UNIVERSIDADE FEDERAL FLUMINENSE

RONALD CHIESSE DE SOUZA

EFFICIENT SEEDING STRATEGIES FOR BUDGETED INFLUENCE MAXIMIZATION IN COMPLEX NETWORKS

NITERÓI

2016

UNIVERSIDADE FEDERAL FLUMINENSE

RONALD CHIESSE DE SOUZA

EFFICIENT SEEDING STRATEGIES FOR BUDGETED INFLUENCE MAXIMIZATION IN COMPLEX NETWORKS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: SISTEMAS DE COMPUTAÇÃO.

Orientador:
ANTONIO AUGUSTO DE ARAGÃO ROCHA

Co-orientador:
DANIEL RATTON FIGUEIREDO

NITERÓI

2016

RONALD CHIESSE DE SOUZA

EFFICIENT SEEDING STRATEGIES FOR BUDGETED INFLUENCE MAXIMIZATION IN COMPLEX NETWORKS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do Grau de Mestre em Computação. Área de concentração: SISTEMAS DE COMPUTAÇÃO.

Aprovada em 21 de dezembro de 2016.

BANCA EXAMINADORA

Prof. ANTONIO AUGUSTO DE ARAGÃO ROCHA – Orientador, UFF

Prof. DANIEL RATTON FIGUEIREDO – Orientador, UFRJ

Prof. FÁBIO PROTTI, UFF

Prof. VALMIR CARNEIRO BARBOSA, UFRJ

Niterói

2016

À minha família.

Agradecimentos

A Deus acima de tudo.
À família e aos amigos por todo o apoio.
À minha namorada, Graciele, pela parceria, no sentido maior da palavra.
Ao Professor Artur Ziviani, pela contribuição ativa e fundamental ao longo de todo o desenvolvimento desse trabalho.
Ao caros Jefferson Simões, Giulio Lacobelli e Don Towsley, pelas ideias sugeridas e comentários realizados, os quais serviram de base para algumas das principais formulações aqui apresentadas.
Aos meus orientadores Antonio Augusto Rocha e Daniel Figueiredo, pela contínua transmissão de experiências e conhecimentos valiosos, já desde antes do início deste Curso de Mestrado.
Aos professores Fábio Protti e Valmir Barbosa, pela presença na banca examinadora.

Resumo

A eficiência de processos de difusão de informação em redes complexas depende muito não apenas da estrutura da rede mas também dos nós escolhidos para iniciar a difusão, os chamados *seeds*. Além disso, em diversos cenários reais, nós distintos impõem custos distintos para iniciar uma difusão, como é o caso do *marketing* viral, onde o custo dos nós correlaciona-se com a estrutura local dos mesmos. O problema conhecido como *budgeted influence maximization* (BIM) consiste em determinar um conjunto de *seeds* cuja difusão maximiza o total de nós influenciados, dado que o custo total dos *seeds* limita-se a um orçamento (*budget*). Este trabalho investiga estrategias eficientes para BIM sob o modelo determinístico de difusão baseado em limiares fixos. Em particular, introduz-se o conceito de *extended surrounding sets* (conjuntos evolventes estendidos): *seeds* relativamente baratos que são vizinhos de nós caros porém estruturalmente privilegiados, os quais serão influenciados a custos mais baixos e assim contribuirão para a difusão. De modo a medir o desempenho de maneira mais eficaz, é introduzido o conceito de "poder de difusão", o qual captura a fração de nós influenciados ao fim do espalhamento descontando-se a fração de *seeds*. É mostrado como essa abordagem pode mudar completamente o entendimento acerca da eficácia de uma dada estratégia. Experimentos numéricos com diversas redes complexas de larga escala (oriundas de bases de dados reais) indicam que o método aqui apresentado é superior a estratégias que se baseiam na razão centralidade/custo dos nós para escolher os *seeds*. Uma ideia chave deste trabalho é que geralmente obtém-se maior poder de difusão quando considera-se a vizinhança de dois saltos de um vértice ao cercá-lo, ao invés de considerar-se apenas os seus vizinhos imediatos.

Palavras-chave: epidemia em redes, maximização de influência, estratégias de semeadura, custo de nó variável, rede social.

Abstract

The efficiency of information diffusion processes on complex networks highly depends on the network structure and the nodes chosen to start the diffusion, known as seeds. Moreover, in various realistic scenarios, different nodes have different costs to start a diffusion, such as in viral marketing, where node costs are correlated with their local structure. The budgeted influence maximization (BIM) problem consists of determining a seed set whose diffusion maximizes the total number of nodes influenced, provided that the total cost of the seeds is within a given budget. We investigate efficient seeding strategies for BIM under the deterministic fixed threshold diffusion model. In particular, we introduce the concept of extended surrounding sets: relatively cheap seeds neighboring expensive, yet structurally privileged nodes that will ultimately be influenced at lower costs and thus contribute to the diffusion. In order to measure performance more effectively, it is introduced the concept of "diffusion power", which captures the fraction of nodes influenced at the end of the spreading discounting the fraction of seeds. We show how this approach may completely change the understanding around a strategy's effectiveness. Numerical experiments with several large-scale complex networks (from real data social networks) indicate our method outperforms strategies that seed nodes based on their centrality/cost ratios. A key insight from our evaluation is that larger diffusion power is generally attained with the surrounding sets that consider the two hop neighborhood of central nodes, as opposed to their immediate neighbors.

Keywords: network epidemic, influence maximization, seeding strategies, variable node cost, social network.

List of Figures

List of Tables

Lista de Abreviaturas e Siglas

DP	:	Diffusion Power;
CS	:	Cost per Seed;
RDP	:	Relative Diffusion Power;
ADP	:	Average Diffusion Power;
ESS	:	Extended Surrounding Set;

Contents

Capítulo 1

Introduction

Spreading phenomena on complex networks, such as the diffusion of rumors or diseases among individuals and the sharing of online content that *go viral* on the Internet, have received an ever-growing attention both from the academia and industry. Understanding how such dynamics can be long lasting, reaching large fractions of the underlying networks, or die out quickly after a negligible spread, is a fundamental step in designing more effective diffusion processes. These include more efficient marketing campaigns across Internet users and more effective disease spreading prevention among individuals in society [28]. Likewise, cascading failures across Internet routers or in power systems can also be studied within the same framework [31, 21].

Models for diffusion processes in networks rely on the network epidemics framework, wherein an *epidemic* refers to any recursive dynamic by which propagators such as viruses, ideas, rumors, and failures, *infect* a fraction of the neighbors of those already *infected* nodes. Most importantly, an epidemic spreads through the edges of the network according to some law. It usually starts from a relatively small set of nodes, infected through some exogenous process. Intuitively, this set of initial spreaders, i.e. the set of *seeds*, plays a fundamental role on the propagation unfolding, as those structurally privileged in some sense are likely to maximize some aspect of the diffusion.

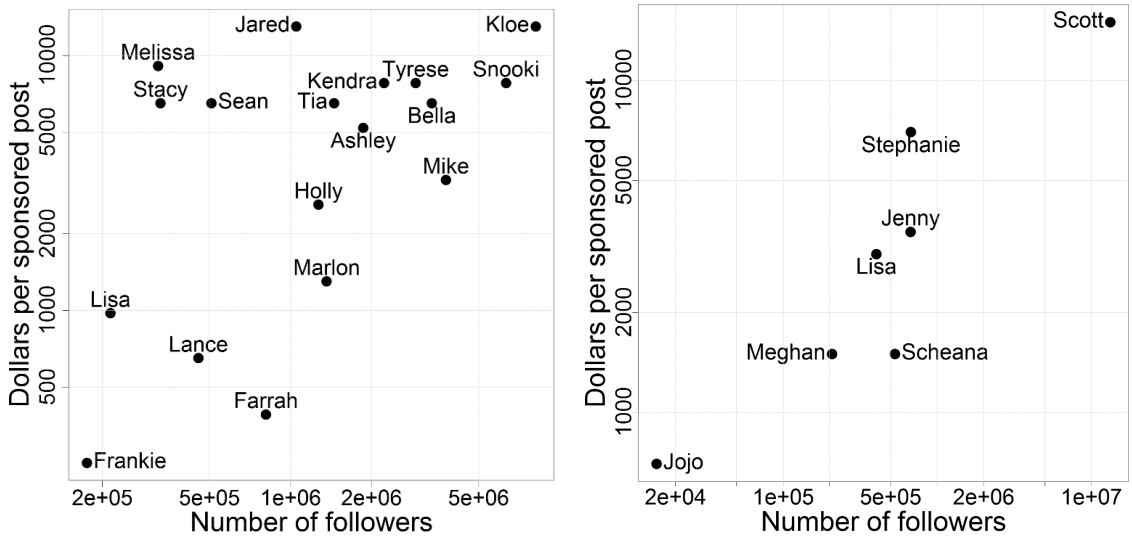
In this sense, to maximize the number of nodes an epidemic reaches corresponds to the *influence maximization* (IM) problem [13], which defines the *influence* of a set of seeds in a network as its expected number of *activations*, i.e. the expected number of nodes infected at the end of the spreading. Classically, IM problems try to identify, for a relatively small number ϕ, which ϕ-seeds are the more influential. Clearly, ϕ represents a constraint in terms of resources to perform the seeding, like an initial budget that covers an intrinsic cost to infect nodes at time zero, i.e. a cost to start an epidemic. Moreover, ϕ

also means this intrinsic cost is the same for all nodes, since it is always possible to seed any ϕ among them, irrespective of their particular properties or attributes.

However, assuming identical node costs is inadequate in many scenarios, most notably in viral marketing campaigns in online social networks (OSN) [16, 26, 6], but also epidemic routing on computer networks [29], for instance. Indeed, popular individuals (celebrities) on Twitter are paid differently to *tweet* sponsored marketing messages [16]. Likewise, through contracts of largely different values, Facebook pays companies and celebrities to sustain live video streaming content, raising users' exposure to advertisements [26]. And similarly, popular Instagram users are paid differently to post photos alongside sponsored products [6]. In another context, studies regarding predictable Delay-Tolerant Networks (DTNs) have also considered epidemic routing assuming variable node costs and limited initial budget [29].

IM problems where nodes have different initial costs to be seeded were recently formalized as the *budgeted influence maximization* (BIM) problem [25]. Given an initial budget and nodes' cost, what is the most influential seed set? Clearly, a seeding strategy must consider not only the estimated influence of a node, but also its cost towards the budget. Let us consider, for instance, both Twitter and Instagram social networks, mentioned above. Figure 1.1 illustrates how largely (orders of magnitude) the values paid for celebrities to post sponsored content may differ from one another in both networks. The exact numerical values and celebrity names are found on [16] and [6] for Twitter and Instagram networks, respectively. The plots of Fig.1.1 are as follows. For each celebrity (labeled dot), the x axis denotes his/her number of followers, while y axis is the value they charge for posting either sponsored *tweets* (Fig. 1.1(a)) or sponsored photos (Fig. 1.1(b)). Both axes are log-scaled. Notice, for instance, on Fig. 1.1(b) a huge difference between Jojo (McCarthy) and Scott (Disick) with respect to both their number of followers and their sponsored posts value—orders of magnitude. The same fact can be observed in Fig. 1.1(a), for instance, between Frankie (Muniz) and Kloe (Kardashian). Thus, it is reasonable to assume that differences among node costs, like those illustrated on Fig. 1.1, are to a certain extent related to the nodes' structural properties (in this case, the number of followers).

Clearly, effective seeding strategies must exploit the inherent cost-effectiveness in BIM. Is it better to seed more nodes at the lowest cost possible? Or to seed a few nodes better positioned at a much higher average cost? Moreover, how to cope with the correlation between node centrality and node cost?



(a) Twitter (b) Instagram

Figure 1.1: Payment received by different celebrities for posting sponsored content in Twitter (left) and Instagram (right) networks, compared to their number of followers (axes log-scaled).

This work investigates efficient seeding strategies for the BIM problem under a correlated node cost model (i.e., higher node degree, higher cost) and the linear threshold model (i.e., a node becomes infected if k or more neighbors are infected). Two key contributions are as follows: (i) the concept of *extended surrounding sets*: relatively cheap seeds near (distance of one or two hops) expensive, yet structurally privileged nodes that will become infected at lower costs, and hence contribute to the epidemic. Surrounding expensive, central nodes by cheap seeds has shown to be advantageous for the spreading independently of how node centrality is assessed; (ii) a more meaningful concept to assess epidemic performance, called *diffusion power*, which subtracts the fraction of seeds from the total fraction of nodes influenced at the end of the spreading. This metric is fundamental to properly compare and understand the effectiveness of different seeding strategies for BIM. Opposite to the unit-cost IM, two different strategies for BIM, upon receiving the same initial budget, may still hugely vary in their number of seeds. In particular, a small set of seeds that influence a large number of nodes should be valued higher than a huge set of seeds that influence just a few nodes. Diffusion power thus captures the benefits (infected non-seeds) of an investment (budget).

We empirically evaluate different seeding strategies on 7 real networks, varying parameters for node cost, budget, and infection threshold. We show that even when the cost of those most central nodes is larger than the budget, effective epidemics can be

still induced. Moreover, we show that cost-aware seeding strategies that select nodes by their cost-effectiveness fail to trigger effective epidemics much earlier than strategies that surround central nodes, considering increasing values of the network threshold. An important insight is that better cost-effectiveness is achieved if the two-hop neighbors of a central node v are also leveraged in order to have v surrounded, as opposite of considering v's direct neighbors only. Last, we believe the main contributions and findings of this work, such as the node surround concept and the measurement of epidemic performance, can be applied to other contexts.

1.1 Related Work

The influence maximization (IM) problem has been broadly investigated since the seminal work of Kempe et al. [13]. This pioneering paper provides framework for the general problem, proving its hardness (it is an NP-Hard problem) and providing an approximate greedy algorithm with provable performance (constant factor of optimal). This approximate algorithm is based on submodular objective functions, which is shown to be the case for some epidemic models. However, despite being polynomial time, the approximate greedy algorithm has very high complexity which has lead to a myriad of approaches to tackle the problem more efficiently [13, 2, 14, 8, 12, 1, 7, 30, 20, 4]. Moreover, there are various scenarios where the epidemic model does not yield submodular objective functions (or cannot be shown to yield) which is for example, the case of the classical Linear Threshold Model [10], considered in this work.

Therefore, many of prior works focused on heuristics to determine the best seed set, exploring structural features of the network as well as features associated with nodes (e.g., labels). For example, computationally-inexpensive heuristics based on node degree [8] and other simple features (such as node homophily) have been considered [1, 17]. Heuristics based on k-core decomposition of the network [27] have also been explored [14], showing a correlation between influential spreaders and highly connected regions of the network. This idea has been explored by various subsequent works that also adapt and augment with node rankings [4], communities [30], disjoint paths [7] , and local neighborhoods [20]. However, all these prior works implicitly assume that nodes have the identical costs, since the constraint is simply the number of seeds.

There are also recent works that have investigated problems related to epidemics where node costs are not fixed (over time) or identical across the network. For example,

Leskovec et al. [18] propose strategies for placing sensors on a network to more quickly detect a diffusion. Arthur et al. [2] propose strategies to price products and provide cashback (discount) to nodes in the network to induce recommendations to neighbors. Miyanchi et al. [22] formulate an optimization problem to allocate a fixed budget to a bipartite network of marketing channels and customers with variable node costs (no diffusion is considered). None of these works is considered the specific BIM problem.

However, BIM has recently been formulated and investigated by Nguyen et al. [25]. The authors depart from the framework introduced by Kempe et al. [13] and tackle the problem using the independent cascades model. They establish a submodular, cost-normalized objective function, from which they determine a greedy algorithm with approximation guarantees to a constant factor. Other works have also investigated the problem [11, 9]. Han et al. [11] tackle BIM with a heuristic combining two seeding strategies based on node influence (and not cost) and on node cost (and not influence). Last, Souza et al. [9] characterize the performance of simple and traditional seeding strategies to solve the BIM problem, motivating the need for more clever strategies.

