \perp, \perp \rangle$, $\mathcal{A}[v] = \langle a_2, \perp, \perp \rangle$, and $T = \{u, v, x\}$, such that $\mathcal{A}[x] = \langle a_3, b_1, c_1 \rangle$. Suppose that node z, with $\mathcal{A}[z] = \langle a_4, \perp, \perp \rangle$, is inserted into S and T, then it holds that: $f_2(S) = 2$, $f_2(T) = 14$, $f_1(S \cup \{z\}) = 6$, and $f_1(T \cup \{z\}) = 24$. It follows that $f_2(S \cup \{z\}) - f_2(S) \not\geq f_2(T \cup \{z\}) - f_2(T)$. Considering $\widehat{f}_2(\cdot)$, we have: $\widehat{f}_2(S) = \frac{1}{2}$, $\widehat{f}_2(T) = \frac{7}{3}$, $\widehat{f}_2(S \cup \{z\}) = 1$, and $\widehat{f}_2(T \cup \{z\}) = 3$; thus, again $\widehat{f}_2(S \cup \{z\}) - \widehat{f}_2(S) \not\geq \widehat{f}_2(T \cup \{z\}) - \widehat{f}_2(T)$. Yet, when using $\widehat{\widehat{f}}_2(\cdot)$, we have: $\widehat{\widehat{f}}_2(S) = 1$, $\widehat{\widehat{f}}_2(T) = \frac{7}{3}$, $\widehat{\widehat{f}}_2(S \cup \{z\}) = 1$, and $\widehat{\widehat{f}}_2(T \cup \{z\}) = 2$; in this case, mononicity is not even satisfied (since $\widehat{\widehat{f}}_2(T \cup \{z\}) \not\geq \widehat{\widehat{f}}_2(T)$).

Alternatively, we considered *Jaccard distance*, i.e., given the profiles of any two nodes u, v:

$$dist^J(u, v) = 1 - \frac{\sum_{j=1}^{m} \mathbb{1}[val_{A_j}(u) = val_{A_j}(v)]}{|\mathcal{A}[u]| + |\mathcal{A}[v]| - \sum_{j=1}^{m} \mathbb{1}[val_{A_j}(u) = val_{A_j}(v)]}.$$

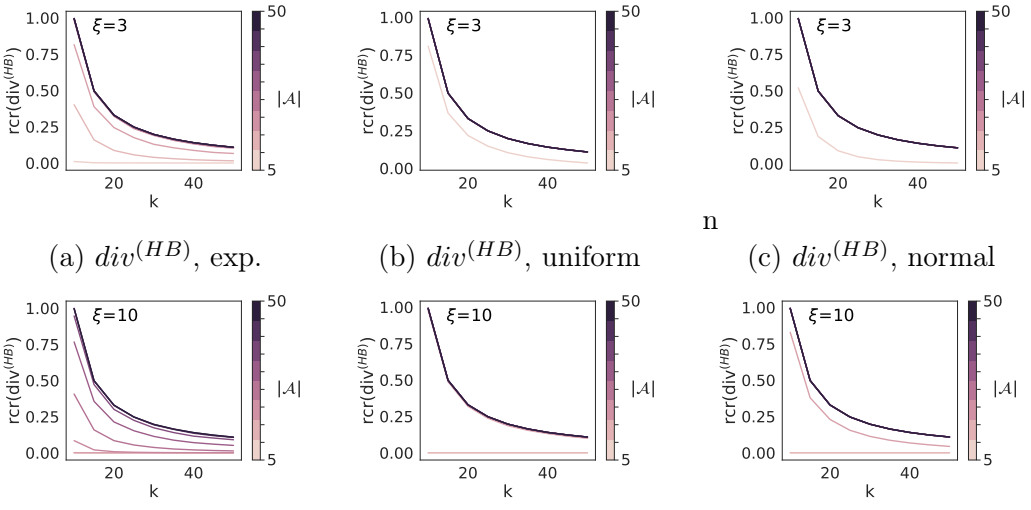
Upon this, let us define $f_3(S) = \sum_{u,v \in S, u \neq v} dist^J(u, v)$, and normalized version: $\widehat{f}_3(S) = (1/(2|S|))f_2(S)$. Like previous functions, it can be empirically shown that $f_3(\cdot)$ and $\widehat{f}_3(\cdot)$ are not appropriate for our purposes. Suppose we are given a schema with five attributes ($m = 5$) and sets $S = \{u, v\}$, such that $\mathcal{A}[u] = \langle a, b, c, \perp, \perp \rangle$, $\mathcal{A}[v] = \langle a, b, \perp, d, \perp \rangle$, and $T = \{u, v, x\}$, such that $\mathcal{A}[x] = \mathcal{A}[v]$. Suppose that node z, with $\mathcal{A}[z] = \langle a, \perp, \perp, d, e \rangle$, is inserted into S and T, then it holds that: $f_3(S) = 1, f_3(T) = 2, f_3(S \cup \{z\}) = \frac{18}{5}, f_3(T \cup \{z\}) = \frac{28}{5}$. It follows that $f_3(S \cup \{z\}) - f_3(S) \not\geq f_3(T \cup \{z\}) - f_3(T)$. Considering $\widehat{f}_3(\cdot)$, we have: $\widehat{f}_3(S) = \frac{1}{4}$, $\widehat{f}_2(T) = \frac{1}{3}$, $\widehat{f}_2(S \cup \{z\}) = \frac{3}{5}$, and $\widehat{f}_2(T \cup \{z\}) = \frac{7}{10}$; thus, again $\widehat{f}_3(S \cup \{z\}) - \widehat{f}_3(S) \not\geq \widehat{f}_3(T \cup \{z\}) - \widehat{f}_3(T)$.

The above Jaccard distance function could also be exploited to allow for measuring the dissimilarity of all profiles in any set $S = \{v_1, \ldots, v_k\} \subseteq V$:

$$f_4(S) = 1 - \frac{\sum_{j=1}^{m} \mathbb{1}[val_{A_j}(v_1) = \ldots = val_{A_j}(v_k)]}{\sum_{j=1}^{m} |\bigcup_{v \in S} \{val_{A_j}(v)\}|}.$$

However, it is straightforward to show that the above function can easily yield useless results; e.g., referring to the previous example, the marginal gains of z w.r.t. S and T are the same. Even worse, a normalization of $f_4(S)$ by set-size does not even ensure monotonicity.

[1] For any nodes u and v, we assume that if either u's or v's profile is not associated with a value in the domain of A (i.e., missing value for A), then the indicator function will be evaluated as 1.



(a) $div^{(HB)}$, exp. (b) $div^{(HB)}$, uniform (c) $div^{(HB)}$, normal

(d) $div^{(HB)}$, exp. (e) $div^{(HB)}$, uniform (f) $div^{(HB)}$, normal

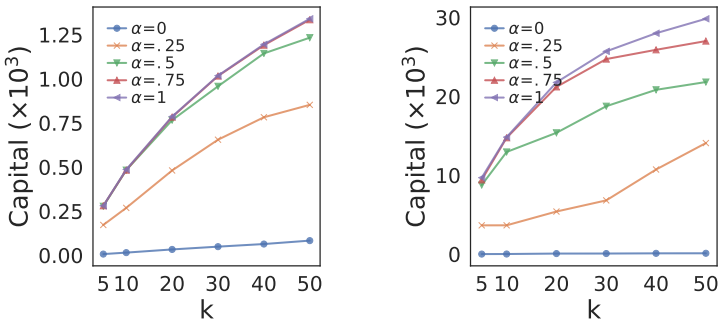
FIGURE D.1: Relative change rate of the Hamming-based diversity function with radius $\xi = 3$ (top) and $\xi = 10$ (bottom) by varying the number of attributes ($|\mathcal{A}|$), on different categorical datasets. Different colors correspond to different projections of the dataset: the darker the color, the higher the number i of attributes selected from the schema, where $i \in [5..50]$ with increments of 5. The number of per-attribute admissible values is set to 15.