Despite addressing the BIM problem, these prior works have the following limitations. The theoretical result of Nguyen et al.[25] assumes that the initial budget is larger than the cost of any node. Moreover, their numerical evaluation uniformly assigns random costs to nodes, from a small range (less than a factor of 10). Similarly, Han et al. [11] assume that the initial budget is larger than the cost of any node, but also that the cost of a node is linearly tied to its argued centrality. These assumptions, however, do not properly capture real-world pricing practices, such as those adopted by celebrities (nodes) for promoting viral marketing in online social networks [26, 16, 6]. In particular, there is no reason to assume that external agents (e.g. sponsors) are always provided large enough resources (budget) to seed even those more expensive nodes. Last, the work of Souza et al. [9] focus on demonstrating (i) how the assumption of nodes with different costs may dramatically impact strategies performances according to their decisions on seeding, and (ii) the strong need of leveraging network structure on BIM problems, by showing that "blind" strategies are highly sensitive to parameter variations. They demonstrate that none of such strategies can be considered consistently superior in terms of inducing broad epidemics. However, no efficient strategies for BIM are proposed.

None of these previous works assesses the total influence of a seed set by considering only the number of nodes influenced during the spreading. Moreover, in this work we propose a flexible node cost model that strictly depends on the network structure. It

allows for an arbitrary range of values, without making strong assumptions on both the budget and the node cost.

1.2 Text Organization

The remainder of this text is organized as follows. In chapter 2 are presented the epidemic model, the node cost model, the budget model and the problem statement. Chapter 3 describes the different seeding strategies, including the node surround concept. Chapter 4 presents the different performance metrics—including diffusion power—, the network datasets, and reports the results of experiments performed over various scenarios. Finally, we conclude with a discussion in Chapter 5.

Capítulo 2

Models

The study of the BIM problem requires a clear definition of the nodes' costs model, the budget model, the epidemic model, the diffusion model, and the network model. In this section we specify the different models used in our study, some of which are first proposed here (cost model and budget model). As for the network, we will not consider any particular model but work directly with different real networks, as described in Section 4.2.

2.1 Epidemic and diffusion models

We adopt the classical SI model for network epidemics, widely used to represent diffusion of ideas or viruses [24]. In this model nodes may assume one of only two states: *susceptible* (S) or *infected* (I). Moreover, nodes can only transition from the S to the I state. Thus, infected nodes never return to the susceptible state.

Let $G = (V, E)$ be a network with node set V and undirected edge set E. Let $N(v)$ denote the set of neighbors of node $v \in V$ and $d(v) = |N(v)|$ the degree of node $v \in V$. We consider a discrete time model, indexed by $t = \{0, 1, 2, ...\}$. Let $S(t)$ and $I(t)$ denote the respective sets of susceptible and infected nodes at time t, such that $S(t) \cup I(t) = V$ and $S(t) \cap I(t) = \emptyset$.

A set of nodes is assumed to be infected at time zero, the seed set, denoted $I(0)$. The epidemic then unfolds on the network until some *quiescence* time t_q, corresponding to the first time t such that $I(t) = I(t+i), \forall i \in \mathbb{N}$. Thus, $I(t_q)$ denotes the set of infected nodes at the end of the epidemic.

State transitions, i.e., conditions under which a susceptible node becomes infected,

are based on the Linear Threshold Model with fixed and identical thresholds [10, 24], wherein the diffusion evolves deterministically. Let $\theta \in \mathbb{N}^*$ denote the fixed threshold, then a susceptible node becomes infected at $t+1$ if its number of infected neighbors at t is greater than or equal to θ. More formally:

$$\forall v \in V, v \in \begin{cases} I(t) & \text{if} \quad v \in I(t-1), \\ I(t) & \text{if} \quad |N(v) \cap I(t-1)| \geq \theta, \\ S(t) & \text{otherwise.} \end{cases} \tag{2.1}$$

2.2 Cost and budget models

A reasonable assumption regarding the cost of a node is to establish a cost dependency on some notion of node centrality. Intuitively, those more central nodes in a network (such as celebrities in Twitter) tend to be the more expensive ones, likewise. The centrality of a node, however, is typically context-dependent, and strongly relates to the phenomenon being studied. A number of metrics, such as *degree*, *closeness* and *PageRank* [24], have been proposed in order to capture how important (central) a node is under some perspective. While actual costs could depend on various metrics for centrality, and also on features not directly encoded in the network structure, this work considers the cost of a node v as being dependent on v's degree. Although a simplification, to consider a node central agreeing to its number of relationships is a quite natural and broadly employed approach for studying network epidemics. Among other examples, degree is related to *popularity* in online social networks, *influential articles* in a citation network, *promiscuity* in a sexual relationship network and *sociability* in a friendship network. Intuitively, the larger the degree of a node v, the more likely v is to either influence or be influenced. Indeed, empirical studies with real data have already pointed out that larger degree nodes are also the more active ones in viral marketing campaigns [12]. Therefore, the cost $c : V \to \mathbb{R}_+$ of a node $v \in V$ is given by:

$$c(v) = d(v)^{\alpha} \, , \tag{2.2}$$

where $d(v)$ is the degree of v and $\alpha \in \mathbb{R}_+$ is a fixed parameter that controls the dependence of the cost on the degree. Note that the unit-cost IM in our model corresponds to the particular case wherein $\alpha = 0$. Is noteworthy that α here controls the impact of $d(v)$ on the cost of v because degree centrality is the chosen proxy of importance.

We also define the cost of a set of nodes, which is simply the sum of the cost of each node in the set. As a useful convention, we consider the cost of an empty set as ∞ (infinity). Formally, given a set $W \subseteq V$ of nodes,

$$c(W) = \begin{cases} \infty & \text{if} \quad W = \emptyset, \\ \sum\limits_{\forall w \in W} d(w)^{\alpha} & \text{otherwise.} \end{cases} \tag{2.3}$$

We propose a budget model that depends on the network structure and node cost model. This allows for a more direct comparison among networks of different sizes and structures and also among different models for node cost. The idea for the budget is to pay for a given fraction of the network nodes at the cost of the average degree. Thus, let $\overline{d} = (\sum_{\forall v \in V} d(v))/n$ denote the average degree of a network $G = (V, E)$ of size $n = |V|$. Then, the budget b is defined as follows:

$$b = kn \cdot c(\overline{d}) = kn(\overline{d})^{\alpha} \ , \tag{2.4}$$

where α is the same applied to the node cost, and k is a fixed fraction. Since it clearly does not make sense to establish an initial budget incapable of performing any seeding, we define a lower limit for k such that b can always seed at least the cheapest node of the network. Thus, let $v_m \in V$ denote a node with the smallest cost. Then, the initial budget b is such that $b \geq c(v_m)$, which implies that $kn(\overline{d})^{\alpha} \geq d(v_m)^{\alpha}$, and hence $k \geq (1/n)(d(v_m)/\overline{d})^{\alpha}$. We therefore consider a fraction k such that $(1/n)(d(v_m)/\overline{d})^{\alpha} \leq k \leq 1$. Note that b is directly proportional to both the cost $(\overline{d})^{\alpha}$ of the (virtual) node with average degree and the size n of the network.

A brief discussion on the models for cost and budget is worth it at this point. Our node cost model (Eq. 2.2), based on the node degree, implies that the average node cost corresponds to $E[d(v)^{\alpha}]$. Note, however, that $\alpha \neq 1 \Rightarrow E[d(v)]^{\alpha} \neq E[d(v)^{\alpha}]$, i.e. the average degree cost does not correspond to the average node cost. Thus, a natural question is why not model the budget as a function of the expected cost rather than a function of the average degree. As previously discussed, in many different real situations there already exists a fair understanding on how central a node is; conversely, it is still an open question how these nodes are assigned costs. For instance, on [16] the value paid for a celebrity to tweet sponsored content weakly correlates with the number of followers.

Indeed, although degree has proven a reasonable proxy of *importance* in real cases [12], it may also not be that influential when assessing how valuable in terms of cost a node is [5]. The adoption of the parameter α is therefore consequence of such lack of accurate cost predictors. Yet, α allows the cost function to capture aspects observed in real cases, such as non-linear relation between centrality and cost, and nodes differing in their costs by orders of magnitude [26, 16, 6]. Furthermore, the budget modeling based on the average degree cost ($\overline{d}^{\alpha}$) also represents an advantage compared to the expected cost, as the latter requires the degree distribution of the network, while the former is immediately derived from one single parameter ($\overline{d}$).

A more realistic model for the budget (as well as node cost) is beyond our scope, as we turn our attention to seeding strategies that yield efficient epidemics within the given constraints by the budget and the cost. Moreover, the different seeding strategies (to be presented) receive the same initial budget, yielding a direct and fair comparison. Last, we will adopt very small values for k, such as $k = 10^{-4}$, leading to really limited budgets and hence stressing the importance of the seeding strategy. Clearly, seeding under large budgets loosen the demand for smart choices on the seeds.

Concerning the node cost model, its sole parameter $\alpha > 0$ plays a fundamental role in the trade-off between node centrality (degree) and cost. Intuitively, if the costs across the network present low variations (small α), then the seeding strategy can focus on the network structure and target those more central nodes. Conversely, if costs differ by many orders of magnitude (large α), these surely must be the strategies' primary concern. In what follows we make a rigorous argument for the role of α in the node cost model.

Proposition 2.2.1. *Given a network $G = (V, E)$, the proposed node cost model, the budget model with budget b, and a sufficiently large α, we have:*

1. *For every $v \in V$, if $d(v) > \overline{d}$ then $c(v) > b$;*
2. *For every $v \in V$, if $d(v) < \overline{d}$ then $b > c(v)$;*
3. *Let $U = \{v \in V | d(v) < \overline{d}\}$. Then $b > c(U)$.*

Proof. Note that the cost of a node $v \in V$ is given by $c(v) = d(v)^{\alpha}$ and the budget is given by $b = kn(\overline{d})^{\alpha}$, where $n = |V|$ and $0 < k \leq 1$. Thus, $c(v) > b \Leftrightarrow d(v)^{\alpha} > kn(\overline{d})^{\alpha} \Leftrightarrow \left(\frac{d(v)}{\overline{d}}\right)^{\alpha} > kn$. Since $\frac{d(v)}{\overline{d}} > 1$ when $d(v) > \overline{d}$, we have that for sufficiently large α the result holds. In particular, $\left(\frac{d(v)}{\overline{d}}\right)^{\alpha} > kn \Leftrightarrow \alpha > \frac{\ln(kn)}{\ln(\frac{d(v)}{\overline{d}})} \Leftrightarrow \alpha > \log_{\frac{d(v)}{\overline{d}}}(kn)$. The proof for case 2 follows along the same lines.

The proof for case 3 follows similarly, since $b > c(U) \Leftrightarrow kn(\overline{d})^{\alpha} > \sum_{v \in U} d(v)^{\alpha}$ (I). Note that $\sum_{v \in U} d(v)^{\alpha} < |U|(d^*)^{\alpha}$, where $d^* = \max_{v \in U} d(v)$. Thus, to accomplish condition (I) it is sufficient to provide that $kn(\overline{d})^{\alpha} > |U|(d^*)^{\alpha}$. Rewriting, it follows that $kn(\overline{d})^{\alpha} > |U|(d^*)^{\alpha} \Leftrightarrow \ln(kn)+\alpha\cdot\ln(\overline{d}) > \ln(|U|)+\alpha\cdot\ln(d^*) \Leftrightarrow \alpha\cdot\ln(\overline{d})-\alpha\cdot\ln(d^*) > \ln(|U|)-\ln(kn) \Leftrightarrow \alpha \cdot \ln(\overline{d}/d^*) > \ln(|U|/kn) \Leftrightarrow \alpha > \frac{\ln(|U|/kn)}{\ln(\overline{d}/d^*)}$, which means that $\alpha > \log_{\frac{\overline{d}}{d^*}}(|U|/kn)$ is sufficiently large to meet case 3.

□

Note that Proposition 2.2.1 states that the budget will never be large enough to pay for nodes with a degree higher than average but at the same time will always be sufficient to pay for nodes with degree smaller than the average, for a sufficiently large α. Thus, increasing α increases likewise the influence of node costs over the strategies' decisions.

It is also interesting to consider the degrees that can be included in the seed set, in the sense that there is enough budget to pay for them. This is stated in the following:

Proposition 2.2.2. *Given a network $G = (V, E)$, the proposed node cost model and the budget model with budget b, it holds that $c(v) \leq b \Leftrightarrow d(v) \leq \overline{d}(kn)^{1/\alpha}, \forall v \in V$.*

Proof. Note that $c(v) \leq b \Leftrightarrow d(v)^{\alpha} \leq kn(\overline{d})^{\alpha}$, which in turn means that $d(v) \leq \overline{d}(kn)^{1/\alpha}$.

□

Proposition 2.2.2 determines what degrees can be included in the seed set. To illustrate, lets consider a real example (*Hep-Ph* network, to be presented later), where $n = 12008$, $\overline{d} = 19.735$, $k = 5 \times 10^{-4}$ and $\alpha = 2$. In this case, if $d(v) \leq 48$ then $c(v) < b$. Thus, only nodes with degree less than or equal to 48 can possibly enter the seed set.

This next result shows the other direction: nodes costs become relatively identical with sufficiently small α.

Proposition 2.2.3. *For a given network $G = (V, E)$, as $\alpha \to 0$ the cost of a node v becomes, at the same proportion, a negligible factor for a seeding strategy to decide whether to seed $v, \forall v \in V$.*

Proof. Let $v_s, v_l \in V$ denote the nodes with the smallest and the largest degrees of the network, respectively. Then $c(v_s) \leq c(v) \leq c(v_l), \forall v \in V$, i.e. $d(v_s)^{\alpha} \leq d(v)^{\alpha} \leq d(v_l)^{\alpha}$. Note that $\forall v \in V, \alpha \to 0 \Rightarrow d(v)^{\alpha} \to 1$ and therefore $d(v_l)^{\alpha} - d(v_s)^{\alpha} \to 0$, which means that $c(v_l) - c(v_s) \to 0$, i.e. all nodes tend to have the same cost, thus allowing the seeding

to be more and more driven by node centrality (which in our model corresponds to node degree). □

It is worth noting that the particular case wherein $\alpha = 0$ implies $|I(0)| = \lfloor kn \rfloor$. Indeed, $\alpha = 0 \Rightarrow d(v)^\alpha = 1, \forall v \in V$, what in turn means that $c(v) = 1, \forall v \in V$, i.e. all nodes have unitary cost. Moreover, for a budget $b = kn(\ \overline{d}\ )^\alpha$ it holds $\alpha = 0 \Rightarrow kn(\ \overline{d}\ )^\alpha = kn$. Trivially, $b = kn$ covers the unitary cost of $\lfloor kn \rfloor$ nodes.

2.3 Problem Statement

In order to measure the effectiveness of a given seed set $I(0) \subset V$, we consider the following metric $\sigma : V \to \mathbb{N}$:

$$\sigma(I(0)) = |I(t_q)| - |I(0)|, \tag{2.5}$$

where t_q is the quiescence time (smallest time at which no other node becomes infected). Note that this metric captures the ability of the seed set to induce an epidemic. In particular, it distinguishes seed nodes (paid influence) from nodes infected along the epidemic (real benefit).

Thus, given a network $G = (V, E)$, the node cost model and the budget model with budget b, the goal is to determine a seed set $I(0) \subset V$ that maximizes $\sigma(I(0))$ such that $c(I(0)) \leq b$. Namely, the seed set $I(0)$ that maximizes the performance as measured by $\sigma(\cdot)$.

Note that this problem is at least as hard as unit-cost IM, which has been shown to be NP-hard [13]. As discussed in Section 1.1, a common approach to tackle epidemic seeding is to show that the objective function ($\sigma(\cdot)$ in this case) is submodular and then propose a polynomial time greedy algorithm that approximates the optimal solution to a constant factor. Unfortunately, $\sigma(\cdot)$ in our case is not submodular, in particular because the linear threshold model does not yield submodular objective functions [13]. Moreover, we are interested in the "diffusion power" of the seed set, and not just the total number of infected nodes. Thus, we resort to another approach, namely to design computationally efficient seeding strategies based on heuristics tailored to the problem at hand, where nodes have variable costs and diffusion follows the linear threshold model.

Capítulo 3

Seeding Strategies

A common approach in the design of seeding strategies for the IM problem is to rank nodes according to some criteria and then consider them sequentially for inclusion in the seed set. We follow this framework, but we divide this process in two steps: (i) nodes are ranked and considered in sequence; (ii) a node is considered to be placed in the seed set, ignored, or *surrounded* (as detailed later). Note that (i) determines a global ordering of the nodes, taking into account their position in the network, but not their cost, while (ii) takes into account the cost and attempts to buy "more for less", i.e. it leverages the importance of node $v \in S(0)$ by not seeding v directly, but its cheapest neighbors to ensure that $v \in I(1)$.

In the following, we first describe the two adopted node rankings. We then formalize the node surrounding process in Sections 3.3 and 3.4, wherein we define the *surrounding set* and the *extended surrounding set*, respectively. Next, we present in Section 3.5 both the seeding policies and the seeding strategies (i.e. "ranking-policy" combinations) that will be evaluated in Chapter 4. Finally, in Section 3.6 we present a cost-aware node ranking metric that considers centrality/cost-to-surround ratio. Benefits and drawbacks of these *cost-weighted rankings* are shown in Chapter 4.

3.1 Degree centrality ranking

As discussed in Chapter 2, degree may not be the best feature to consider when estimating nodes' importance on spreading processes [5]. Under the linear threshold model with fixed thresholds such a fact is perhaps easier to visualize due to its diffusion dynamics, wherein the sole condition for a node v to become infected at time t is the number of neighbors already infected at time $t-1$. The node degree therefore acts like a valve: if v has less

than θ neighbors then it will never become infected unless v itself is a seed. Nevertheless, as also discussed in Chapter 2, it is also unclear what are the main node features that yield large epidemics, thus making node degree a simple yet reasonable alternative. Moreover, degree centrality as a measure of node importance presents three major benefits. First, a common property to a wide variety of real networks, irrespective of their nature, is a heavy-tailed degree distribution, meaning that nodes with degrees orders of magnitude larger than the average occur with non-negligible probability. Although this condition may impose a constraint for seeding (since the cost of such nodes may be prohibitive), it also means that frequently there exists a set of a few nodes that, together, are neighbors of a large fraction of the network. The *surrounding set* approach thus becomes specially attractive since important, yet cost-prohibitive, nodes will surely get infected at time $t = 1$, if conveniently *surrounded* by seeds. Second, on real networks, the set of neighbors of the largest degree nodes is mostly formed by nodes with small degrees. This observation is captured by the notion of *degree assortativity* [23], i.e. the extent to which neighbors in a network are similar with respect to their degree. Finally, large degree nodes naturally tend to have numerous neighbors in common, meaning that part of the nodes used to surround a large degree node may also contribute to surround other large degree nodes, allowing them to be surrounded by using less new seed nodes. Therefore, one of the node rankings adopted is *degree centrality*, wherein network nodes are sorted with respect to their degree, from the largest to the smallest.

3.2 Triangle centrality ranking

When considering the very condition that allows an influence to propagate upon a network under the Linear Threshold Model with fixed thresholds, it is immediate to notice that, at each time step t, it is necessary that at least one susceptible neighbor of some infected node has at least $\theta - 1$ other infected neighbors to become infected at $t+1$. This highlights the importance of *triangles* on the network, i.e. subsets of three nodes connected to each other or, more formally, cliques of size 3. Indeed, a node is more likely to adopt the behavior of its neighbors when these are also neighbors from each other [3], an interesting phenomenon already referred to as a "gravitational force" [15] a node's neighbors exert on it. We therefore propose a node ranking based on triangles, defined as follows. First, we compute for every node v_i the number τ_i' of triangles that have v_i as a vertex. Formally, let $N(v_i)$ be the set of neighbors of v_i and $\mathbb{1}(\cdot)$ the indicator function. Then:

$$\tau_i' = \begin{cases} 0 & \text{if} \quad |N(v)| < 2, \\ \sum\limits_{\forall u_j, u_k \in N(v_i), j<k} \mathbb{1}\left(u_k \in N(u_j)\right) & \text{otherwise.} \end{cases} \tag{3.1}$$

Next, we determine τ_i'', which is the sum of τ_j' from every neighbor v_j of v_i, plus τ_i':

$$\tau_i'' = \tau_i' + \sum_{\forall u_j \in N(v_i)} \tau_j'. \tag{3.2}$$

The triangle centrality of node v_i is given by τ_i'', and larger is more central. Note that the definition of τ_i'' implies that every triangle of τ_i' is being counted three times, while triangles common to two of v_i's neighbors but not to v_i itself are being counted twice. Thus, τ_i'' reflects a higher appreciation of v_i's own triangles, a desirable feature since at each step of the diffusion, the sole condition for a new infection to occur relates only with those susceptible nodes directly neighboring the infected ones.

The second node ranking adopted, therefore, is *triangle centrality*, wherein the assessed importance of each node corresponds to its τ'' index. It is worth noting that such a score provides a very different information from that of the node's *clustering coefficient* (CC) [24], as the latter captures a *relative* measure—nodes with same CC may hugely differ with respect to their absolute number of triangles.

Hereafter, to distinguish the node rankings, we will denote degree centrality by V^d and triangle centrality by V^t.

3.3 The Surrounding Set

Intuitively, under different scenarios, cost-unaware strategies are more likely to exhaust the budget prior to achieving the minimum number of seeds needed to induce a large epidemic. A fundamental question is "how to obtain a relatively numerous set of seeds while leveraging those expensive, well-ranked ones?". To tackle this problem, we propose *surrounding sets*, defined as follows. Let θ denote the network epidemic threshold and $\gamma : S \to \mathbb{N}$ denote the "distance" of a node $v \in S(0)$ from $I(1)$, i.e. $\gamma(v)$ is the number of v's neighbors that still need to be seeded in order to ensure $v \in I(1)$, as follows:

$$\gamma(v) = \max(0, \theta - |N(v) \cap I(0)|). \tag{3.3}$$

A *surrounding set* of node v has size $\gamma(v)$ and is given by the set function $\Gamma : S \to S$, defined as

$$\Gamma(v) = \begin{cases} \emptyset & \text{if} \quad |N(v) \cap S(0)| < \gamma(v), \\ \arg\min(c(\{w_1, w_2, \cdots, w_{\gamma(v)}\}))\forall w \in \{N(v) \cap S(0)\} & \text{otherwise.} \end{cases} \tag{3.4}$$

Note that $\Gamma(v) \neq \emptyset$ means that $\Gamma(v)$ contains the cheapest $\gamma(v)$ susceptible neighbors of v. Also, note that if all nodes in $\Gamma(v)$ are seeded, v gets infected at $t = 1$. Thus, for each $v \in S(0)$, if $|N(v) \cap I(0)| \geq \theta$, v is said to be directly surrounded, since certainly $v \in I(1)$. Algorithm 1 describes the construction of the surrounding set.

Computational complexity: Prior to executing any of the later described algorithms, including Alg. 1, a pre-processing is made, only once, for each node in the network, such that its neighbors are sorted ascending by their costs. To sort the $d(v)$ neighbors of a single node v is $O(d(v) \log d(v))$. The overall computation for all nodes is therefore $O(\sum_i d(v_i) \log d(v_i))$. However, note that this operation depends on the number of neighbors of each node, what ultimately depends on the number $m = |E|$ of edges of the network. Indeed, every edge determines two neighbors, hence a total of $2m$ neighbors must be sorted. Thus, the pre-processing is $O(m \log m)$. Regarding Alg. 1, its complexity is dominated by iterating over $\gamma(v)$ neighbors of v (lines 7-13), which in the worst case corresponds to θ. Thus Alg. 1 is $O(\theta)$.

3.3.1 Tiebreakers

Tiebreak is an important matter during the above mentioned pre-processing stage of sorting each node's neighbors with respect to their cost. The reason is that nodes with identical costs are frequently different with respect to their neighborhoods. Every time the minimum cost can be achieved by more than one node, preference is given as follows. For V^d, preference is given to the node with the largest sum over the degree of its neighbors, and for V^t, the node with largest τ' score. Formally, let $\psi : S \times S \to S$ denote the function that receives a pair of susceptible nodes eligible for composing $\Gamma(v)$ and returns

the preferred node:

$$\psi(\{w_1, w_2\}) = \begin{cases} w_1 & \text{if } c(w_1) < c(w_2), \\ w_2 & \text{if } c(w_2) < c(w_1), \\ \arg\max_{w \in \{w_1, w_2\}} \sum_{\forall u \in N(w)} d(u) & \text{if } V^d, \\ \arg\max_{w \in \{w_1, w_2\}} \tau'_w & \text{if } V^t. \end{cases} \tag{3.5}$$

Note that to use the τ'' score as the tiebreaker for V^t would be less appropriate than τ', as τ'' carries information about a *region*, not a node. A high τ'' score means that the subgraph induced by v, v's neighbors, and v's neighbors' neighbors has a large number of triangles, which does not imply that τ' is also large (particularly, even $\tau' = 0$ is possible!). The τ'' score therefore is a good indicator of highly triangulated regions. However, it is better for a surrounding node to have itself a large number of triangles, as this increases the chance that this same node can also contribute to surround others, what in turn increases the chances of triggering the epidemic.

3.4 The Extended Surrounding Set

We now describe the extended surrounding set concept proposed in this work. The goal is to achieve a more effective budget usage by extending the surrounding set approach (described in Section 3.3) to the two-hop neighborhood of each node, as follows. Let $\Gamma(\cdot)$ be as described in Eq. 3.4. Now let $\rho : S \to S$ denote the function that receives a node $w \in S(0)$ and returns the cheaper set between $\{w\}$ and w's surrounding set $\Gamma(w)$ as follows

$$\rho(w) = \begin{cases} \{w\} & \text{if } \; c(w) < c(\Gamma(w)), \\ \Gamma(w) & \text{otherwise.} \end{cases} \tag{3.6}$$

Remembering that $\mathrm{c}(\emptyset) = \infty$ (as in Eq. 2.3) and $\gamma(v)$ is the number of seeds needed to surround v, the *extended surrounding set (ESS)* $\Gamma^+(v)$ of a node v is

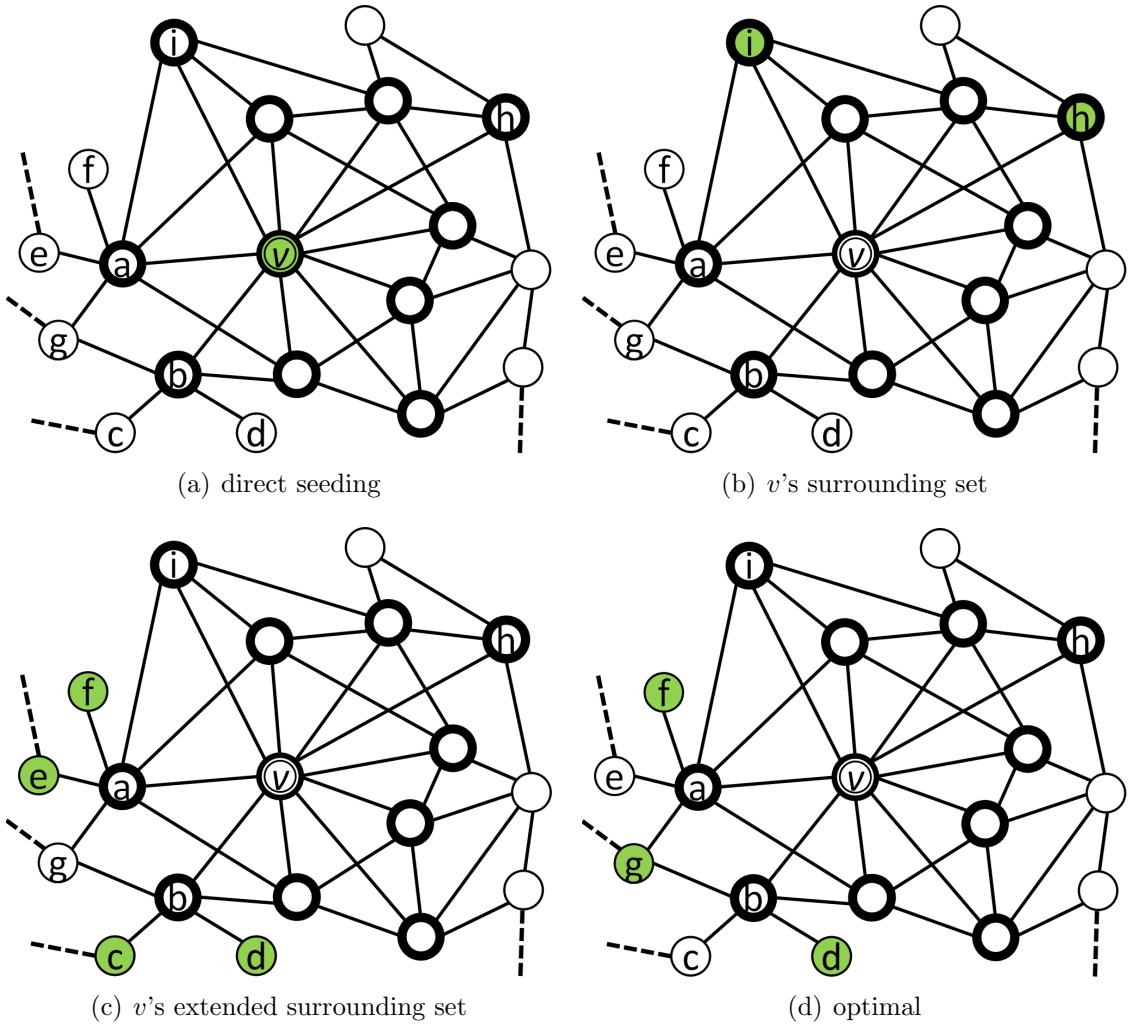
$$\Gamma^{+}(v) = \begin{cases} \emptyset & \text{if} \quad |N(v) \in S(0)| < \gamma(v), \\ \arg\min(c(\rho(w_1) \cup \rho(w_2) \cup \cdots & \\ \qquad \cup \rho(w_{\gamma(v)}))) \forall w \in \{N(v) \cap S(0)\} & \text{otherwise.} \end{cases} \tag{3.7}$$

It is immediate to notice that if $\Gamma^{+}(v)$ is seeded, v becomes infected, at most, when $t = 2$. Moreover, note that it is always the case that $c(\Gamma^{+}(v)) \leq c(\Gamma(v))$. Yet, $c(\Gamma^{+}(v))$ may not be optimal with respect to cost minimization in the two-hop neighborhood of v. Since the formation of the surrounding set $\Gamma(w)$ of each $w \in N(v)$ does not predict which other nodes are also going to be surrounded, it is not clear whether some 2^{nd} hop neighbors of v, once seeded, would contribute in surrounding more than one 1^{st} hop neighbor w of v. To illustrate such a fact, let us consider the hypothetical network of Figure 3.1. Consider its epidemic threshold is $\theta = 2$. Also, assume that the cost of a node is linearly proportional to its importance (degree), i.e., the cost function $c(v) = d(v)^{\alpha}$ is such that $\alpha = 1$. Finally, consider a seeding strategy that aims at leveraging the importance of the most central node, labeled v. For better visualization, borders of the 1^{st} hop neighbors of v are thicker.

Starting by Figure 3.1(a) and supposing arbitrary units of cost, note that to seed v directly would cost 10. Now, by observing Figure 3.1(b), note that to surround v with its $\gamma(v) = 2$ cheapest direct neighbors (surrounding set) would reduce such a cost to 8, since $\Gamma(v) = \{i, h\}$ and $c(\Gamma(v)) = c(\{i, h\}) = d(i)^{\alpha} + d(h)^{\alpha} = 8$. When adopting the ESS for this purpose, however (Figure 3.1(c)), the overall cost drops to 6 since $\Gamma^{+}(v) = \{\Gamma(a), \Gamma(b)\} = \{e, f, c, d\}$ and $c(\{e, f, c, d\}) = 6$. Yet, $c(\Gamma^{+}(v))$ is not the minimum achievable from v's two-hop neighborhood. Indeed, Figure 3.1(d) shows that further savings are still possible, as seeding $\{d, f, g\}$ would yield a total cost of 5 for having v infected (in this case, when $t = 2$).