D.3 Additional experimental results

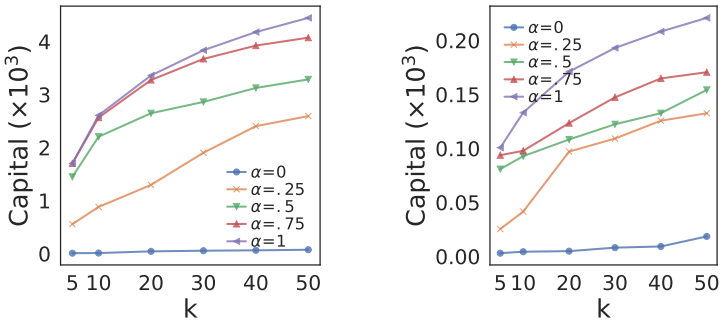
Figure D.1 shows additional results on the relative change rate of the Hamming-based diversity function (cf. Section 6.7.1.1). As supplementary material for Section 6.7.2, Figure D.2 shows additional results on the relation between capital and diversity functions in ADITUM, while Figures D.3-D.4 and Figures D.6-D.5 show results on normalized overlap of seed sets and results on a comparison between exponential and uniform distributions, respectively, for top-5% and top-10% target selection thresholds.

D.4 Effect of the attribute distribution

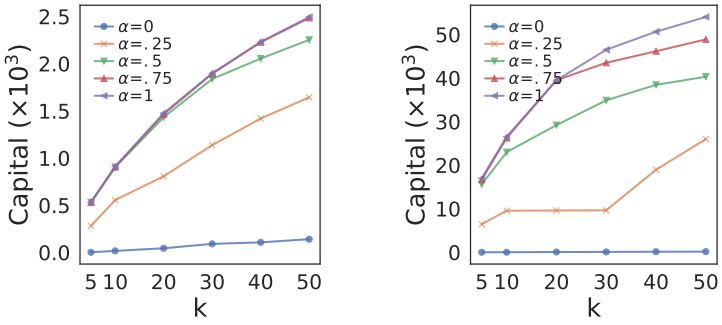
As supplementary material for the analysis on effects due to the attribute distribution discussed in Section 6.7.2, Figures D.6-D.5 shows further results on comparison between exponential and uniform distributions, for top-5% and top-10% target selection threshold.



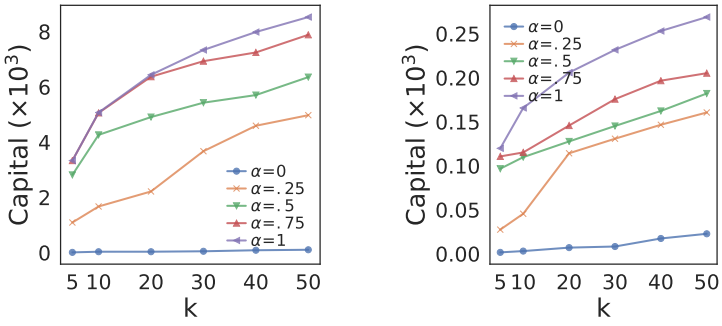
(a) Instagram (b) FriendFeed



(c) GooglePlus (d) Reddit

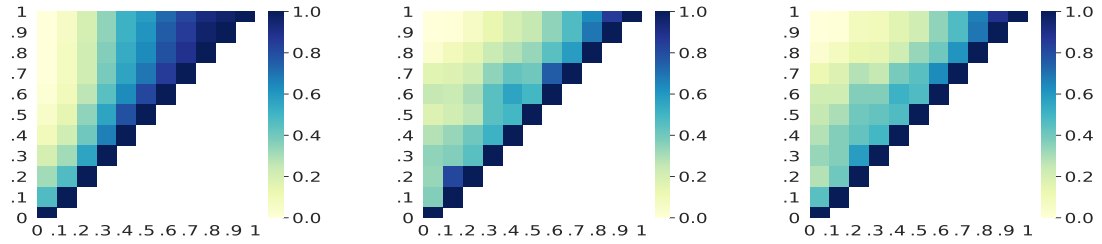


(e) Instagram (f) FriendFeed

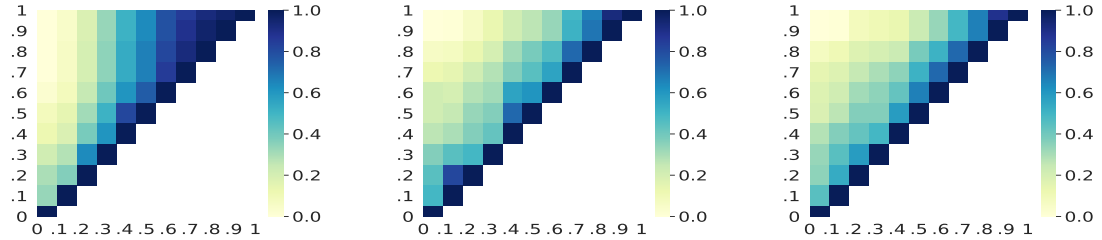


(g) GooglePlus (h) Reddit

FIGURE D.2: Expected capital, by varying $\alpha \in \{0, 0.25, 0.5, 1\}$, with $k \in [5, 50]$, top-5% (a–d) and top-10% (e–h) target selection, and exponential distribution of attributes (except Reddit).

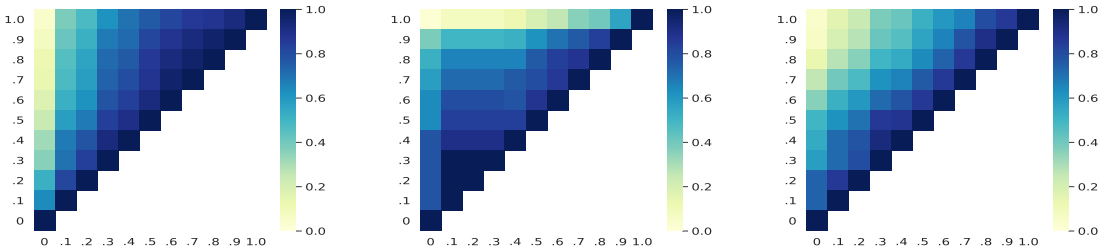


(a) Instagram (b) FriendFeed (c) GooglePlus

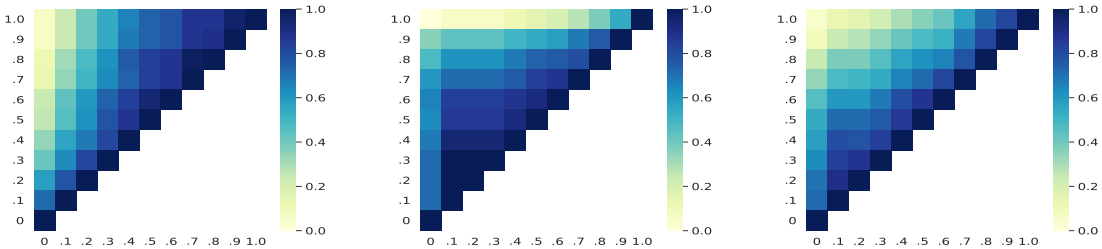


(d) Instagram (e) FriendFeed (f) GooglePlus

FIGURE D.3: Normalized overlap of seed sets, for $\alpha \in [0, 1]$ (with increments of 0.1), $k = 50$, top-5% (top) and top-10% (bottom) target selection, and exponential distribution of attributes.



(a) Instagram (b) FriendFeed (c) GooglePlus



(d) Instagram (e) FriendFeed (f) GooglePlus

FIGURE D.4: Normalized overlap of seed sets, for $\alpha \in [0, 1]$ (with increments of 0.1), $k = 50$, top-5% (top) and top-10% (bottom) target selection, and exponential distribution of attributes.