Finally, Algorithm 2 describes the construction of the extended surrounding set (ESS), which is the actual approach employed in this work.

Computational complexity: Two scopes of Alg. 2 present an overall dominant complexity, as follows. To form an ESS it is necessary to firstly determine, for each neighbor w of v (line 8), w's surrounding set (line 10). Therefore, the $O(\theta)$-complex surrounding set computation is performed $d(v)$ times. Moreover, after such iterations, the array of size $d(v) = n$ containing the surrounding set of each of v's neighbors is then sorted ascending by cost (line 20), which is $O(n \log n)$. Thus, Alg. 2 is $O(n\theta + n \log n) = O(n(\theta + \log n))$.



(a) direct seeding (b) v's surrounding set

(c) v's extended surrounding set (d) optimal

Figure 3.1: Comparison of methods for leveraging the importance of a central node in a network. $\alpha = 1.0$ and $\theta = 2$.

Algorithm 1 SurroundingSet.

Require: v //{Input. v = node to be surrounded}
Require: Γ //{Output. The surrounding set Γ}
1: $\Gamma \leftarrow \emptyset$
2: **se** $v \notin S(0)$ **or** $d(v) < \gamma(v)$ **então**
3: RETURN Γ
4: **fim se**
5: $total \leftarrow 0$
6: $i \leftarrow 0$ //{Accesses the element $N^s(v)_i$ of the set $N^s(v)$ of v's neighbors, sorted ascending by cost and descending by the proper tiebreak criterion. }
7: **enquanto** $total < \gamma(v)$ **faça**
8: **se** $N^s(v)_i \in S(0)$ **então**
9: $\Gamma \leftarrow \Gamma \cup N^s(v)_i$
10: $total \leftarrow total + 1$
11: **fim se**
12: $i \leftarrow i + 1$
13: **fim enquanto**
14: RETURN Γ

Algorithm 2 ExtendedSurroundingSet.

Require: v, //{Input. v = the node to be surrounded}
Require: Γ^+ //{Output. The set Γ^+ of eligible nodes (no seeding performed)}
$\Gamma^+ \leftarrow \emptyset$
se $v \notin S(0)$ **or** $d(v) < \gamma(v)$ **então**
  RETURN Γ^+
fim se
$array\Gamma \leftarrow array[d(v)]$ //{Creates an array of size $|N(v)|$. Each position is a set.}
// At this point it is guaranteed v can be surrounded.
// We thus determine Γ for each neighbor of v, in order to determine Γ^+.
para $i = 1, \cdots, d(v)$ **faça**
  se $N(v)_i \in S(0)$ **então**
    $\Gamma_i \leftarrow surroundingSet(N(v)_i)$
    se $c(N(v)_i) < c(\Gamma_i)$ **então**
      $array\Gamma[i] \leftarrow \{N(v)_i\}$
    senão
      $array\Gamma[i] \leftarrow \Gamma_i$
    fim se
  senão
    $array\Gamma[i] \leftarrow \emptyset$
  fim se
fim para
$sortCostAscending(array\Gamma)$
para $i = 1, \cdots, \gamma(v)$ **faça**
  $\Gamma^+ \leftarrow \Gamma^+ \cup array\Gamma[i]$
fim para
RETURN Γ^+

3.5 Seeding Policies and Seeding Strategies

In the following, we describe the two different seeding policies adopted. Each seeding strategy further evaluated corresponds to the combination of a node ranking and a seeding policy.

- *Node surround (NS)*: This policy tries to surround each visited node, simply skipping those for which such task is impossible. First, it determines the ESS $\Gamma^+_{v_1}$ of v_1—the first node of a given node ranking—, seeding it in case $b > c(\Gamma^+_{v_1})$. It then evaluates v_2 in the same way, trying to surround it. Every time a node v_i cannot be surrounded (either because $|N(v_i) \cap S(0)| < \gamma(v_i)$ or $c(\Gamma^+_{v_i}) > b$) it is skipped and the ranking traversing continues. If the ranking is completely traversed prior to the exhaustion of the effective budget, then it is traversed again, but this time the NS policy tries to seed each node directly until the budget is no longer effective. When the budget becomes residual the process stops and the current set $I(0)$ of seeds is regarded as complete. Two seeding strategies derive from the NS policy, one for each node ranking. Indeed, we denote NS-D and NS-T the seeding strategies which combine NS with V^d and V^t, respectively. Finally, Algorithm 4 describes both strategies.

Computational complexity: Prior to analyze Alg. 4, we need to first determine the computation required to try to seed a given number of susceptible nodes, as described in Alg. 3. The iteration over the set S' of size $|S'| = n$ (line 2) dominates the overall complexity. Thus, Alg. 3 is $O(n)$. Considering now Alg. 4, its complexity is dominated by iterating, in the worst case, over all n network nodes (lines 4-11), in order to determine their corresponding ESS Γ^+ (line 7) and then try to seed those nodes in Γ^+ (line 8). Notice that, because of its definition, the largest size possible for Γ^+ occurs when it is formed by θ surrounding sets of size θ each. Therefore, in the worst case, Alg. 3 will try to seed $|\Gamma^+| = \theta^2$ nodes. That considered, Alg. 4 main complexity thus arises from n formations of the $O(n(\theta + \log n))$-complex ESS summed to n trials of seeding θ^2 nodes. Thus, Alg. 4 is $O(n(n(\theta + \log n) + \theta^2))$, which is $O(n\theta(n + \theta))$.

- *Cheapest nodes (CH)*: This simple policy is combined only with the degree ranking V^d, but in ascending order, thus starting from the smallest degree of the network. For each visited node v_i, it attempts to seed it directly. Note that the number of nodes this strategy seeds is the largest possible for a budget b. The budget here surely becomes residual after one single ranking traversal. The moment when the budget b is no longer effective, the process stops and the current set $I(0)$ of seeds is regarded as complete. Because of its simplicity, we have omitted its related algorithm. Finally, for its uniqueness we will also refer to the seeding strategy formed by combining CH with V^d simply as CH.

  Computational complexity: In the worst case, the budget is large enough for the strategy to keep its network seeding until having ultimately seeded all $n = |V|$ nodes of the network. Since CH always tries to seed one node at a time, the complexity of Alg. 3 here is $O(1)$. Thus, CH is $O(n)$.

It is worth to mention that other two seeding policies were also considered along our investigations, namely "Ranking Preserver" (RP) and "Cost Moderator" (CM). Both represent variations towards the NS policy, as follows. RP tries always not to skip an evaluated node in the case it cannot be surrounded, trying then to seed such node directly prior to evaluate another worse-ranked node. CM in turn evaluates, for each node v, how much the cost of having v surrounded by k other nodes will be. It then compares this cost to a ceiling that corresponds to have v surrounded by k virtual nodes whose degree is $\overline{d}$. CM then surrounds v if $c(v)$ is less than the ceiling, skipping v otherwise. Such a ceiling is based on the fact that the budget also derives from $\overline{d}$.

None of these policies, however, presented results fundamentally distinct to the point of being included here. Indeed, RP has proven not to be a good policy, as no single of its results presented superior performance compared to the others, apart from CH. More than that: RP presented, for almost all considered scenarios, the worst performance among those policies that surround nodes. CM, on the other hand, had punctually presented the best results, but by "best" we mean a very small edge compared to NS. For the vast majority of scenarios, however, it performed either as well as NS or worse than the latter, revealing that it is not that consistent. For these reasons, and also to provide a cleaner set of results with fewer curves in the plots, in this work we present strategies based only on CH and NS. Ultimately, the latter is the core idea throughout the experiments.

Algorithm 3 tryToSeed.

Require: $S' \in S(0)$ //{Input. The set of nodes to be seeded.}
Require: $S(0)$, $I(0)$, b //{Output. Updated $S(0)$, $I(0)$ and b}
se $b \geq c(S')$ **então**
  para todo $v \in S'$ **faça**
    $S(0) \leftarrow S(0) \setminus \{v\}$
    $I(0) \leftarrow I(0) \cup \{v\}$
    $b \leftarrow b - c(v)$
  fim para
fim se

Algorithm 4 Strategy NS (Node Surround).

Require: $G = (V^r, E)$, δ, θ, b //{Input. δ = minimum cost; V^r is either V^d or V^t}
$S(0) \leftarrow V^r$ //{Set of susceptible nodes}
$I(0) \leftarrow \emptyset$ //{Set of infected nodes}
$i \leftarrow 1$
enquanto $b \geq \delta$ **and** $i \leq |V^r|$ **faça**
  $v \leftarrow V_i^r$ //{Next node in the ranking}
  se $v \in S(0)$ **então**
    $\Gamma^+ \leftarrow extendedSurroundingSet(v)$
    $tryToSeed(\Gamma^+, S(0), I(0), b)$ //{Algorithm 3}
  fim se
  $i \leftarrow i + 1$
fim enquanto
$i \leftarrow 1$ //{Ranking is traversed again to exhaust the remaining effective budget}
enquanto $b \geq \delta$ **and** $i \leq |V^r|$ **faça**
  $v \leftarrow V_i^r$
  $tryToSeed(\{v\}, S(0), I(0), b)$
  $i \leftarrow i + 1$
fim enquanto
RETURN $I(0)$

3.6 Cost-weighted rankings

The seeding performed by the strategies above described is solely based on local information, i.e. the policy that guides whether to seed or surround a node v considers uniquely

whether the cost of such tasks can be covered by the budget. A clear downside of such approach is that the moment the strategy evaluates v it may opt to surround it and, right after, visit a node that, although slightly less central, is likewise much less costly for being surrounded. Intuitively, if the strategies could always firstly surround those nodes with better centrality/cost-to-surround ratio, the budget would be used more effectively, allowing further seeding and hence inducing larger epidemics.

We thus introduce *cost-weighted rankings*. Their goal is to deliver to NS policy a node ranking that already embeds a global relation between nodes with respect to both their centrality and their cost for being surrounded. Cost-weighted rankings are defined as follows. Given a network $G = (V, E)$, each node $v \in V$ is visited prior to any seeding attempt. For each visited node v_i, its argued centrality (either $d(v)$ or τ_i'') is divided by the cost of having v_i surrounded (i.e. $d(v)/c(\Gamma^+_{v_i})$ or $\tau_i''/c(\Gamma^+_{v_i})$). Such a division yields a second score $\lambda_i \in \mathbb{R}_+$ for v_i, upon which V is sorted, yielding a ranking V_Γ. For every v_i such that $\Gamma^+_{v_i} = \emptyset$ (nodes that cannot be surrounded), its original score is divided by a constant $C = \theta \cdot c(d_L)$, where d_L is the largest degree of the network and hence $c(d_L)$ is the highest cost. Note that this provokes all of such nodes to become low ranked but still ordinary comparable.

The adoption of cost-weighted rankings may lead strategies' performances to either improve or degrade, depending on whether node costs are more or less dependent on node centrality, respectively. These behaviors are demonstrated and discussed in Section 4.3. Last, because no actual seeding is performed during V_Γ's formation, and also because $\Gamma^+(\cdot)$ is not cost-optimal (as shown in Section 3.4), note that V_Γ does not yield an optimal centrality/cost-to-surround relation among nodes either.

Hereafter, cost-weighted rankings will be referred to as V^d_Γ and V^t_Γ for degree-based and triangle-based rankings, respectively.

Capítulo 4

Evaluation

We start this section by presenting some performance metrics, in particular the concept of "diffusion power". We then present the different networks and parametrization used, followed by evaluation and main findings.

4.1 Performance metrics

Let $G = (V, E)$ denote the network with $n = |V|$ nodes, θ the epidemic threshold, $I(0)$ the seed set, and t_q the quiescence time. We consider the following metrics in order to evaluate the performance of the seeding strategies:

- *Cost per Seed (CS)*, $\overline{c} = \frac{c(I(0))}{|I(0)|}$.
- Fraction of infected nodes at time t, namely $|I(t)|/n$.
- *Diffusion Power (DP)*: fraction of nodes infected until t_q. discounting the fraction of seeds. Thus, the DP $\Delta : I \to \mathbb{R}_+$ of a set $I(0)$ of seeds is

$$\Delta(I(0)) = \frac{\sigma(I(0))}{n} = \frac{|I(t_q)| - |I(0)|}{n}, \tag{4.1}$$

- *Relative Diffusion Power (RDP)*: fraction of nodes with degree equal to or greater than θ infected until t_q, discounting the fraction of seeds. More formally, let $B = \{v \in V | d(v) < \theta\}$ denote the set of nodes whose degree is smaller than θ. Then, the RDP $\Delta^r : I \to \mathbb{R}_+$ of a set $I(0)$ is

$$\Delta^r(I(0)) = \frac{\sigma(I(0))}{n - |B|}. \tag{4.2}$$

- *Average Diffusion Power (ADP)* across all epidemic thresholds for a given G. More formally, let θ_1 and θ_M denote, respectively, the minimum and the maximum values of θ for a given network G that triggers an epidemic. Note that $\theta_1 = 2$ for all networks and θ_M assumed different values as indicated in Table 4.1. Yet, let $I(0, \theta_i)$ denote the seed set of a given strategy when $\theta = \theta_i$. Thus, the average diffusion power $\overline{\Delta}$ is given by:

$$\overline{\Delta} = \frac{\sum_{i=1}^{M} \Delta(I(0, \theta_i))}{\theta_M - 1}. \tag{4.3}$$

Notice that the three last metrics are based on the concept of "diffusion power". They will be instrumental later in Section 4.3 on clarifying the importance of such a concept.

4.2 Network datasets and evaluation scenarios

The proposed seeding strategies were evaluated empirically with various real, undirected networks using different scenarios (parametrization). Table 4.1 lists the networks considered in our evaluation and some basic statistics for each of them, but more information for those datasets may be found at [19]. Networks had their original names abbreviated and is refereed henceforward as follows: *Fb* = "ego-Facebook"; *hep* = "ca-HepPh"; *astro* = "ca-AstroPh"; *cmat* = "ca-CondMat"; *enron* = "email-Enron"; *Bk* = "loc-Brightkite" and *dblp* = "com-DBLP". For each network, we report the number $|V|$ of nodes, number $|E|$ of edges, Degree Assortativity (DA) [23], percentage of nodes in the largest connected component (%LCC), average degree $\overline{d}$, maximum degree d_M, and maximum epidemic threshold θ_M used during the experiments. It is worth noting that the minimum degree in all networks is 1.

Note that the networks have quite different structures, not only with respect to their sizes (orders of magnitude) but also their connectedness, expressed both by their average degree $\overline{d}$ and degree assortativity (DA). For instance, *astro* has less than 10% of *dblp*'s size, but also a 3 times larger average degree. The DA for *enron* reveals that such network is slightly disassortative, meaning that neighbors tend to have different degrees. Conversely,

hep is quite assortative, which indicates neighbors tend to have similar degrees. Also note that the average degree is more than 20 times smaller than the maximum degree for all networks, and in some cases more than 100 times smaller (Bk). Those networks, therefore, provide a structurally rich context to better assess the concept of extended surrounding sets and the proposed strategies in general.

Table 4.1: Network datasets and some basic statistics.

Network	$\|V\|$	$\|E\|$	DA	%LCC	$\overline{d}$	d_M	θ_M
astro	18772	198050	0.21	95.4%	21.10	504	15
cmat	23133	93439	0.13	92.3%	8.08	281	10
Fb	4039	88234	0.064	100%	43.69	1045	13
Bk	58228	214078	0.011	97.4%	7.35	1134	20
dblp	317080	1049867	0.27	100%	6.62	343	11
enron	36692	183832	-0.11	91.8%	10.02	1383	35
hep	12008	118489	0.63	93.3%	19.73	491	15

Besides the used networks, we also consider different scenarios by varying the parameters associated with the models. In particular, for the node cost ($c(v) = d(v)^\alpha$) we consider $\alpha = \{0.5, 1.0, 2.0\}$, which implies node costs have highly different dependence on node degree (sub-linear, linear and quadratic, respectively). The initial budget ($b = kn\overline{d}^\alpha$) will be based on $k = \{0.0005, 0.0002\}$. These values of k were chosen in order to intentionally restrict strategies to seed only a very small fraction of the network. In addition to the fact that it raises the importance of selecting seeds well (as discussed in Section 2.2), it also corresponds to a more realistic scenario, considering possible applications. Yet, such a condition is not guaranteed, as we later demonstrate. The range of epidemic threshold values adopted is $\theta = \{2, 3, 4, ..., \theta_M\}$ where θ_M varies according to the network investigated. It corresponds to the value to which none of the considered strategies manages to induce epidemics, considering the possible values of the other parameters. This per-network maximum threshold θ_M is described in Table 4.1.

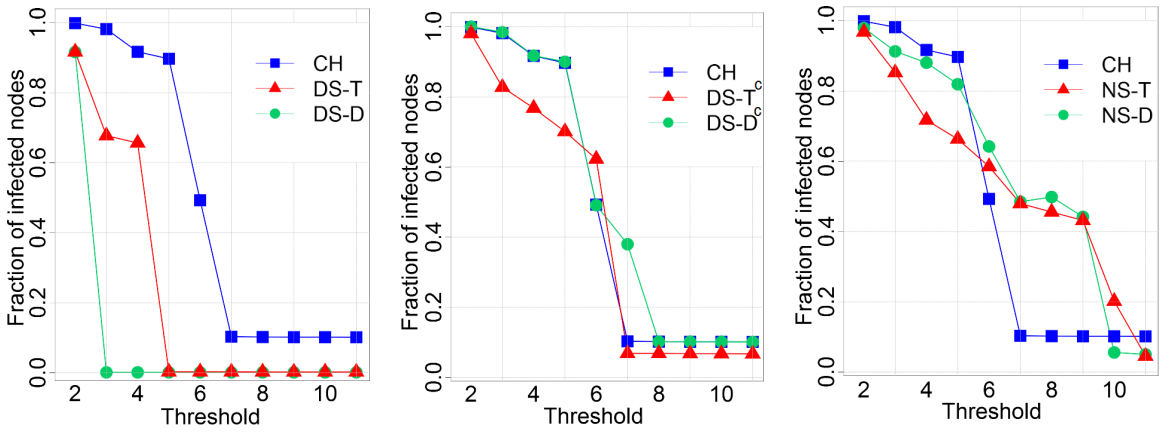
A *simulation scenario* corresponds to a network and the parameters k and α upon which the different strategies and θ values are tested. Note that, given any specific scenario (i.e. network, k and α), the decision of any considered seeding strategy is deterministic, and thus a single simulation run is necessary for the evaluation. The considered strategies are NS-D, NS-T and CH (as described in Section 3.5). Last, because we have extensively evaluated several combinations for the various parameters, a huge amount of results was generated. However, to better illustrate and highlight our main findings, we present here only a subset of all results.

4.3 Results

We start by reporting how cost aware strategies are superior but still limited when directly targeting nodes. In particular, we show that NS-based strategies outperform those based on centrality/cost ratio. Figure 4.1 shows the *fraction of infected nodes per threshold* at the quiescence time t_q for the networks *Fb* and *astro*. Every curve from the different plots relates to a particular seeding strategy. Each (connected) dot of a given curve denotes, for a network threshold θ (x axis), the corresponding fraction of infected nodes at t_q, i.e. $I(t_q)/|V|$ (y axis). The impact of a node degree on its cost is $\alpha = 2.0$ (Eq. 2.2) for *Fb* and $\alpha = 0.5$ for *astro*. We will denote by DS the illustrative policy shown in figures 4.1(a), 4.1(b), 4.1(d) and 4.1(e), which performs *direct seeding* of each node visited, budget permitting, skipping the node otherwise. The rankings associated to DS in figures 4.1(a) and 4.1(d) are V^d and V^t (as defined in sections 3.1 and 3.2), thus forming the strategies DS-D and DS-T, respectively. In figures 4.1(b) and 4.1(e), rankings are such that the ordinary position of each node v is given by its centrality/cost ratio, from largest to smallest, namely $d(v)/c(v)$ and $\tau''_v/c(v)$ for degree and triangle centrality, respectively. We will thus denote the corresponding strategies by DS-D^c and DS-T^c, respectively.

Comparing Figures 4.1(a) and 4.1(b), it becomes clear the importance of cost-awareness for BIM. Indeed, by neglecting nodes' costs (Figure 4.1(a)), DS-T's performance becomes already negligible for any $\theta > 4$; conversely, the pondering of nodes' centrality/cost ratio (Figure 4.1(b)) led DS-T^c to activate more than 60% of the network until $\theta = 6$. The difference becomes even larger between DS-D and DS-D^c strategies, as the former induce epidemics already negligible for every $\theta \geq 3$ (Figure 4.1(a)) whereas the latter still manages to cover almost 40% of the network when $\theta = 7$ (Figure 4.1(b)). Nevertheless, strategies' performances may still be significantly further improved: Figure 4.1(c) shows that broad epidemics are still attainable for considerably larger values of θ when strategies adopt the NS policy over V^d and V^t. Indeed, $\theta = 7$ is the maximum threshold to which DS-D^c still manages to induce broad epidemics, influencing almost 40% of the network (Figure 4.1(b)), whereas NS-D manages to influence more than 40% even when $\theta = 9$ (Figure 4.1(c)). Considering DS-T^c, note that its spreading is hindered by every $\theta > 6$ (Fig 4.1(b)), whereas NS-T still manages to influence 20% of the network when $\theta = 10$ (Figure 4.1(c)). Consider now the different plots for the *astro* network. Remember that the cost of a node here grows sub-linearly with the node degree ($\alpha = 0.5$)—a completely different cost regime than that of the just analyzed *Fb* network. Note that, interestingly, to seed nodes according to their centrality/cost ratio (Fig.4.1(e)) led DS-D^c

to perform worse than the cost-unaware DS-D strategy (Fig.4.1(d)). This illustrates the fact captured by Proposition 2.2.3: under diminishing values of α, the need of concerning about node costs also diminishes. Even then, to surround those more central nodes (Fig.4.1(f)) allowed, for instance, NS-T to induce broad epidemics to more than twice as many thresholds surpassed by DS-T^c (Fig.4.1(e)).



(a) central nodes first (b) best centrality/cost first (c) central nodes surrounded

(d) central nodes first (e) best centrality/cost first (f) central nodes surrounded

Figure 4.1: Fraction of infected nodes per threshold on the networks *Fb* (top; $\alpha = 2.0$) and *astro* (bottom; $\alpha = 0.5$) from cost-unaware strategies (left), cost-aware strategies that leverage nodes' centrality/cost ratio (center), and cost-aware strategies that surround central nodes (right). For both networks, $k = 0.0005$.

We now report how both the different node rankings and the different values of α adopted lead strategies to yield both different seed sets and different spreading dynamics. Thus, we show in Figure 4.2 the log-scaled *degree distribution of seeds* from each strategy on the networks *Fb* (top) and *hep* (bottom) for $\alpha = 0.5$ (left), 1.0 (center) and 2.0 (right). Budget is based on $k = 0.0005$ and the network threshold is $\theta = 6$ for *Fb* and $\theta = 8$ for *hep*. Note that at every single plot of Figure 4.2 apart from Figure 4.2(a) strategies have clearly seeded different nodes for a same scenario. Regarding NS-D and NS-T, the reason lies in their respective node rankings, as both strategies adopt the same policy (NS).

Moreover, it also becomes clear how the sole variation of α (each column of figures) also leads NS-strategies to different decisions on seeding, since all other parameters remain unchanged from one column to another. Focusing now on NS-T, it can be noticed that such strategy have generally seeded a larger fraction of nodes that are more expensive, compared to NS-D, as seen on all figures apart from Figure 4.2(a). Such a fact is a consequence of the ability of the triangle centrality (Eq. 3.2) of locating those regions of a network where every node tends to be well connected to many others, even those nodes that compose the surrounding sets. Indeed, later in the evaluation we demonstrate that NS-T tends to present higher *cost per seed*, meaning that its seed set tends to be smaller than NS-D's (since the budget is same).

It is noteworthy to stress that to increase the value of α provokes the budget ($b = kn(\overline{d})^{\alpha}$) to become larger as well. As a consequence, the observed variation on α (from 0.5 to 2.0) allows all strategies—specially CH—to seed much more nodes from the network. It even gets to the point of having all nodes of the smallest degrees from both networks to be seeded. Indeed, when $\alpha = 1.0$ on *Fb* (Figure 4.2(b)), almost all CH seeds have degree $d = 1$, and some few have degree $d = 2$, meaning that the budget was sufficient to cover the cost of all nodes whose degree is 1. When α is raised to 2.0 (Figure 4.2(c)), the cost of those nodes whose degree is close to the smallest grows much less than the budget. Thus, the fraction of seeds from CH that have degree $d = 1$ decreases although all nodes of such degree were seeded. The reason is that now all nodes of degree $d \leq 4$ were seeded, and budget still remained to seed some few nodes of degree $d = 5$. Curiously, these seeds from CH on Figure 4.2(c) also reveal that the *Fb* network have a larger number of nodes with degree 2, 3 and 4 than nodes with degree $d = 1$.

Interestingly, despite the different decisions taken by each strategy, their induced spreading may either be totally different or reasonably similar. To illustrate that, Figure 4.3 shows the *fraction of infected nodes over time* of the corresponding epidemics induced by each set of seeds from Figure 4.2. Thus, Figure 4.3(a) shows the diffusion produced by seeds from Figure 4.2(a), and so on.

Risings on the network threshold value impact strategies differently. Figure 4.4 shows the *fraction of infected nodes over time* for the networks *Cmat* and *HepPh*. Curves here are related to a specific threshold, thus allowing to visualize more than one epidemic at a time for the same strategy. Therefore, names for each curve additionally include their respective value of θ. Note that the strategy CH presents higher sensibility to θ. It transits from a large, long lasting epidemic to a short, negligible one long before a similar

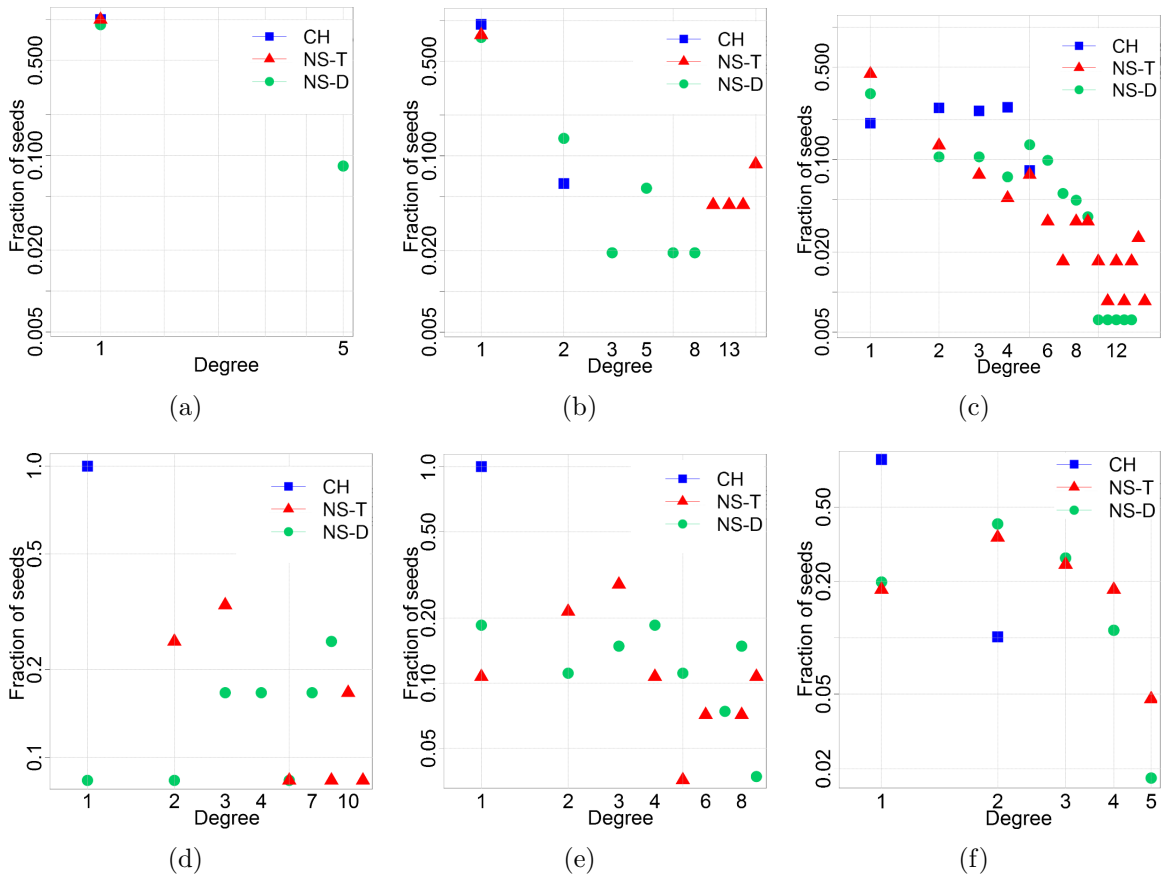


Figure 4.2: Degree distribution of seeds for the networks *Fb* (top) and *hep* (bottom) when $\alpha = 0.5$ (left), 1.0 (center), and 2.0 (right). $\theta = 6$ for *Fb* and $\theta = 8$ for *hep*.

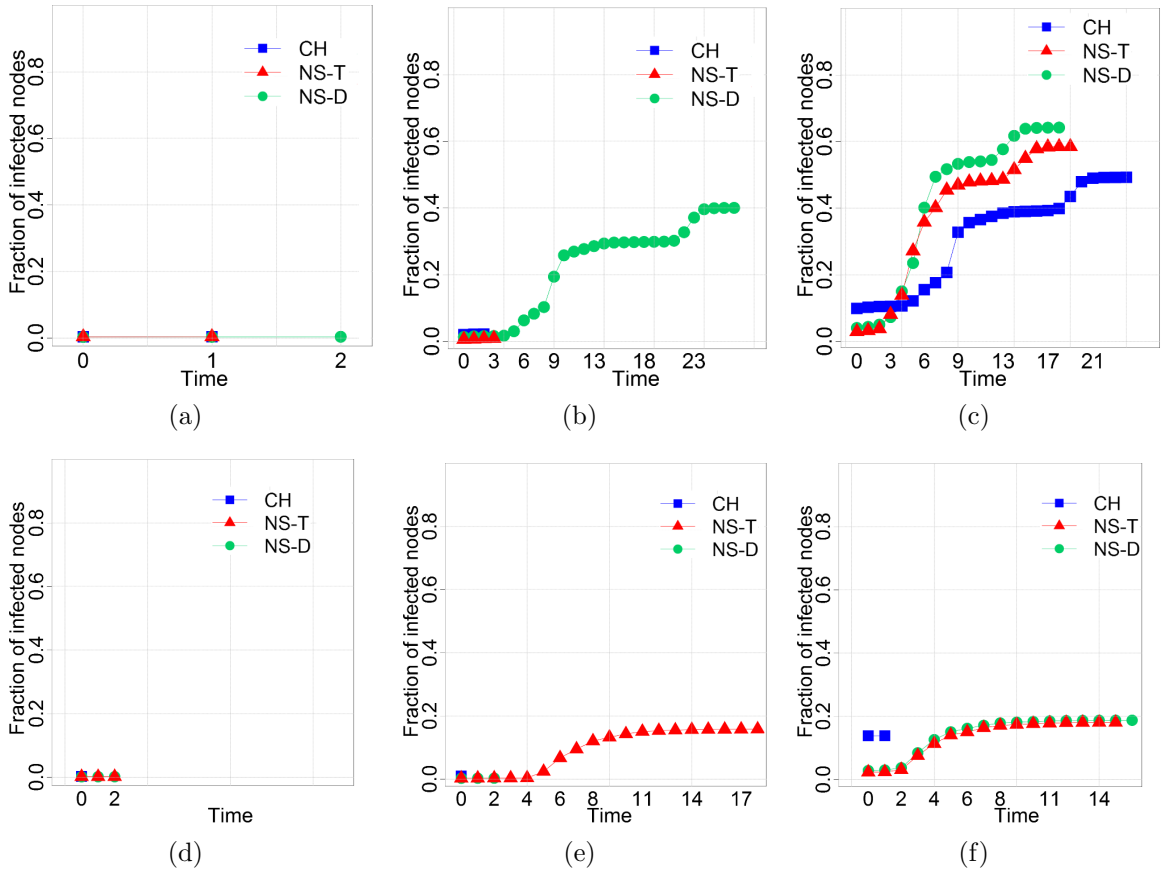
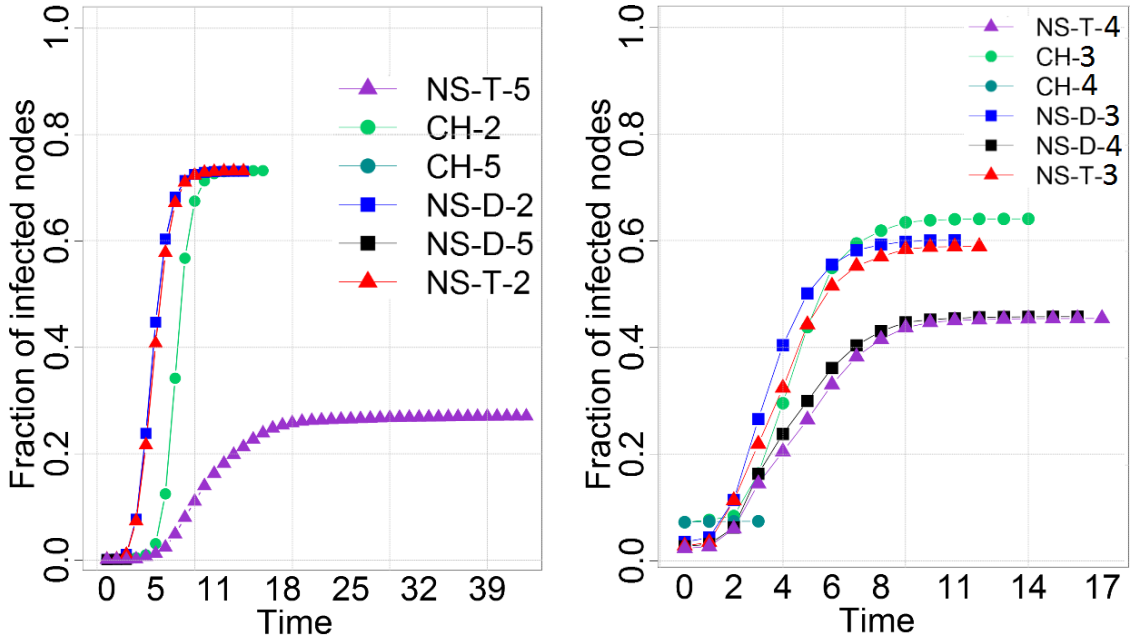