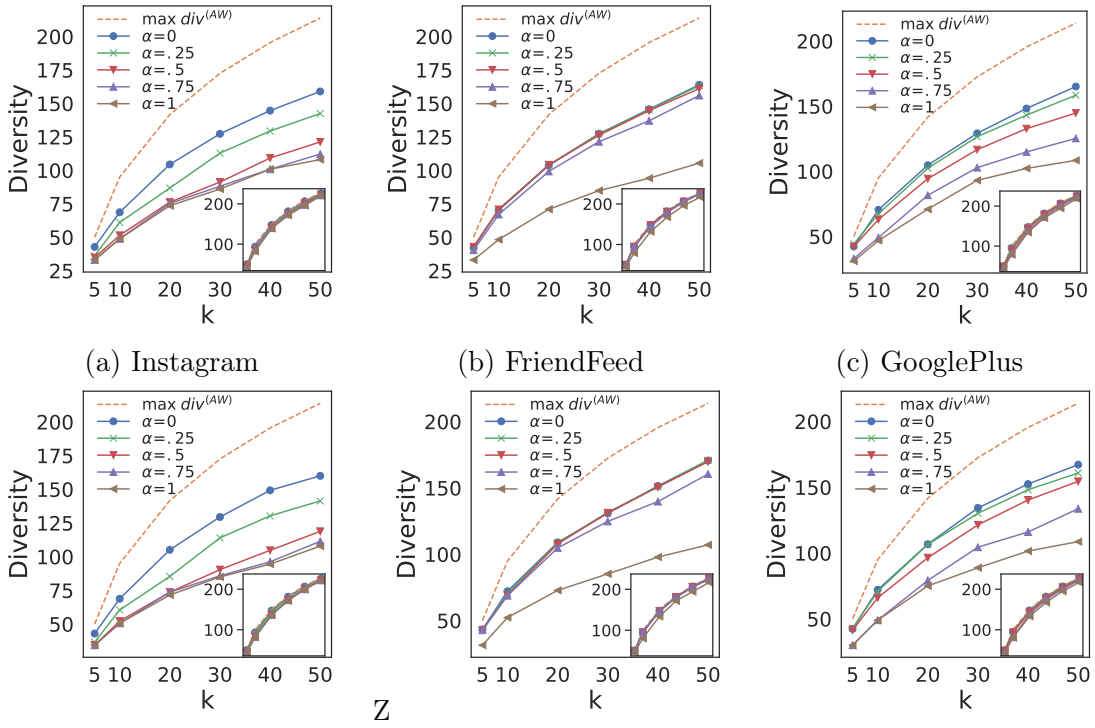


FIGURE D.5: Exponential (main) vs. uniform (inset) distribution: seed-set diversity for varying k and α, top-5% (a–c) and top-10% (d–f) target selection, and comparison to maximum diversity value.

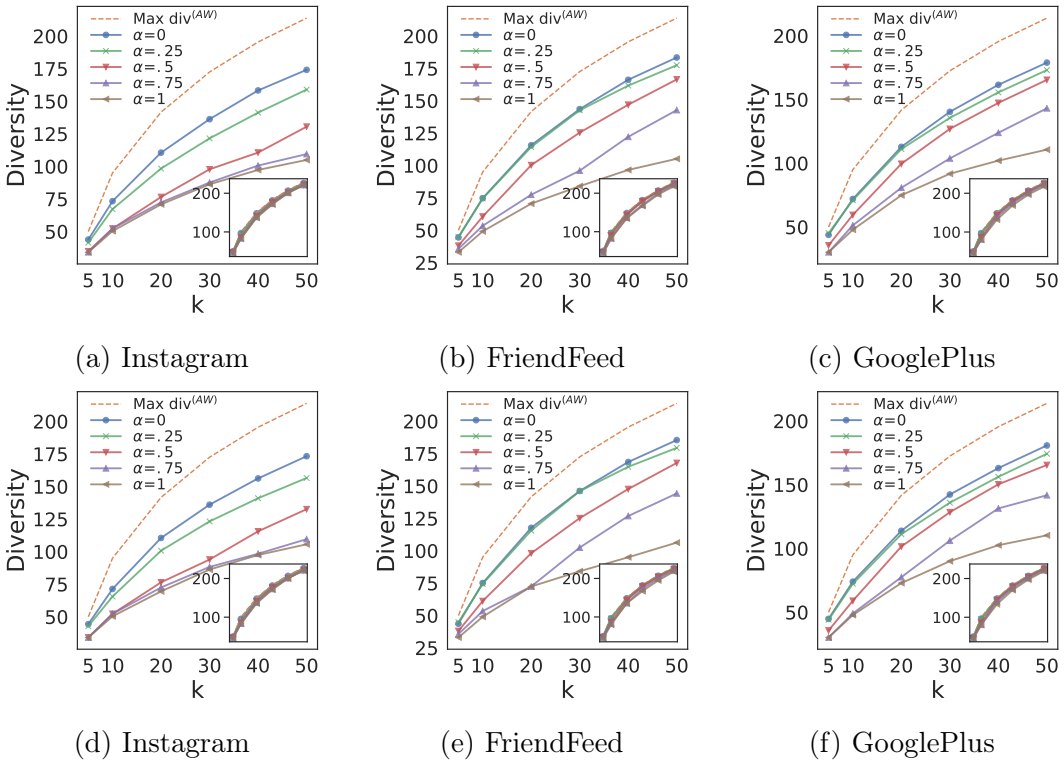


FIGURE D.6: Exponential (main) vs. uniform (inset) distribution: seed-set diversity for varying k and α, top-5% (a–c) and top-10% (d–f) target selection, and comparison to maximum diversity value. —-

Bibliography

[1] R. Alhajj and J. Rokne. *Boundary spanning*. 2014, p. 82.

[2] A. Anagnostopoulos et al. "Viral Misinformation: The Role of Homophily and Polarization". In: *Proc. World Wide Web Conf.* 2015, pp. 355–356.

[3] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. "Influence and Correlation in Social Networks". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: Association for Computing Machinery, 2008, 7–15. ISBN: 9781605581934. DOI: 10.1145/1401890.1401897. URL: https://doi.org/10.1145/1401890.1401897.

[4] Ashton Anderson et al. "Steering User Behavior with Badges". In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13. Rio de Janeiro, Brazil: Association for Computing Machinery, 2013, 95–106. ISBN: 9781450320351. DOI: 10.1145/2488388.2488398. URL: https://doi.org/10.1145/2488388.2488398.

[5] Akhil Arora, Sainyam Galhotra, and Sayan Ranu. "Debunking the Myths of Influence Maximization: An In-Depth Benchmarking Study". In: *Proceedings of the 2017 ACM International Conference on Management of Data*. SIGMOD '17. Chicago, Illinois, USA: Association for Computing Machinery, 2017, 651–666. ISBN: 9781450341974. DOI: 10.1145/3035918.3035924. URL: https://doi.org/10.1145/3035918.3035924.

[6] Çigdem Aslay et al. "Maximizing the Diversity of Exposure in a Social Network". In: *Proc. IEEE Int. Conf. on Data Mining (ICDM)*. 2018, pp. 863–868.

[7] J. Bae and S. Kim. "Identifying and ranking influential spreaders in complex networks by neighborhood coreness". In: *Physica A: Statistical Mechanics and its Applications* 395 (2014), pp. 549–559.

[8] Eytan Bakshy et al. "Everyone's an Influencer: Quantifying Influence on Twitter". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: Association for Computing Machinery, 2011, 65–74. ISBN: 9781450304931. DOI: 10.1145/1935826.1935845. URL: https://doi.org/10.1145/1935826.1935845.

[9] S. Banerjee, M. Jenamani, and D. K. Pratihar. "ComBIM: A community-based solution approach for the Budgeted Influence Maximization Problem". In: *Expert Systems with Applications* 125 (2019), pp. 1–13.

[10] Q. Bao, W. K. Cheung, and Y. Zhang. "Incorporating Structural Diversity of Neighbors in a Diffusion Model for Social Networks". In: *Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence*. 2013, pp. 431–438.

[11] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. "Cascade-Based Community Detection". In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM '13. Rome, Italy: Association for Computing Machinery, 2013, 33–42. ISBN: 9781450318693. DOI: 10.1145/2433396.2433403. URL: https://doi.org/10.1145/2433396.2433403.

[12] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. "Topic-Aware Social Influence Propagation Models". In: *Proc. IEEE Int. Conf. on Data Mining (ICDM)*. 2012, pp. 81–90.

[13] V. Batagelj and M. Zaversnik. "An O(m) Algorithm for Cores Decomposition of Networks". In: *CoRR* cs.DS/0310049 (2003). URL: http://arxiv.org/abs/cs.DS/0310049.

[14] A. Bessi et al. "Social Determinants of Content Selection in the Age of (Mis)Information". In: *Proc. Int. Conf. on Social Informatics (SocInfo)*. 2014, pp. 259–268.

[15] Jonathan Bishop. "Increasing participation in online communities: A framework for human-computer interaction". In: *Computers in Human Behavior* 23.4 (2007), pp. 1881–1893.

[16] F. Bonchi. "Influence Propagation in Social Networks: A Data Mining Perspective". In: *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol. 1. 2011, pp. 2–2. DOI: 10.1109/WI-IAT.2011.286.

[17] F. Bonchi, A. Khan, and L. Severini. "Distance-generalized Core Decomposition". In: *Proc. ACM SIGMOD*. 2019, pp. 1006–1023.

[18] F. Bonchi et al. "Core Decomposition of Uncertain Graphs". In: *Proc. ACM KDD*. 2014, pp. 1316–1325.

[19] Francesco Bonchi et al. "Probabilistic Causal Analysis of Social Influence". In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM '18. Torino, Italy: Association for Computing Machinery, 2018, 1003–1012. ISBN: 9781450360142. DOI: 10.1145/3269206.3271756. URL: https://doi.org/10.1145/3269206.3271756.

[20] C. Borgs et al. "Maximizing Social Influence in Nearly Optimal Time". In: *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*. 2014, pp. 946–957.

[21] Arastoo Bozorgi et al. "INCIM: A community-based algorithm for influence maximization problem under the linear threshold model". In: *Inf. Process. Manage.* 52.6 (2016), pp. 1188–1199.

[22] U. Brandes et al. "Network Analysis of Collaboration Structure in Wikipedia". In: *Proc. World Wide Web Conf.* 2009, pp. 731–740.

[23] Linda Briesemeister, Patrick Lincoln, and Phillip Porras. "Epidemic Profiles and Defense of Scale-Free Networks". In: *Proceedings of the 2003 ACM Workshop on Rapid Malcode*. WORM '03. Washington, DC, USA: Association for Computing Machinery, 2003, 67–75. ISBN: 1581137850. DOI: 10.1145/948187.948200. URL: https://doi.org/10.1145/948187.948200.

[24] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. "Limiting the Spread of Misinformation in Social Networks". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. Hyderabad, India: Association for Computing Machinery, 2011, 665–674. ISBN: 9781450306324. DOI: 10.1145/1963405.1963499. URL: https://doi.org/10.1145/1963405.1963499.

[25] A. Caliò, A. Tagarelli, and F. Bonchi. "Cores matter? An analysis of graph decomposition effects on influence maximization problems". In: *Proc. 12th ACM Conf. on Web Science (WebSci)*. 2020. DOI: 10.1145/3394231.3397908.

[26] A. Caliò et al. "Topology-driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks". In: *IEEE Trans. Knowl. Data Eng.* 30.12 (2018), pp. 2421–2434.

[27] Tianyu Cao et al. "OASNET: an optimal allocation approach to influence maximization in modular social networks". In: *Proc. ACM Symposium on Applied Computing (SAC)*. 2010, pp. 1088–1094.

[28] Fabio Celli et al. "Social Network Data and Practices: The Case of Friendfeed". In: vol. 6007. Mar. 2010, pp. 346–353. ISBN: 978-3-642-12078-7. DOI: 10.1007/978-3-642-12079-4_43.

[29] Damon Centola and Michael Macy. "Complex Contagions and the Weakness of Long Ties". In: *American Journal of Sociology* 113.3 (2007), pp. 702–734. DOI: 10.1086/521848. URL: https://doi.org/10.1086/521848.

[30] Deepayan Chakrabarti et al. "Epidemic Thresholds in Real Networks". In: *ACM Trans. Inf. Syst. Secur.* 10.4 (Jan. 2008). ISSN: 1094-9224. DOI: 10.1145/1284680.1284681. URL: https://doi.org/10.1145/1284680.1284681.

[31] S. Chen and K. He. "Influence Maximization on Signed Social Networks with Integrated PageRank". In: *Proc. IEEE Social Computing Conf.* 2015, pp. 289–292.

[32] Shuo Chen et al. "Online Topic-Aware Influence Maximization". In: *PVLDB* 8.6 (2015), pp. 666–677.

[33] W. Chen, L. V. S. Lakshmanan, and C. Castillo. *Information and Influence Propagation in Social Networks*. Morgan & Claypool, 2013.

[34] W. Chen, W. Lu, and N. Zhang. "Time-critical influence maximization in social networks with time-delayed diffusion". In: *Proc. AAAI Conf.* 2012, pp. 592–598.

[35] W. Chen, Y. Yuan, and L. Zhang. "Scalable Influence Maximization in Social Networks under the Linear Threshold Model". In: *2010 IEEE International Conference on Data Mining*. 2010, pp. 88–97. DOI: 10.1109/ICDM.2010.118.

[36] W. Chen et al. "Influence maximization in social networks when negative opinions may emerge and propagate". In: *Proc. SIAM Conf. on Data Mining*. 2011.

[37] Wei Chen, Chi Wang, and Yajun Wang. "Scalable influence maximization for prevalent viral marketing in large-scale social networks". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 1029–1038.

[38] Wei Chen, Chi Wang, and Yajun Wang. "Scalable Influence Maximization for Prevalent Viral Marketing in Large-Scale Social Networks". In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '10. Washington, DC, USA: Association for Computing Machinery, 2010, 1029–1038. ISBN: 9781450300551. DOI: 10.1145/1835804.1835934. URL: https://doi.org/10.1145/1835804.1835934.

[39] Wei Chen, Chi Wang, and Yajun Wang. "Scalable influence maximization for prevalent viral marketing in large-scale social networks". In: *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 2010, pp. 1029–1038.

[40] Wei Chen, Yajun Wang, and Siyu Yang. "Efficient Influence Maximization in Social Networks". In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '09. Paris, France: Association for Computing Machinery, 2009, 199–208. ISBN: 9781605584959. DOI: 10.1145/1557019.1557047. URL: https://doi.org/10.1145/1557019.1557047.

[41] Wei Chen, Yifei Yuan, and Li Zhang. "Scalable Influence Maximization in Social Networks under the Linear Threshold Model". In: *Proc. IEEE Int. Conf. on Data Mining (ICDM)*. 2010, pp. 88–97.

[42] Yi-Cheng Chen et al. "CIM: Community-Based Influence Maximization in Social Networks". In: *ACM TIST* 5.2 (2014), 25:1–25:31.

[43] Suqi Cheng et al. "StaticGreedy: Solving the Scalability-Accuracy Dilemma in Influence Maximization". In: *Proceedings of the 22nd ACM International Conference on Information amp; Knowledge Management*. CIKM '13. San Francisco, California, USA: Association for Computing Machinery, 2013, 509–518. ISBN: 9781450322638. DOI: 10.1145/2505515.2505541. URL: https://doi.org/10.1145/2505515.2505541.

[44] N.A. Christakis and J.H. Fowler. *Connected: The Surprising Power of our Social Networks and How They Shape Our Lives*. Little, Brown, 2009.

[45] Nicholas A. Christakis and James H. Fowler. "The Spread of Obesity in a Large Social Network over 32 Years". In: *N Engl J Med* 357.4 (2007), pp. 370–379. ISSN: 0028-4793. DOI: 10.1056/NEJMsa066082.

[46] Ben-Avraham Cohen R Havlin S. "D. Efficient immunization strategies for computer networks and populations." In: (2003), 91(24). DOI: 10.1103/PhysRevLett.91.247901.

[47] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[48] David Crandall et al. "Feedback Effects between Similarity and Social Influence in Online Communities". In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA: Association for Computing Machinery, 2008, 160–168. ISBN: 9781605581934. DOI: 10.1145/1401890.1401914. URL: https://doi.org/10.1145/1401890.1401914.

[49] A. Das et al. "Information dissemination in heterogeneous-intent networks". In: *Proc. ACM Conf. on Web Science*. 2016, pp. 259–268.

[50] X. Deng et al. "Credit distribution for influence maximization in online social networks with node features". In: *J Intell Fuzzy Syst* 31.2 (2016), pp. 979–990.

[51] P. S. Dodds. "Slightly Generalized Contagion: Unifying Simple Models of Biological and Social Spreading". In: *Complex Spreading Phenomena in Social Systems*. Ed. by S. Lehman and Y.-Y. Ahn. Computational Social Sciences. 2018, pp. 67–80.

[52] Marina Drosou and Evaggelia Pitoura. "DisC diversity: result diversification based on dissimilarity and coverage". In: *PVLDB* 6.1 (2012), pp. 13–24.

[53] Marina Drosou et al. "Diversity in Big Data: A Review". In: *Big data* 5 2 (2017), pp. 73–84.