Figure 4.3: Fraction of infected nodes over time for the networks *Fb* (top) and *hep* (bottom), for the same parameters of Figure 4.2.

behavior can be observed on another strategy, for growing values of θ (Figure 4.4(b)). This behavior is not surprising as CH performs "blind" seeding, i.e. it does not leverage the network structure, and this in turn means that to merely focus on cost many times does not pay off, specially for larger values of θ. In particular, note that on Figure 4.4(b) CH-4 forms a set of seeds more than two times larger than NS-T-4 ($t = 0$) and even then it does not manage to induce a large spreading; CH-5 (Figure 4.4(a)) does not even come to start, thus being unnoticeable in the plot. Figure 4.4 also illustrates, again, (i) the trend of having NS-D showing slightly quicker and larger propagation compared to NS-T for smaller values of θ (Figure 4.4(b)); and (ii) the trend of having NS-T showing higher resilience for growing values of θ compared to NS-D. Indeed, this is captured in Figure 4.4(a) by a larger threshold gap (from $\theta = 2$ to $\theta = 5$), showing that while NS-D-5 influence is negligible, NS-T-5 epidemic still manages to slowly reach more than a quarter of the network.



(a) CMat (b) HepPh

Figure 4.4: *Fraction of infected nodes over time* for the networks *CMat* and *HepPh*, considering two different thresholds on each plot: $\theta = \{2, 5\}$ and $\theta = \{3, 4\}$, respectively. In both plots, $k = 0.0002$. For *Cmat*, $\alpha = 0.5$ and for *HepPh*, $\alpha = 2.0$.

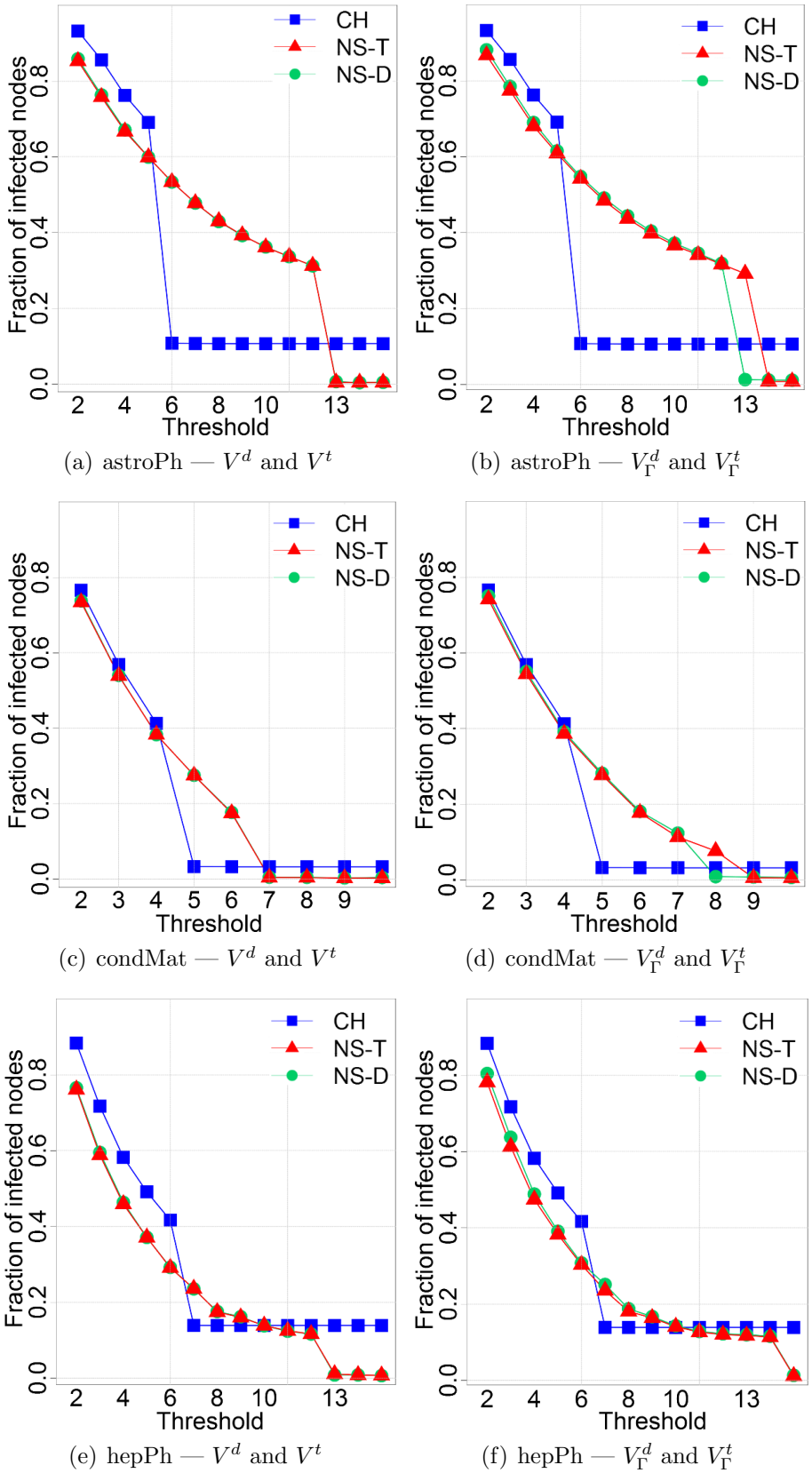
The cost-weighted ranking V_Γ (Section 3.6) may significantly increase the resilience of the different strategies to growing values of θ, i.e. the maximum θ under which the strategies still manage to induce large epidemics, thus being adopted in many scenarios. Figure 4.5 shows the *fraction of infected nodes per threshold* from epidemics simulated over the networks *astro*, *cmat* and *hep*, when $k = 0.0005$. Costs are based on $\alpha = 2.0$,

hence node degree enormously impacts on the cost. Note that to use V_Γ led almost all strategies to induce epidemics for greater values of θ. Indeed, as mentioned in Section 3.6, Interestingly, our results also demonstrate that the use of V_Γ when $\alpha \leq 1$ does not necessarily increases strategies' resilience, and may even reduce it. Figure 4.6 shows the *fraction of infected nodes per threshold* from epidemics simulated over the networks *cmat* and *Bk*, using the regular V^d and V^t rankings (figures 4.6(a) and 4.6(c)) and their cost-weighted versions V^d_Γ and V^t_Γ, respectively (figures 4.6(b) and 4.6(d))when $k = 0.0005$. Here, $\alpha = 0.5$, meaning that the cost variation across the network is much lower, compared to Figure 4.5. Note that V^t_Γ does not deliver any visible improvements on NS-D's epidemics. Furthermore, considering NS-T, the adoption of V^t_Γ diminished its resilience to epidemic thresholds, as this strategy no longer manages to induce non-negligible epidemics when $\theta = 5$ for *cmat* (Figure 4.6(b)) and $\theta = 10$ for *Bk* (Figure 4.6(d)). Figures 4.5 and 4.6 therefore illustrate what Propositions 2.2.1 and 2.2.3 capture: how the cost gradually becomes either a predominant or negligible feature when deciding whether to seed a node, as the node costs tend to differ or equalize among each other, respectively.

Plots presented up to this point did not consider two important issues:

- despite using the same k for the various networks favors their comparison, these largely vary in both the size $n = |V|$ (orders of magnitude) and the average degree $\overline{d}$. Consequently, it may be the case that a strategy, upon receiving the budget $(kn(\overline{d})^\alpha)$, ends up forming a set $I(0)$ with non-negligible size, specially strategy CH. As already mentioned, the unit-cost IM problem does not distinguish from the final set of activated nodes, those which were the initial spreaders. This is not a concern in there, actually, since its $|I(0)|$ is the same irrespective of the seeding strategy. In BIM problems, however, $|I(0)|$ not only varies but may assume huge sizes. It would not be reasonable, therefore, to consider this potentially numerous amount of paid influence (seeds) when assessing a strategy's performance. Thus, to properly capture how influential $I(0)$ is, we exclude $|I(0)|$ from $|I(t_q)|$, in order to obtain $I(0)$'s *diffusion power* (Eq. 4.1). Note that the division by n (network size) allows to compare different networks.

- for each value of θ, $\forall v \in S(0)$, $d(v) < \theta \Rightarrow v \in S(t_q)$, which means that although these plots show how many nodes were infected, they provide no indication on the upper limit of how many *could have been* infected. To determine this precise value is a complex task since it depends not only on $d(v)$ (numerical condition) but also on whether the propagation unfolding will eventually reach v (structural condition).



(a) astroPh — V^d and V^t

(b) astroPh — V_Γ^d and V_Γ^t

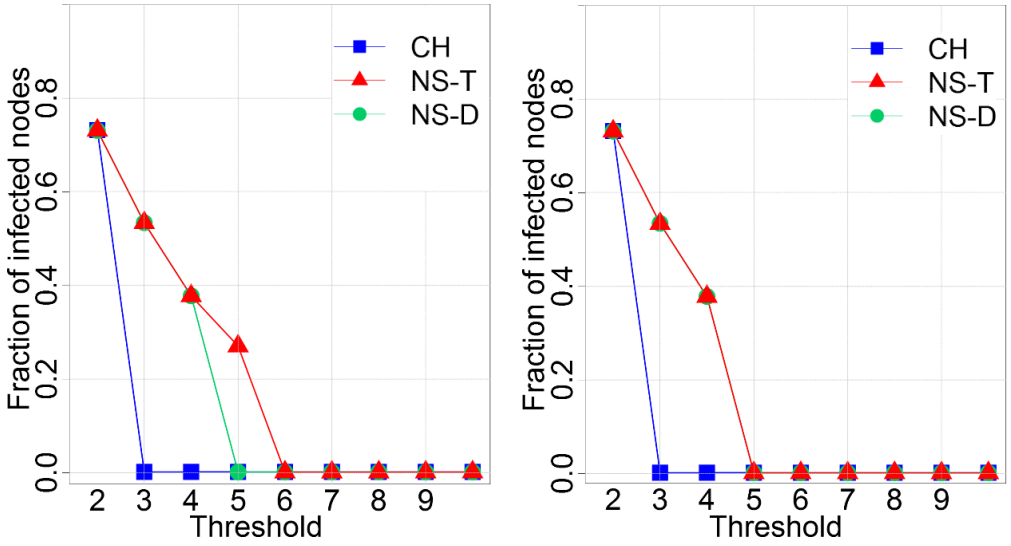
(c) condMat — V^d and V^t

(d) condMat — V_Γ^d and V_Γ^t

(e) hepPh — V^d and V^t

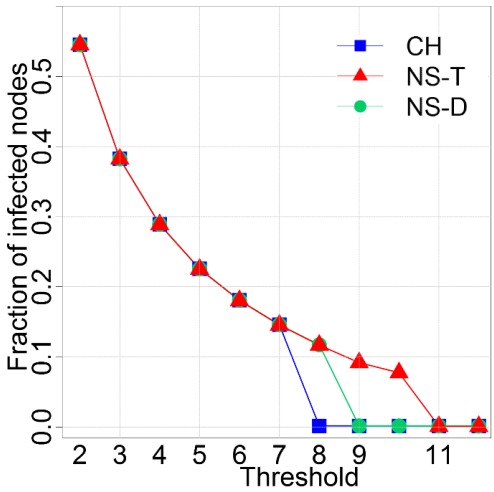
(f) hepPh — V_Γ^d and V_Γ^t

Figure 4.5: *Comparison of fraction of infected nodes per threshold for the networks astro, cmat and hep, when strategies are given regular rankings (left) and cost weighted rankings (right), for* $k = 0.0005$ *and* $\alpha = 2.0$.