[54] Elizabeth Dubois and Grant Blank. "The echo chamber is overstated: the moderating effect of political interest and diverse media". In: *Information, Communication & Society* 21.5 (2018), pp. 729–745. DOI: 10.1080/1369118X.2018.1428656. eprint: https://doi.org/10.1080/1369118X.2018.1428656. URL: https://doi.org/10.1080/1369118X.2018.1428656.

[55] Richard Durrett. *Lecture notes on particle systems and percolation*. The Wadsworth & Brooks/Cole statistics/probability series. Pacific Grove, Calif: Wadsworth & Brooks/Cole Advanced Books & Software, 1988. ISBN: 9780534094621.

[56] Noella Edelmann. "Reviewing the Definitions of "Lurkers" and Some Implications for Online Research". In: *Cyberpsychology, behavior and social networking* 16 (July 2013). DOI: 10.1089/cyber.2012.0362.

[57] L. Fan et al. "Least cost rumor blocking in social networks". In: *Proc. Distr. Comp. Syst. Conf.* 2013, pp. 540–549.

[58] M. Fazli et al. "On non-progressive spread of influence through social networks". In: *Theor. Comput. Sci.* 550 (2014), pp. 36–50.

[59] Linton C. Freeman. "Centrality in social networks: Conceptual clarification". In: *Social Networks* 1.3 (1979), pp. 215–239. DOI: 10.1016/0378-8733(78)90021-7. URL: http://dx.doi.org/10.1016/0378-8733(78)90021-7.

[60] Y.-H. Fu, C.-Y. Huang, and C.-T. Sun. "Using global diversity and local topology features to identify influential network spreaders". In: *Physica A: Statistical Mechanics and its Applications* 433.C (2015), pp. 344–355.

[61] S. Fujishige. "Polymatroid dependence structure of a set of random variables". In: *Inform. Contr.* 39 (1978), pp. 55–72.

[62] Sainyam Galhotra et al. "ASIM: A Scalable Algorithm for Influence Maximization under the Independent Cascade Model". In: May 2015. DOI: 10.1145/2740908.2742725.

[63] A. Ganesh, L. Massoulie, and D. Towsley. "The effect of network topology on the spread of epidemics". In: *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.* Vol. 2. 2005, 1455–1466 vol. 2. DOI: 10.1109/INFCOM.2005.1498374.

[64] K. Garimella et al. "The Effect of Collective Attention on Controversial Debates on Social Media". In: *Proc. ACM Conf. on Web Science.* 2017, pp. 43–52.

[65] C. Giatsidis, Di. M. Thilikos, and M. Vazirgiannis. "D-cores: measuring collaboration of directed graphs based on degeneracy". In: *Knowl. Inf. Syst.* 35.2 (2013), pp. 311–343.

[66] Eric Gilbert and Karrie Karahalios. "Predicting tie strength with social media". In: *Proc. Int. Conf. on Human Factors in Computing Systems (CHI).* 2009, pp. 211–220.

[67] J. Golbeck and J. A. Hendler. "Inferring binary trust relationships in Web-based social networks". In: *ACM Trans. Internet Techn.* 6.4 (2006), pp. 497–529.

[68] Jacob Goldenberg and Barak Libai. "Using Complex Systems Analysis to Advance Marketing Theory Development: Modeling Heterogeneity Effects on New Product Growth through Stochastic Cellular Automata". In: *Acad. Market. Sci. Rev* 9 (Jan. 2001).

[69] Jacob Goldenberg, Barak Libai, and Eitan Muller. "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth". In: *Marketing Letters* (2001).

[70] Sreenivas Gollapudi and Aneesh Sharma. "An axiomatic approach for result diversification". In: *Proc. ACM Conf. on World Wide Web (WWW)*. 2009, pp. 381–390.

[71] P. Govindan et al. "The K-peak Decomposition: Mapping the Global Structure of Graphs". In: *Proc. WebConf*. 2017.

[72] A. Goyal, W. Lu, and L. V. S. Lakshmanan. "SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model". In: *2011 IEEE 11th International Conference on Data Mining*. 2011, pp. 211–220.

[73] Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. "Learning influence probabilities in social networks". In: *Proc. Int. Conf. on Web Search and Web Data Mining (WSDM)*. 2010, pp. 241–250.

[74] Amit Goyal, Wei Lu, and Laks V.S. Lakshmanan. "CELF++: Optimizing the Greedy Algorithm for Influence Maximization in Social Networks". In: *Proceedings of the 20th International Conference Companion on World Wide Web*. WWW '11. Hyderabad, India: Association for Computing Machinery, 2011, 47–48. ISBN: 9781450306379. DOI: 10.1145/1963192.1963217. URL: https://doi.org/10.1145/1963192.1963217.

[75] M. Granovetter. "Threshold Models of Collective Behavior". In: *The American Journal of Sociology* 83.6 (1978), pp. 1420–1443.

[76] B. Guler et al. "Optimal strategies for targeted influence in signed networks". In: *Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*. 2014, pp. 906–911.

[77] J. Guo et al. "Personalized influence maximization on social networks". In: *Proc. ACM Conf. on Information and Knowledge Management (CIKM)*. 2013, pp. 199–208.

[78] Furkan Gursoy and Dilek Günneç. "Influence maximization in social networks under Deterministic Linear Threshold Model". In: *Knowl.-Based Syst.* 161 (2018), pp. 111–123.

[79] S. Hamdi et al. "Trust Inference Computation for Online Social Networks." In: *Proc. IEEE TrustCom/ISPA/IUCC*. 2013, pp. 210–217.

[80] L. Han, Z. Ma, and T. Shi. "An SIRS epidemic model of two competitive species". In: *Mathematical and Computer Modelling* 37.1-2 (2003), pp. 87–108.

[81] David Harrison and Katherine Klein. "What's the Difference? Diversity Constructs as Separation, Variety, or Disparity in Organizations". In: *Academy of Management Review* 32 (Oct. 2007). DOI: 10.5465/AMR.2007.26586096.

[82] Xinran He et al. "Influence Blocking Maximization in Social Networks under the Competitive Linear Threshold Model". In: *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 463–474. DOI: 10.1137/1.9781611972825.40.

[83] H. Hethcote. "The Mathematics of Infectious Diseases". In: *SIAM Review* 42.4 (2000), pp. 599–653.

[84] Shawndra Hill, Foster Provost, and Chris Volinsky. "Network-Based Marketing: Identifying Likely Adopters Via Consumer Networks". In: *Statistical Science* 21 (July 2006). DOI: 10.1214/088342306000000222.

[85] M. Hu et al. "The analysis of epidemic disease propagation in competition environment". In: *Proc. Intelligent Automation Conf.* 2013, pp. 227–234.

[86] Huimin Huang et al. "Community-based influence maximization for viral marketing". In: *Appl. Intell.* 49.6 (2019), pp. 2137–2150.

[87] Keke Huang et al. "Revisiting the Stop-and-Stare Algorithms for Influence Maximization". In: *Proc. VLDB Endow.* 10.9 (May 2017), 913–924. ISSN: 2150-8097. DOI: 10.14778/3099622.3099623. URL: https://doi.org/10.14778/3099622.3099623.

[88] P.-Y. Huang et al. "The Impact of Social Diversity and Dynamic Influence Propagation for Identifying Influencers in Social Networks". In: *Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence*. 2013, pp. 410–416.

[89] Jehad Imlawi and Dawn Gregg. "Engagement in Online Social Networks: The Impact of Self-Disclosure and Humor". In: *International Journal of Human-Computer Interaction* 30 (Feb. 2014). DOI: 10.1080/10447318.2013.839901.

[90] G. Iniguez et al. "Service Adoption Spreading in Online Social Networks". In: *Complex Spreading Phenomena in Social Systems*. Ed. by S. Lehman and Y.-Y. Ahn. Computational Social Sciences. 2018, pp. 151–175.

[91] R. Interdonato, C. Pulice, and A. Tagarelli. "Community-based delurking in social networks". In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016, pp. 263–270. DOI: 10.1109/ASONAM.2016.7752244.

[92] R. Interdonato, C. Pulice, and A. Tagarelli. "Got to have faith! : The DEvOTION algorithm for delurking in social networks". In: *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2015, pp. 314–319.

[93] J.L. Iribarren and E. Moro. "Impact of human activity patterns on the dynamics of information diffusion". In: *Phys. Rev. Lett.* 103.3 (2009).