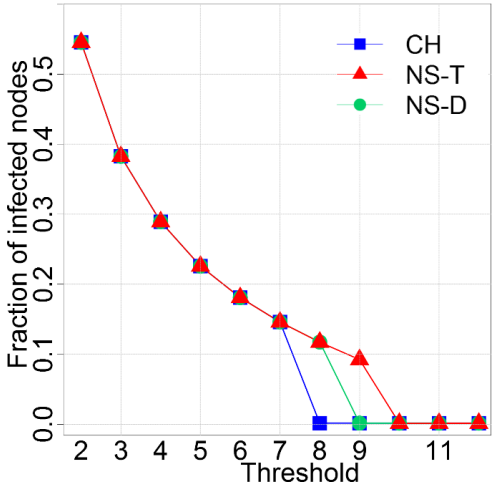


(a) $\alpha = 0.5$ — V^d and V^t

(b) $\alpha = 0.5$ — V_Γ^d and V_Γ^t



(c) $\alpha = 0.5$ — V^d and V^t



(d) $\alpha = 0.5$ — V_Γ^d and V_Γ^t

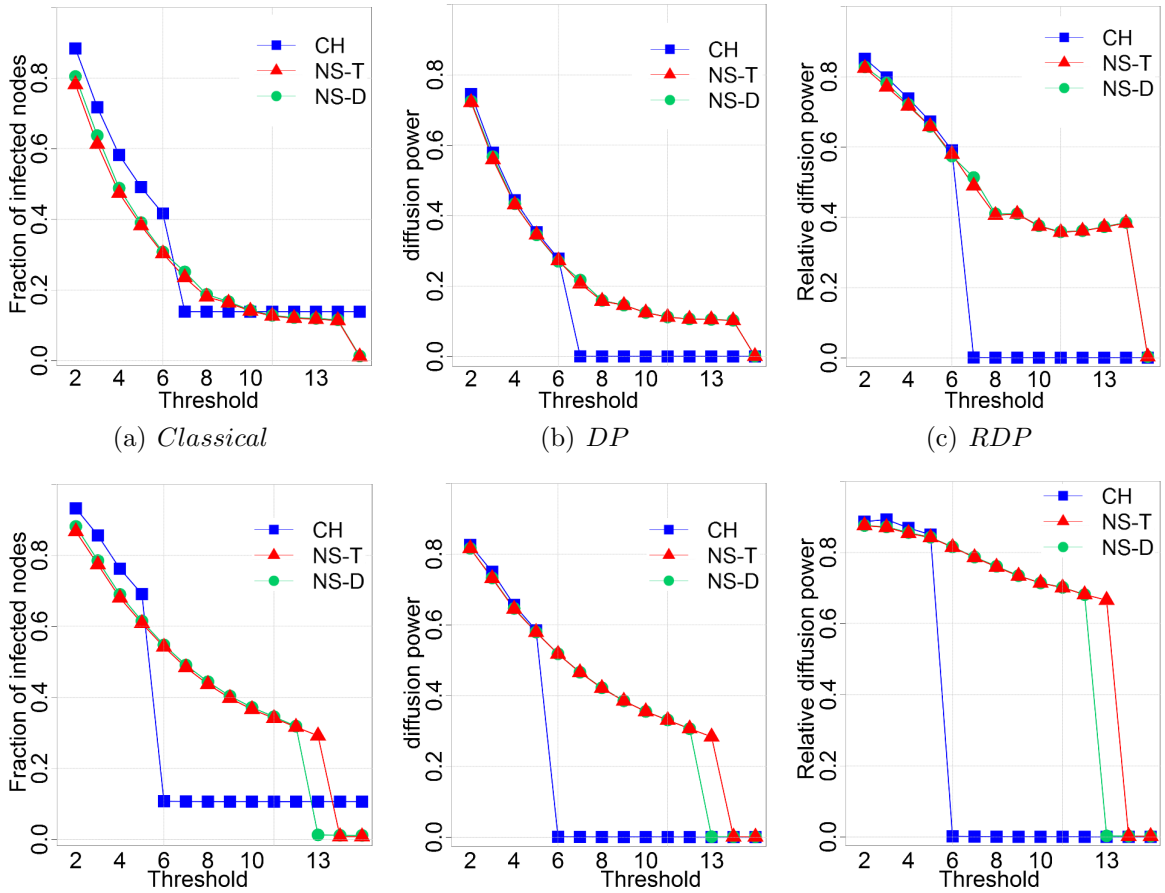
Figure 4.6: *Comparison of fraction of infected nodes per threshold for the networks* cmat *and* Bk *when strategies are given regular rankings (left) and cost-weighted rankings (right), for* $k = 0.0005$ *and* $\alpha = 0.5$.

Despite the latter being hard to determine, the degree of the node is not only trivially obtained, but also quite useful as it defines a ceiling for the number of nodes that can be infected by the epidemic. The *relative diffusion power* (RDP) of a set $I(0)$ (Eq. 4.2) provides such indication.

Figure 4.7 compares these three different measures of influence (classical, DP and RDP) over *hep* and *astro* networks, and illustrates how the perception around strategies' performances changes according to the metric. Note, for instance, that on *hep* network the classical approach (Figure 4.7(a)) indicates for any $\theta > 10$ that the CH strategy performs better than the others, and such a perception is even larger for $\theta = \{14, 15\}$, wherein CH seemingly influences around 15% more nodes than any other strategy. A completely different understanding is captured by DP (Figure 4.7(b)), whereby it becomes evident that what the classical metric indicated as being larger influence (CH) is, in fact, a huge number of seeds whose influence is already nonexistent ever since $\theta = 7$. A key property of DP becomes visible here: it does not take into account the wide-varying number of initial spreaders when assessing how influential they are.

As an example, the performance of 1000 seeds that manage to influence other 100 nodes is exactly the same of that of 5 nodes that also manage to influence 100, as assessed by the DP metric. In another example, the classical metric would point out that 1000 promoters that manage to influence 10 other nodes in a viral marketing campaign over a social network present much better performance than 100 promoters that manage to influence 100 other nodes, whereas DP properly captures that the latter case is more advantageous. It is worth to note that the fraction expressed by DP is never equal to 1, as the occurrence of infections imposes the existence of at least θ seeds, and these are always discounted from the total. Continuing the evaluation, a behavior resemblant of that of Figure 4.7(b) is observed on *astro*, wherein the classical assessment (Figure 4.7(d)) points out that CH always manages to influence at least around 11% ($\approx$ 2000 nodes) of that network, being particularly advantageous when $\theta = \{14, 15\}$. Here again, DP provides a completely different understanding, making it clear that no single strategy has actually managed to induce epidemics under such thresholds. Considering now the seeds' RDP, it reveals that by considering the DP of the different strategies in comparison to the maximum number of nodes that could possibly be infected for a given θ, their score may be significantly raised compared to DP. For instance, when $\theta = 13$ on *hep* network, NS-T influences around 17% of its nodes (Figure 4.7(b)), but these correspond to almost 40% of those nodes whose degree allows for infection (Figure 4.7(c)). Interestingly, note that, because of its definition, RDP may not decay monotonically (as observed in Figure

4.7(c) between $\theta = 11$ and $\theta = 14$). Indeed, this can possibly happen in a network whenever its number of nodes whose degree is d is greater than those with degree $d - 1$.



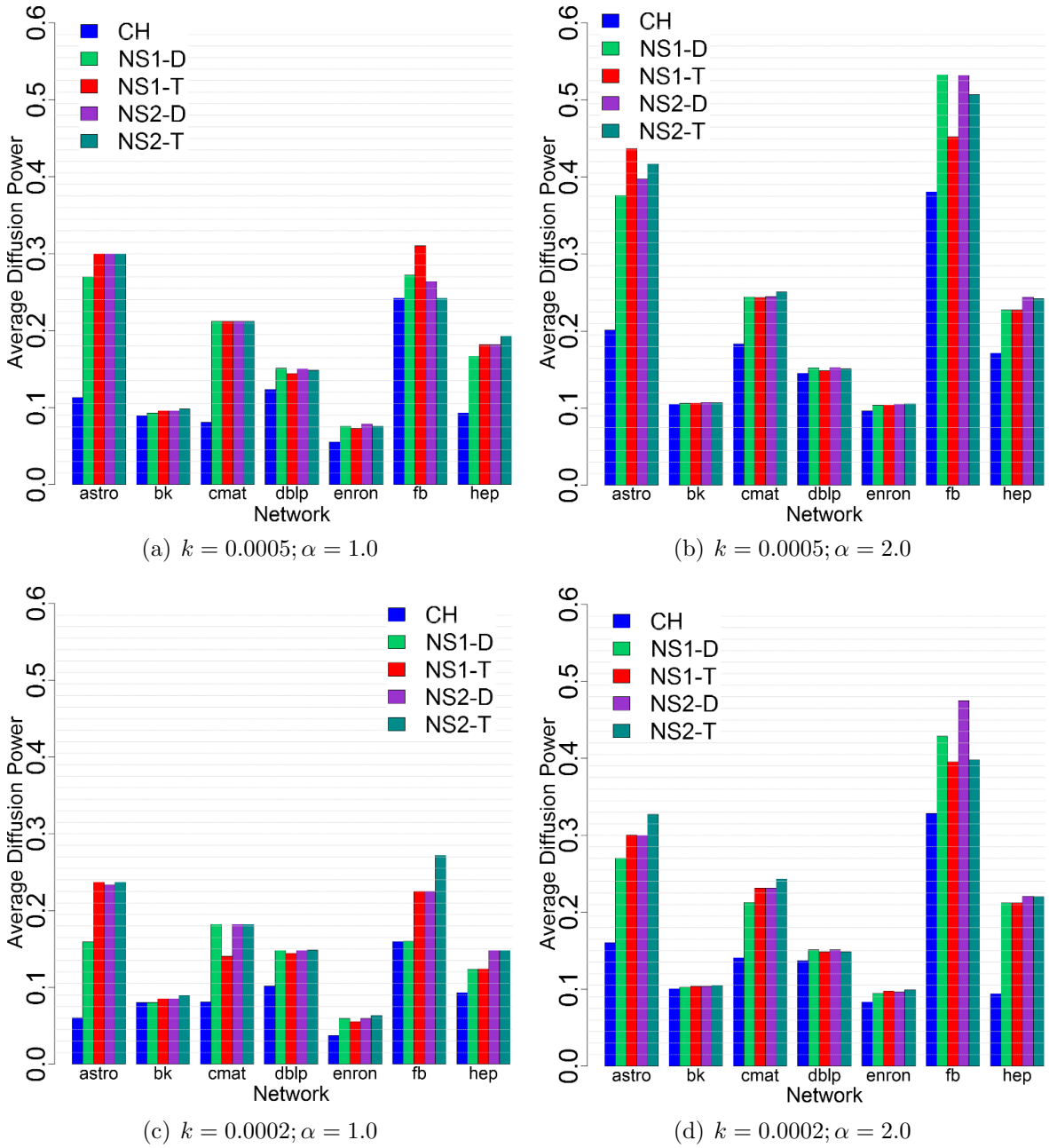
(a) *Classical* (b) *DP* (c) *RDP*

(d) *Classical* (e) *DP* (f) *RDP*

Figure 4.7: *Fraction of infected nodes (left), Diffusion Power (center) and Relative Diffusion Power (Right) per Threshold* for epidemics simulated on the networks *hep* (top) and *astro* (bottom), when $\alpha = 2.0$ and $k = 0.0005$.

We continue by reporting the *Average Diffusion Power* (Eq.4.3) of every strategy in each network, for $\alpha = \{1.0, 2.0\}$ and $k = \{0.0005, 0.0002\}$, as shown in Figure 4.8. It illustrates how strategies' performances tend to increase when these consider two-hop neighborhoods when surrounding nodes, as opposite to strategies which surround nodes considering only their one-hop neighbors. Preserving name conventions, we will denote the former strategies by NS2-D and NS2-T, and the latter, NS1-D and NS1-T. The comparison is made over two different budget regimes, namely $k = \{0.0005, 0.0002\}$, in order to better illustrate how each strategy behaves as the budget diminishes. On each plot, the x axis is composed by groups of bars—one group for each network and one bar for each strategy. The y axis is their corresponding average DP (Eq.4.3). For instance: Table 4.1 shows the maximum θ applied to the network *dblp* is 11, thus the average DP of a given strategy

for *dblp* shown in Figure 4.8 considers $\theta = [2, 11]$. Although the x axis contains all the networks considered, our purpose is not to compare them directly, but to provide the intuition, given a specific network, on the benefits of adopting the extended surrounding sets. In particular, note that the number of cases at which NS2 strategies present better performance compared to NS1 tends to be greater as both α increases (figures 4.8(b) and 4.8(d)) and k diminishes (figures 4.8(c) and 4.8(d)).



(a) $k = 0.0005; \alpha = 1.0$

(b) $k = 0.0005; \alpha = 2.0$

(c) $k = 0.0002; \alpha = 1.0$

(d) $k = 0.0002; \alpha = 2.0$

Figure 4.8: Average Diffusion Power of the different strategies across all evaluated networks under two different budgets.

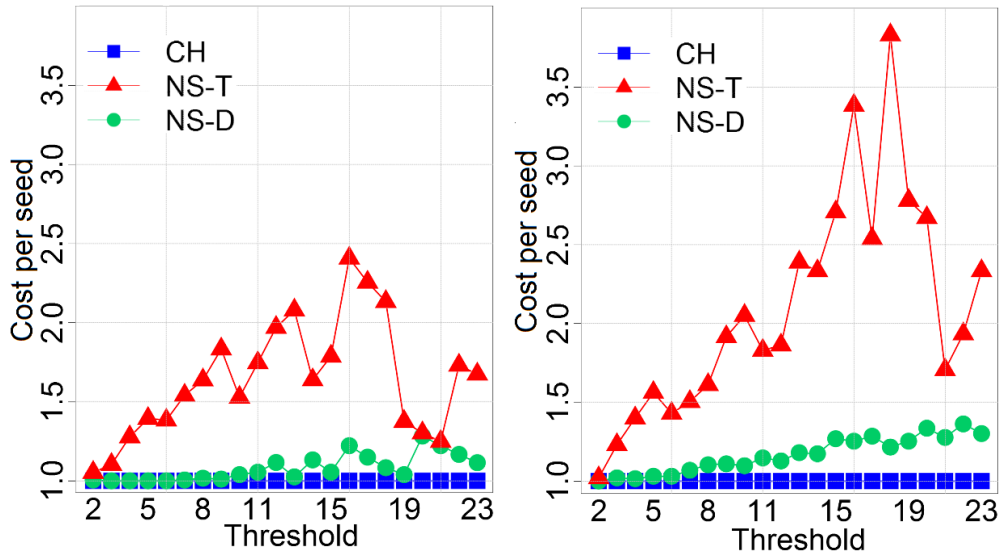
As mentioned in Section 2.2, we are mainly interested in study the influence spreading

on scenarios for which the budget b is relatively small, i.e. b allows for seeding only a negligible fraction of the network. Note that since the same k is applied to every network, b varies according to the corresponding $|V|$ and $\overline{d}$. The size of the different networks varies by orders of magnitude, and this in turn strongly influences b.

Interestingly, in scenarios wherein NS-T have induced larger or equivalent epidemics compared to NS-D, their set of seeds generally corresponds to those with higher *cost per seed* (CS), which in turn implies a smaller number of seeds. For instance, Figure 4.9 shows the *CS per threshold* of the different strategies for the networks *Bk*, when $\alpha = \{0.5, 1.0\}$, and *cmat*, when $\alpha = \{1.0, 2.0\}$. Note that, back on Figure 4.8(a), NS2-T have presented higher DP for that network. The corresponding CS per threshold is on Figure 4.9(b), which reveals that the cost per seed for NS-T was consistently and significantly higher than that of NS-D for all considered thresholds. Similarly, let us now consider NS2-T's DP for *cmat*, as shown on figures 4.8(a) and 4.8(b) for $\alpha = 1.0$ and $\alpha = 2.0$, respectively. Note that, on Figure 4.8(a) it presents a performance similar to that of NS2-D, and on Figure 4.8(b) it manages to be slightly superior. Figures 4.9(c) and 4.9(d) present their respective CS per threshold.

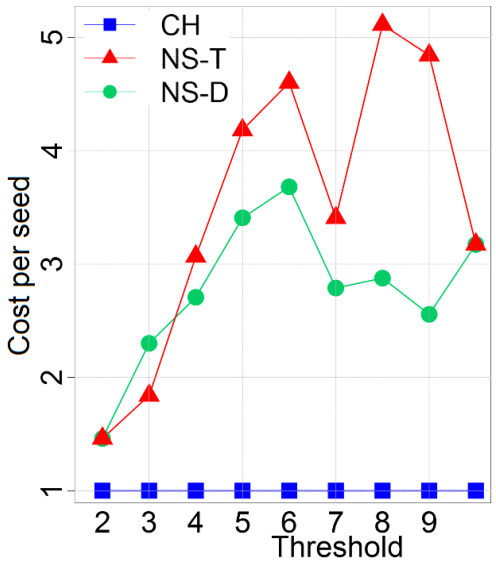
As indicated by Figure 4.9, note that NS-T tends to consistently seed fewer nodes at higher costs, as compared to NS-D, and this in turn clarifies why, for many times, the overall performances of NS-D and NS-T have shown to be similar: the triangle centrality focuses on the network's most connected—and hence most expensive—regions. Although beneficial for the spreading, to seed such regions impose likewise a faster budget exhaustion, leading to a significantly smaller number of initial spreaders, compared to NS-D. Conversely, the degree centrality does not explore epidemic principles as directly as NS-T does, but on the other hand it exploits those three major benefits discussed on Section 3.1 (briefly: fraction of nodes directly reached, trend of neighboring cheap nodes, and trend of having cheap neighbors in common). NS-D therefore counts on a much larger set of initial spreaders.