[94] Chunyan Ji et al. "Dynamics of a multigroup SIR epidemic model with stochastic perturbation". In: *Automatica* 48.1 (2012), pp. 121–131.

[95] W. Jiang, G. Wang, and J. Wu. "Generating trusted graphs for trust evaluation in online social networks." In: *Future Generation Comp. Syst.* (2014), pp. 48–58.

[96] K. Jung, W. Heo, and W. Chen. "IRIE: Scalable and Robust Influence Maximization in Social Networks". In: *2012 IEEE 12th International Conference on Data Mining*. 2012, pp. 918–923. DOI: 10.1109/ICDM.2012.79.

[97] D. Kempe, J. M. Kleinberg, and É. Tardos. "Maximizing the spread of influence through a social network". In: *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 2003, pp. 137–146.

[98] Dongeun Kim et al. "Influence maximization based on reachability sketches in dynamic graphs". In: *Inf. Sci.* 394 (2017), pp. 217–231.

[99] J. Kim et al. "Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation". In: *Proc. ACM Conf. on Web Search and Data Mining*. 2018, pp. 324–332.

[100] Masahiro Kimura and Kazumi Saito. "Tractable Models for Information Diffusion in Social Networks". In: *Knowledge Discovery in Databases: PKDD 2006*. Ed. by Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 259–271. ISBN: 978-3-540-46048-0.

[101] M. Kitsak et al. "Identification of influential spreaders in complex networks". In: *Nature Physics* 6.11 (2010), pp. 888–893.

[102] D. Koutra, P. N. Bennett, and E. Horvitz. "Events and Controversies: Influences of a Shocking News Event on Information Seeking". In: *Proc. World Wide Web Conf.* 2015, pp. 614–624.

[103] S. Krishnan et al. "Seeing the Forest for the Trees: New Approaches to Forecasting Cascades". In: *Proc. ACM Conf. on Web Science*. 2016, pp. 249–258.

[104] K P Krishna Kumar and G Geethakumari. "Detecting misinformation in online social networks using cognitive psychology". In: *Human-cent. Comp. and Inf. Sci.* 4 (2014), pp. 1–22.

[105] S. Kumar, R. West, and J. Leskovec. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes". In: *Proc. World Wide Web Conf.* 2016, pp. 591–602.

[106] Srijan Kumar et al. "Community interaction and conflict on the web". In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. 2018, pp. 933–943.

[107] Jérôme Kunegis et al. "Diversity Dynamics in Online Networks". In: *Proceedings of the 23rd ACM Conference on Hypertext and Social Media*. HT '12. Milwaukee, Wisconsin, USA: Association for Computing Machinery, 2012, 255–264. ISBN: 9781450313353. DOI: 10.1145/2309996.2310039. URL: https://doi.org/10.1145/2309996.2310039.

[108] Timothy La Fond and Jennifer Neville. "Randomization Tests for Distinguishing Social Influence and Homophily Effects". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, 601–610. ISBN: 9781605587998. DOI: 10.1145/1772690.1772752. URL: https://doi.org/10.1145/1772690.1772752.

[109] C. Lagnier et al. "Predicting information diffusion in social networks using content and user's profiles". In: *Proc. European Conf. on Information Retrieval (ECIR)*. 2013, pp. 74–85.

[110] J. Lee and C. Chung. "A Query Approach for Influence Maximization on Specific Users in Social Networks". In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 340–353. DOI: 10.1109/TKDE.2014.2330833.

[111] J.R. Lee and C.W. Chung. "A query approach for influence maximization on specific users in social networks". In: *IEEE Trans Knowl Data Eng* 27.2 (2015), pp. 340–353.

[112] J. Leskovec, D. Huttenlocher, and J. Kleinberg. "Governance in Social Media: A Case Study of the Wikipedia Promotion Process". In: *Proc. Conf. on Weblogs and Social Media*. 2010.

[113] J. Leskovec, D. Huttenlocher, and J. Kleinberg. "Signed networks in social media". In: *Proc. Conf. on Human Factors in Computing Systems*. 2010, pp. 1361–1370.

[114] J. Leskovec et al. "Cost-effective Outbreak Detection in Networks". In: *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 2007, pp. 420–429.

[115] J. Leskovec et al. "Cost-effective Outbreak Detection in Networks". In: *Proc. ACM KDD*. 2007, pp. 420–429.

[116] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. "The dynamics of viral marketing". In: *ACM Transactions on the Web (TWEB)* 1.1 (2007), p. 5.

[117] Jure Leskovec and Julian J. Mcauley. "Learning to Discover Social Circles in Ego Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 539–547. URL: http://papers.nips.cc/paper/4532-learning-to-discover-social-circles-in-ego-networks.pdf.

[118] Jure Leskovec, Ajit Singh, and Jon Kleinberg. "Patterns of Influence in a Recommendation Network". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Wee-Keong Ng et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 380–389. ISBN: 978-3-540-33207-7.

[119] Stephan Lewandowsky et al. "Misinformation and Its Correction". In: *Psychological Science in the Public Interest* 13.3 (2012), pp. 106–131.

[120] F. Li, C. Li, and M. Shan. "Labeled Influence Maximization in Social Networks for Target Marketing". In: *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 2011, pp. 560–563. DOI: 10.1109/PASSAT/SocialCom.2011.152.

[121] H. Li et al. "Conformity-aware influence maximization in online social networks". In: *The VLDB Journal* 24 (2015), pp. 117–141.

[122] R. Li and J. X. Yu. "Scalable Diversified Ranking on Large Graphs". In: *2011 IEEE 11th International Conference on Data Mining*. 2011, pp. 1152–1157. DOI: 10.1109/ICDM.2011.126.

[123] X. Li et al. "Why approximate when you can get the exact? Optimal targeted viral marketing at scale". In: *Proc. IEEE Conf. on Computer Communications (INFOCOM)*. 2017, pp. 1–9.

[124] Xiao Li et al. "Community-based seeds selection algorithm for location aware influence maximization". In: *Neurocomputing* 275 (2018), pp. 1601–1613.

[125] Y. Li et al. "Influence Maximization on Social Graphs: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* 30.10 (2018), pp. 1852–1872.

[126] Yiqing Li et al. "Conformity-Aware Influence Maximization with User Profiles". In: *Proc. Int. Conf. on Wireless Communications and Signal Processing (WCSP)*. 2018, pp. 1–6.

[127] Yuchen Li, Dongxiang Zhang, and Kian-Lee Tan. "Real-time Targeted Influence Maximization for Online Advertisements". In: *Proc. VLDB Endow.* 8.10 (2015), pp. 1070–1081.

[128] Yuchen Li, Dongxiang Zhang, and Kian-Lee Tan. "Real-time targeted influence maximization for online advertisements". In: *Proceedings of the VLDB Endowment* 8 (June 2015), pp. 1070–1081. DOI: 10.14778/2794367.2794376.

[129] D. Liben-Nowell and J. Kleinberg. "Tracing information flow on a global scale using Internet chain-letter data". In: *Procs. of the National Academy of Sciences* 105.12 (2008), pp. 4633–4638.

[130] I. Litou et al. “Real-Time and Cost-Effective Limitation of Misinformation Propagation”. In: *Proc. IEEE Conf. on Mobile Data Management*. 2016, pp. 158–163.

[131] B. Liu et al. “Time constrained influence maximization in social networks”. In: *Proc. IEEE Conf. on Data Mining*. 2012, pp. 439–448.

[132] H. Liu et al. “Predicting trusts among users of online communities: an epinions case study”. In: *Proc. ACM Conf. on Electronic Commerce (EC)*. 2008, pp. 310–319.

[133] Q. Liu et al. “Social Marketing Meets Targeted Customers: A Typical User Selection and Coverage Perspective”. In: *2014 IEEE International Conference on Data Mining*. 2014, pp. 350–359. DOI: 10.1109/ICDM.2014.93.

[134] Qi Liu et al. “Influence Maximization over Large-Scale Social Networks”. In: Nov. 2014, pp. 171–180. DOI: 10.1145/2661829.2662009.

[135] V.Y. Lou et al. “Modeling non-progressive phenomena for influence propagation”. In: *Proc. ACM Conf. on Online Social Networks*. 2014, pp. 131–137.

[136] L. Lovász. “Submodular functions and convexity”. In: *Mathematical Programming: The State of the Art*. Ed. by A. Bachem, B. Korte, and M. Grötschel. Springer-Verlag Berlin Heidelberg, 1983, pp. 235–257.

[137] W. Lu, W. Chen, and L. V. S. Lakshmanan. “From Competition to Complementarity: Comparative Influence Diffusion and Maximization”. In: *Proc. VLDB Endow.* 9.2 (2015), pp. 60–71.

[138] W. Lu, W. Chen, and L. V. S. Lakshmanan. “From Competition to Complementarity: Comparative Influence Diffusion and Maximization”. In: *Proc. VLDB Endow.* 9.2 (2015), pp. 60–71.

[139] Wei Lu, Wei Chen, and Laks V. S. Lakshmanan. “From Competition to Complementarity: Comparative Influence Diffusion and Maximization”. In: *Proc. VLDB Endow.* 9.2 (Oct. 2015), 60–71. ISSN: 2150-8097. DOI: 10.14778/2850578.2850581. URL: https://doi.org/10.14778/2850578.2850581.

[140] Wei Lu and Laks V. S. Lakshmanan. “Profit Maximization over Social Networks”. In: *Proc. IEEE Int. Conf. on Data Mining (ICDM)*. 2012, pp. 479–488.

[141] Wei Lu et al. *Refutations on "Debunking the Myths of Influence Maximization: An In-Depth Benchmarking Study"*. 2017. arXiv: 1705.05144 [cs.SI].

[142] Guowei Ma et al. “Identifying Hesitant and Interested Customers for Targeted Social Marketing”. In: vol. 9077. May 2015, pp. 576–590. ISBN: 978-3-319-18037-3. DOI: 10.1007/978-3-319-18038-0_45.

[143] Q. Ma and J. Ma. “Identifying and ranking influential spreaders in complex networks with consideration of spreading probability”. In: *Physica A: Statistical Mechanics and its Applications* 465 (2017), pp. 312–330.

[144] Nilly Madar et al. “Immunization and Epidemic Dynamics in Complex Networks”. In: *Physics of Condensed Matter* 38 (Mar. 2004), pp. 269–276. DOI: 10.1140/epjb/e2004-00119-8.

[145] F.D. Malliaros, M.-E.G. Rossi, and M. Vazirgiannis. “Locating influential nodes in complex networks”. In: *Scientific Reports* 6 (2016).

[146] Fragkiskos D. Malliaros and Michalis Vazirgiannis. "To Stay or Not to Stay: Modeling Engagement Dynamics in Social Graphs". In: *Proceedings of the 22nd ACM International Conference on Information amp; Knowledge Management*. CIKM '13. San Francisco, California, USA: Association for Computing Machinery, 2013, 469–478. ISBN: 9781450322638. DOI: 10.1145/2505515.2505561. URL: https://doi.org/10.1145/2505515.2505561.

[147] Newman ME. "Spread of epidemic disease on networks." In: ().

[148] Yasir Mehmood et al. "CSI: Community-Level Social Influence Analysis". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Hendrik Blockeel et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 48–63. ISBN: 978-3-642-40991-2.

[149] P. Metaxas and E. Mustafaraj. "From Obscurity to Prominence in Minutes: Political Speech and Real-Time Search". In: *Proc. ACM Conf. on Web Science*. 2010.

[150] R. Mohamadi-Baghmolaei, N. Mozafari, and A. Hamzeh. "Trust based latency aware influence maximization in social networks". In: *Engineering Applications of Artificial Intelligence* 41 (2015), pp. 195–206.

[151] E. Mustafaraj and P. T. Metaxas. "The Fake News Spreading Plague: Was It Preventable?" In: *Proc. ACM Conf. on Web Science*. 2017, pp. 235–239.

[152] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. "An analysis of approximations for maximizing submodular set functions-I". In: *Mathematical Programming* 14.1 (1978), pp. 265–294.

[153] H. T. Nguyen, T. N. Dinh, and M. T. Thai. "Cost-aware Targeted Viral Marketing in billion-scale networks". In: *Proc. IEEE Conf. on Computer Communications (INFOCOM)*. 2016, pp. 1–9.

[154] H. T. Nguyen, M. T. Thai, and T. N. Dinh. "Stop-and-Stare: Optimal Sampling Algorithms for Viral Marketing in Billion-scale Networks". In: *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD)*. 2016, pp. 695–710.

[155] Lan N. Nguyen, Kunxiao Zhou, and My T. Thai. "Influence Maximization at Community Level: A New Challenge with Non-submodularity". In: *Proc. IEEE Int. Conf. on Distributed Computing Systems (ICDCS)*. 2019, pp. 327–337.

[156] Naoto Ohsaka et al. "Fast and Accurate Influence Maximization on Large Networks with Pruned Monte-Carlo Simulations". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 28.1 (2014). URL: https://ojs.aaai.org/index.php/AAAI/article/view/8726.

[157] J. Overgoor, E. Wulczyn, and C. Potts. "Trust Propagation with Mixed-Effects Models". In: *Proc. Int. Conf. on Weblogs and Social Media (ICWSM)*. 2012.

[158] M. R. Padmanabhan et al. "Influence Maximization in Social Networks With Non-Target Constraints". In: *Proc. IEEE Int. Conf. on Big Data*. 2018, pp. 771–780.

[159] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, 1999. URL: http://ilpubs.stanford.edu:8090/422/.

[160] Zhao Pan, Yaobin Lu, and Sumeet Gupta. "How heterogeneous community engage newcomers? The effect of community diversity on newcomers' perception of inclusion: An empirical study in social media service". In: *Computers in Human Behavior* 39 (2014), pp. 100 –111. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2014.05.034. URL: http://www.sciencedirect.com/science/article/pii/S0747563214003203.

[161] S. Peng et al. "Influence analysis in social networks: A survey". In: *Journal of Network and Computer Applications* 106 (2018), pp. 17–32.

[162] K. W. Phillips. "How Diversity Makes Us Smarter". In: 311.4 (2014).

[163] M. A. Porter and J. P. Gleeson. *Dynamical Systems on Networks*. Frontiers in Applied Dynamical Systems: Reviews and Tutorials. Gewerbestrasse 11, 6330 Cham, Switzerland: Springer Int. Publishing, 2016.

[164] Adarsh Prasad, Stefanie Jegelka, and Dhruv Batra. "Submodular meets Structured: Finding Diverse Subsets in Exponentially-Large Structured Item Sets". In: *arXiv CoRR* abs/1411.1752 (2014).

[165] L. Qiu et al. "PHG: A Three-Phase Algorithm for Influence Maximization Based on Community Structure". In: *IEEE Access* 7 (2019), pp. 62511–62522.

[166] Lionel Robert and Daniel M. Romero. "Crowd Size, Diversity and Performance". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2015, 1379–1382. ISBN: 9781450331456. URL: https://doi.org/10.1145/2702123.2702469.

[167] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. "Differences in the Mechanics of Information Diffusion across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter". In: *Proceedings of the 20th International Conference on World Wide Web*. WWW '11. Hyderabad, India: Association for Computing Machinery, 2011, 695–704. ISBN: 9781450306324. DOI: 10.1145/1963405.1963503. URL: https://doi.org/10.1145/1963405.1963503.

[168] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. "Prediction of Information Diffusion Probabilities for Independent Cascade Model". In: *Proc. Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES)*. 2008, pp. 67–75.

[169] Francisco C Santos, Marta D Santos, and Jorge M Pacheco. "Social diversity promotes the emergence of cooperation in public goods games." In: *Nature* 454.7201 (2008).

[170] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. "Search Result Diversification". In: *Found. Trends Inf. Retr.* 9.1 (Mar. 2015), pp. 1–90. ISSN: 1554-0669.

[171] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. "Search Result Diversification". In: *Found. Trends Inf. Retr.* 9.1 (Mar. 2015), 1–90. ISSN: 1554-0669. DOI: 10.1561/1500000040. URL: https://doi.org/10.1561/1500000040.

[172] S. B. Seidman. "Network structure and minimum degree". In: *Social Networks* 5.3 (1983), pp. 269–287.

[173] Jiaxing Shang et al. "CoFIM: A community-based framework for influence maximization on large-scale networks". In: *Knowl.-Based Syst.* 117 (2017), pp. 88–100.

[174] W. Sherchan, S. Nepal, and C. Paris. "A survey of trust in social networks". In: *ACM Comput. Surv.* (2013), p. 47.