The adoption of extended surrounding sets (ESS — Section 3.4) may significantly lower the cost per seed (CS) when compared to surrounding sets (Section 3.3), as illustrated in Figure 4.10. Note, for instance, that for $\theta = 7$ the CS for NS-T significantly drops from approximately 12 to around 5, leading its RDP to raise from zero to nearly 19%. The degree distribution of the seeds shows how the adoption of ESS increases the fraction of cheaper nodes in $I(0)$. Note yet that this information is also implicitly given by the CS, since it varies at the same proportion of $|I(0)|$. Last, Table 4.2 shows a direct

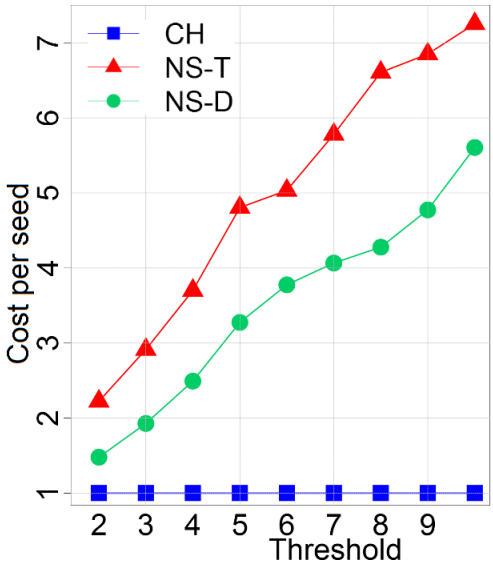


(a) *Bk*; $\alpha = 0.5$

(b) *Bk*; $\alpha = 1.0$



(c) *cmat*; $\alpha = 1.0$



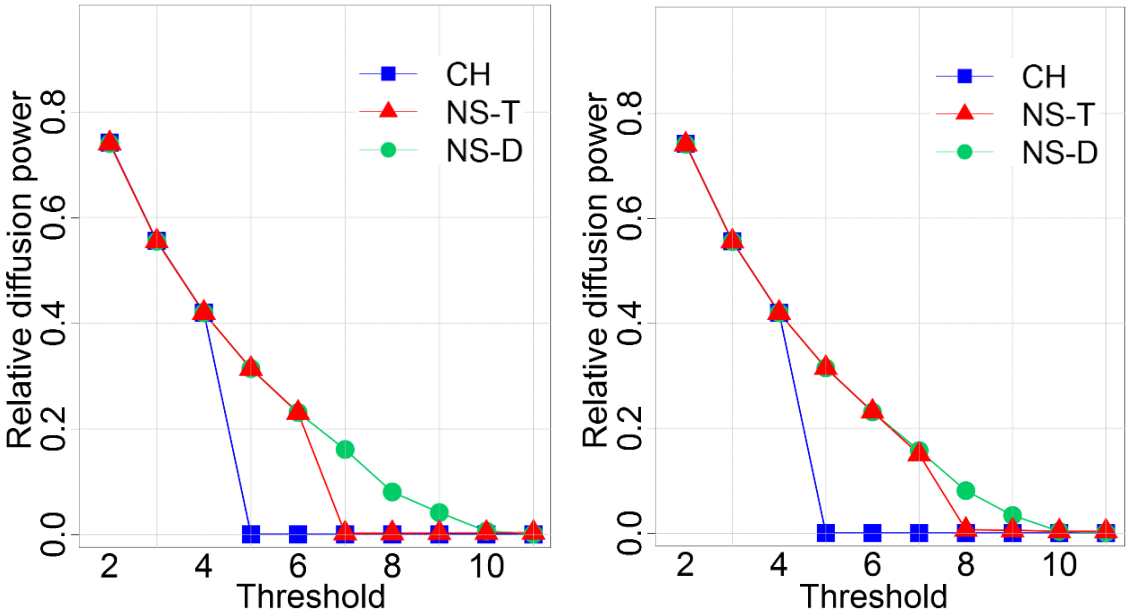
(d) *cmat*; $\alpha = 2.0$

Figure 4.9: Cost per seed for each network threshold applied over *Bk* and *cmat* networks. In both plots, $k = 0.0005$.

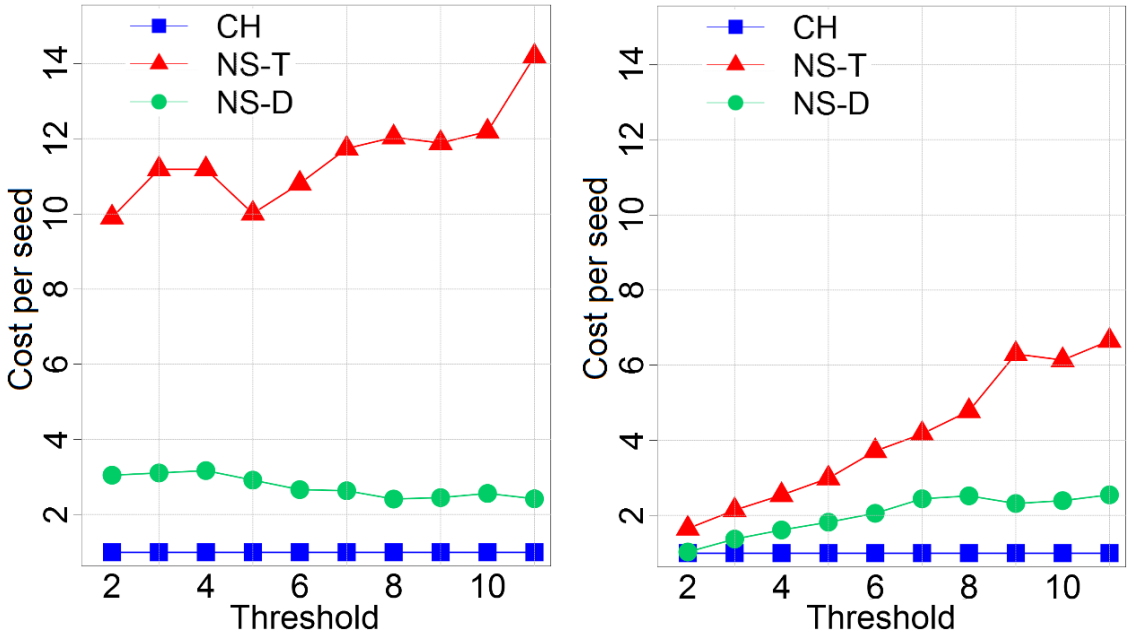
comparison regarding to Figure 4.10, wherein a significant increasing on the number of nodes surrounded is observed for NS strategies when ESS are used.

Table 4.2: A numeric comparison of the seeding for both surrounding sets and extended surrounding sets of Figure 4.10 when $\theta = 7$. $|I(0)|$ = number of initial spreaders.

	Related figures	NS-D	NS-T	CH
nodes surrounded	4.10(a) and 4.10(c)	53	28	0
seeded directly	4.10(a) and 4.10(c)	1	0	951
$\|I(0)\|$	4.10(a) and 4.10(c)	361	81	951
nodes surrounded	4.10(b) and 4.10(d)	74	74	0
seeded directly	4.10(b) and 4.10(d)	1	1	951
$\|I(0)\|$	4.10(b) and 4.10(d)	400	199	951



(a) RDP – SS (b) RDP – ESS



(c) CS – SS (d) CS – ESS

Figure 4.10: *RDP* (top) and *CS* (bottom) for the *dblp* network, employing surrounding sets (SS — left) and extended surrounding sets (ESS — right) when $alpha = 1.0$ and $k = 0.0005$.

Capítulo 5

Conclusion

The understanding on how propagators, like ideas and information, spread in complex networks lies in the core of many applications, such as viral marketing and information diffusion. The reach of an epidemic, however, strongly depends on where it starts—the initial spreaders, i.e. the seeds—, what settles influence maximization as a fundamental, largely-studied problem.

In this work, we considered a deterministic influence maximization problem where nodes have variable costs to be seeded, and proposed efficient strategies to tackle this budgeted influence maximization (BIM) problem. We consider SI epidemics with fixed thresholds over real networks. The cost of a node is proportional to how central (important) the node is, and degree centrality is used as a proxy for node importance. We propose and evaluate the efficiency of different seeding policies, applied over nodes under different sequences, according to some node ranking. Each heuristic was given an initial budget and a cost function to perform the seeding. We focused on scenarios where the initial budget was relatively scarce, as it forces better decisions on seeding and also better relates to real cases.

Our main results show that heuristics that leverage both cost and network structure perform better and are more robust, as they present less sensitivity to parameter variations. We also show that, under risings of the network threshold, to focus solely on either the node cost or the node centrality leads to seeds whose influence becomes negligible much earlier, compared with heuristics that consider the cost-centrality trade-off. Particularly, we show that, for increasing values of the network epidemic threshold, to seed nodes that are not the cheapest ones nor the most central, but that jointly explore some known epidemic principle, tends to induce larger epidemics. Indeed, that was the case regarding the performed exploitation of those more triangulated regions from the

various complex networks. Moreover, we proposed a novel approach to further leverage cost-effectiveness on BIM problems, namely the *extended surrounding sets* (ESS). Our results indicate ESS generally favors seeding strategies to yield broad influence spreading through a wider range of network thresholds. Finally, we introduced the concept of *diffusion power*—a more meaningful metric to assess epidemic performances for BIM. We demonstrated that, opposite to the unit-cost IM, different strategies in BIM problems, upon receiving the same initial budget, may still yield seed sets of largely different sizes. Thus, by considering only the number of non-seeds infected at the end of the spreading, *diffusion power* manages to capture the real benefit (infected non-seeds) of an investment (budget), and hence better suits possible real-world applications. We showed how this approach may completely change the understanding around a strategy's effectiveness.

5.1 Future Work

Despite our effort to cover as much as possible of its intrinsic challenges, BIM is still far from being well understood in real complex networks. In this work we focused on capturing real-world aspects not before considered, and provided some key insights as already discussed. Yet, the problem is under many ways still open. In the following we point out a few aspects either unexplored or not deeply approached in this work, and that we believe could form the basis for future work towards the design of strategies for BIM.

First, we believe that to study improvements on heuristics towards the ESS formation is a promising direction, based on our preliminary results. Indeed, although the ESS adoption has proven advantageous for epidemic performances under many scenarios, it was shown likewise both that its cost-effectiveness is not optimal regarding an entire two-hop neighborhood, and that its current algorithm can be computationally demanding, depending on the network. Second, it would be very opportune to study the problem considering other network epidemic models, under which the whole idea around ESS could be further researched, thus providing a better understanding on its applicability. Last, we believe the study upon new heuristics for ESS could possibly allow to generalize the leveraged neighborhood radius, currently fixed in two hops.

References

[1] Aral, S.; Muchnik, L.; Sundararajan, A. Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science 1* (Aug. 2013), 125–153.

[2] Arthur, D.; Motwani, R.; Sharma, A.; Xu, Y. Pricing strategies for viral marketing on social networks. In *Internet and Network Economics*, vol. 5929 of *Lecture Notes in Computer Science*. 2009, pp. 101–112.

[3] Backstrom, L.; Huttenlocher, D. P.; Kleinberg, J. M.; Lan, X. Group formation in large social networks: membership, growth, and evolution. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2006), KDD '06, pp. 44–54.

[4] Bae, J.; Kim, S. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A: Statistical Mechanics and its Applications 395* (Feb. 2014).

[5] Boguñá, M.; Pastor-Satorras, R.; Vespignani, A. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett. 90* (Jan 2003), 028701.

[6] Brown, K. Here's how much celebrities make in the instagram product placement machine. `http://jezebel.com/heres-how-much-celebrities-make-in-the-instagram-produc-1740632946`, Jan. 2016. Jezebel (visited on 12/02/2016).

[7] Chen, D.-B.; Xiao, R.; Zeng, A.; Zhang, Y.-C. Path diversity improves the identification of influential spreaders. *Europhysics letters 104* (Jan. 2014).

[8] Chen, W.; Wang, Y.; Yang, S. Efficient influence maximization in social networks. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2009), KDD '09, pp. 199–208.

[9] de Souza, R. C.; Figueiredo, D. R.; de A. Rocha, A. A.; Ziviani, A. Evaluation of epidemic seeding strategies under variable node costs. In *SBC Workshop em Desempenho de Sistemas Computacionais e de ComunicaÃ§Ã£o* (2014), WPerformance '14.

[10] Granovetter, M. Threshold models of collective behavior. *American Journal of Sociology 83* (May 1978), 489–515.

[11] Han, S.; Zhuang, F.; He, Q.; Shi, Z. Balanced seed selection for budgeted influence maximization in social networks. In *Advances in Knowledge Discovery and Data Mining - Pacific-Asia Conference* (2014), PAKDD '14, pp. 65–77.

[12] Hinz, O.; Skiera, B.; Barrot, C.; Becker, J. U. Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing 75* (Nov. 2011), 55–71.

[13] Kempe, D.; Kleinberg, J.; Tardos, E. Maximizing the spread of influence through a social network. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2003), KDD '03, pp. 137–146.

[14] Kitsak, M.; Gallos, L. K.; Havlin, S.; Liljeros, F.; Muchnik, L.; Stanley, H. E.; Makse, H. A. Identification of influential spreaders in complex networks. *Nature Physics 6* (2010), 888–893.

[15] Kleinberg, J. Cascading behavior in networks: Algorithmic and economic issues. In *Algorithmic Game Theory*. Sept. 2007, pp. 613–632.

[16] Kornowski, L. Celebrity sponsored tweets: What the stars get paid for advertising in 140 characters. `http://www.huffingtonpost.com/2013/05/30/celebrity-sponsored-tweets_n_3360562.html`, May 2013. The Huffington Post (visited on 12/02/2016).

[17] Kostka, J.; Oswald, Y.; Wattenhofer, R. Word of mouth: Rumor dissemination in social networks. In *Structural Information and Communication Complexity*, vol. 5058 of *Lecture Notes in Computer Science*. 2008, pp. 185–196.

[18] Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J. M.; Glance, N. S. Cost-effective outbreak detection in networks. In *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)* (2007), KDD '07, pp. 420–429.

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

[20] Liu, Y.; Wei, B.; Du, Y.; Xiao, F.; Deng, Y. Identifying influential spreaders by weight degree centrality in complex networks. *Chaos, Solitons and Fractals 86* (05 2016).

[21] Manzano, M.; Calle, E.; Torres-Padrosa, V.; Segovia, J.; Harle, D. Endurance: A new robustness measure for complex networks under multiple failure scenarios. *Computer Networks 57*, 17 (2013), 3641–3653.

[22] Miyauchi, A.; Iwamasa, Y.; Fukunaga, T.; Kakimura, N. Threshold influence model for allocating advertising budgets. In *International Conference on Machine Learning* (2015), ICML '15, pp. 1395–1404.

[23] Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett. 89* (Oct 2002), 208701.

[24] Newman, M. E. J. *Networks: An Introduction*. 2010.

[25] Nguyen, H.; Zheng, R. On budgeted influence maximization in social networks. *IEEE Journal on Selected Areas in Communications 31*, 6 (2013), 1084–1094.

[26] Perlberg, S. Facebook signs deals with media companies, celebrities for facebook live. http://www.wsj.com/articles/facebook-signs-deals-with-media-companies-celebrities-for-facebook-live-146653 June 2016. The Wall Street Journal (visited on 12/02/2016).

[27] Seidman, S. B. Network structure and minimum degree. *Social Networks 5*, 3 (Sept. 1983).

[28] Socievole, A.; Rango, F. D.; Scoglio, C.; Mieghem, P. V. Assessing network robustness under {SIS} epidemics: The relationship between epidemic threshold and viral conductance. *Computer Networks 103* (2016), 196–206.

[29] Tang, S.; Yuan, J.; Li, X.; Wang, Y.; Wang, C.; Liu, X. MINT: maximizing information propagation in predictable delay-tolerant network. In *ACM International Symposium on Mobile Ad Hoc Networking and Computing* (2013), MobiHoc '13, pp. 253–256.

[30] Wang, S.; Wang, F.; Chen, Y.; Liu, C.; Li, Z.; Zhang, X. Exploiting social circle broadness for influential spreaders identification in social networks. *World Wide Web 18*, 3 (2015), 681–705.

[31] Watts, D. J. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences 99*, 9 (2002), 5766–5771.