[175] Shashank Sheshar Singh et al. "C2IM: Community based context-aware influence maximization in social networks". In: *Physica A* 514 (2019), pp. 796–818.

[176] Vladimir Soroka and Sheizaf Rafaeli. "Invisible Participants: How Cultural Capital Relates to Lurking Behavior". In: *Proceedings of the 15th International Conference on World Wide Web*. WWW '06. Edinburgh, Scotland: Association for Computing Machinery, 2006, 163–172. ISBN: 1595933239. DOI: 10.1145/1135777.1135806. URL: https://doi.org/10.1145/1135777.1135806.

[177] Ana-Andreea Stoica and Augustin Chaintreau. "Fairness in Social Influence Maximization". In: *Proc. ACM Conf. on World Wide Web (WWW)*. 2019, pp. 569–574.

[178] N. Sumith, B. Annappa, and S. Bhattacharya. "Influence maximization in large social networks: Heuristics, models and parameters". In: *Future Generation Computer Systems* 89 (2018), pp. 777–790.

[179] Na Sun, Patrick Pei-Luen Rau, and Liang Ma. "Understanding lurkers in online communities: A literature review". In: *Computers in Human Behavior* 38 (2014), pp. 110 –117. ISSN: 0747-5632. DOI: https://doi.org/10.1016/j.chb.2014.05.022. URL: http://www.sciencedirect.com/science/article/pii/S0747563214003008.

[180] Andrea Tagarelli and Roberto Interdonato. "Lurking in Social Networks: Topology-based Analysis and Ranking Methods". In: *Social Network Analysis and Mining* 4 (Sept. 2014). DOI: 10.1007/s13278-014-0230-4.

[181] Andrea Tagarelli and Roberto Interdonato. "Time-aware Analysis and Ranking of Lurkers in Social Networks". In: *Social Network Analysis and Mining* 5 (Dec. 2015). DOI: 10.1007/s13278-015-0276-y.

[182] Andrea Tagarelli and Roberto Interdonato. ""Who's out There?": Identifying and Ranking Lurkers in Social Networks". In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM '13. Niagara, Ontario, Canada: Association for Computing Machinery, 2013, 215–222. ISBN: 9781450322409. DOI: 10.1145/2492517.2492542. URL: https://doi.org/10.1145/2492517.2492542.

[183] M. Talluri, H. Kaur, and J.S. He. "Influence maximization in social networks: Considering both positive and negative relationships". In: *Proc. Conf. on Collaboration Technologies and Systems*. 2015, pp. 479–480.

[184] F. Tang et al. "Diversified social influence maximization". In: *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. 2014, pp. 455–459. DOI: 10.1109/ASONAM.2014.6921625.

[185] F. Tang et al. "Diversified social influence maximization". In: *Proc. IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining (ASONAM)*. 2014, pp. 455–459.

[186] Jiliang Tang and Huan Liu. *Trust in Social Media*. Synthesis Lectures on Information Security, Privacy, & Trust. 1210 Fifth Ave 250, San Rafael, US: Morgan & Claypool Publishers, 2015.

[187] Jing Tang, Xueyan Tang, and Junsong Yuan. "Profit Maximization for Viral Marketing in Online Social Networks: Algorithms and Analysis". In: *IEEE Trans. Knowl. Data Eng.* 30.6 (2018), pp. 1095–1108.

[188] Y. Tang, Y. Shi, and X. Xiao. "Influence Maximization in Near-Linear Time: A Martingale Approach". In: *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD)*. 2015, pp. 1539–1554.

[189] Y. Tang, X. Xiao, and Y. Shi. "Influence Maximization: Near-optimal Time Complexity Meets Practical Efficiency". In: *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD)*. 2014, pp. 75–86.

[190] Steven J. J. Tedjamulia et al. "Motivating Content Contributions to Online Communities: Toward a More Comprehensive Theory". In: *Proc. Int. Conf. on System Sciences (HICSS)*. 2005.

[191] G. Tong et al. "An Efficient Randomized Algorithm for Rumor Blocking in Online Social Networks". In: *Proc. IEEE Conf. on Computer Communications*. 2017, pp. 1–9.

[192] Leslie G Valiant. "The complexity of enumeration and reliability problems". In: *SIAM Journal on Computing* 8.3 (1979), pp. 410–421.

[193] Sharan Vaswani et al. "Model-Independent Online Learning for Influence Maximization". In: *Proc. Int. Conf. on Machine Learning (ICML)*. 2017, pp. 3530–3539.

[194] Alexei Vazquez et al. "Impact of Non-Poissonian Activity Patterns on Spreading Processes". In: *Phys. Rev. Lett.* 98 (15 2007), p. 158702.

[195] N. Vedula, S. Parthasarathy, and V. L. Shalin. "Predicting Trust Relations Within a Social Network: A Case Study on Emergency Response". In: *Proc. ACM Conf. on Web Science*. 2017, pp. 53–62.

[196] J. Wang and J. Cheng. "Truss Decomposition in Massive Networks". In: *Proc. VLDB Endow.* 5.9 (2012), pp. 812–823.

[197] Yu Wang et al. "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks". In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 1039–1048.

[198] Z. Wang et al. "Exploiting social influence for context-aware event recommendation in event-based social networks". In: *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*. 2017, pp. 1–9. DOI: 10.1109/INFOCOM.2017.8057167.

[199] Duncan Watts. "A Simple Model of Global Cascades on Random Networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 99 (May 2002), pp. 5766–71. DOI: 10.1073/pnas.082090499.

[200] Lilian Weng et al. "The Role of Information Diffusion in the Evolution of Social Networks". In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '13. Chicago, Illinois, USA: Association for Computing Machinery, 2013, 356–364. ISBN: 9781450321747. DOI: 10.1145/2487575.2487607. URL: https://doi.org/10.1145/2487575.2487607.

[201] X. Weng, Z. Liu, and Z. Li. "An efficient influence maximization algorithm considering both positive and negative relationships". In: *Proc. IEEE TrustCom/BigDataSE/ISPA*. 2016, pp. 1931–1936.

[202] Le Wu et al. "Relevance Meets Coverage: A Unified Framework to Generate Diversified Recommendations". In: *ACM Trans. Intell. Syst. Technol.* 7.3 (Feb. 2016), 39:1–39:30. ISSN: 2157-6904.

[203] Le Wu et al. "Relevance Meets Coverage: A Unified Framework to Generate Diversified Recommendations". In: *ACM Trans. Intell. Syst. Technol.* 7.3 (Feb. 2016). ISSN: 2157-6904. DOI: 10.1145/2700496. URL: https://doi.org/10.1145/2700496.

[204] D.-N. Yang et al. "Maximizing acceptance probability for active friending in online social networks". In: *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*. 2013, pp. 713–721.

[205] G. Yin et al. "AUTrust: A Practical Trust Measurement for Adjacent Users in Social Networks". In: *Proc. Int. Conf. on Cloud and Green Computing (CGC)*. 2012, pp. 360–367.

[206] Yuan Yuan, Ahmad Alabdulkareem, and Alex Pentland. "An interpretable approach for social network formation among heterogeneous agents". In: *Nature Communications* 9 (Nov. 2018). DOI: 10.1038/s41467-018-07089-x.

[207] Muhammad Bilal Zafar et al. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proc. ACM Conf. on World Wide Web (WWW)*. 2017, pp. 1171–1180.

[208] Richard S. Zemel et al. "Learning Fair Representations". In: *Proc. Int. Conf. on Machine Learning (ICML)*. 2013, pp. 325–333.

[209] K. Zhang et al. "A Core Theory Based Algorithm for Influence Maximization in Social Networks". In: *2017 IEEE International Conference on Computer and Information Technology (CIT)*. 2017, pp. 31–36.

[210] Xiaohang Zhang et al. "Identifying influential nodes in complex networks with community structure". In: *Knowl.-Based Syst.* 42 (2013), pp. 74–84.

[211] C. Zhou et al. "On the Upper Bounds of Spread for Greedy Algorithms in Social Network Influence Maximization". In: *IEEE Transactions on Knowledge and Data Engineering* 27.10 (2015), pp. 2770–2783. DOI: 10.1109/TKDE.2015.2419659.

[212] J. Zhou, Y. Zhang, and J. Cheng. "Preference-based mining of top-k influential nodes in social networks". In: *Future Generation Comp. Syst.* 31 (2014), pp. 40–47.