Cores matter? An analysis of graph decomposition effects on influence maximization problems

Antonio Caliò
DIMES Dept., University of Calabria
Rende (CS), Italy
a.calio@dimes.unical.it

Andrea Tagarelli
DIMES Dept., University of Calabria
Rende (CS), Italy
andrea.tagarelli@unical.it

Francesco Bonchi
ISI Foundation
Torino, Italy
francesco.bonchi@isi.it

ABSTRACT

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

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

CCS CONCEPTS

• **Mathematics of computing** → **Graph theory**; • **Information systems** → **Web searching and information discovery**.

ACM Reference Format:
Antonio Caliò, Andrea Tagarelli, and Francesco Bonchi. 2020. Cores matter? An analysis of graph decomposition effects on influence maximization problems. In *12th ACM Conference on Web Science (WebSci '20), July 6–10, 2020, Southampton, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394231.3397908

WebSci '20, July 6–10, 2020, Southampton, United Kingdom

ACM ISBN 978-1-4503-7989-2/20/07…$15.00
https://doi.org/10.1145/3394231.3397908

1 INTRODUCTION

Measuring and understanding the spread of "contagion" has attracted tremendous attention as a universal phenomenon that is extensively studied in physical, biological, and social networks. Exemplary application domains are related to social influence, diffusion of information, misinformation or rumors, spread of viruses etc. In this context, a key problem is the identification of the most effective spreaders in a social network. In order to estimate the spreading potential of nodes in a social network, several heuristics have been studied: centrality measures, such as degree or PageRank, or mesoscale-structure-based properties of nodes, such as *core decomposition*. One important study by Kitsak et al. [19] showed that the influential spreaders are those located in the inner-most core of the network, in contrast to the fact that high-degree or high-betweenness nodes could have little effect on the extent of a spreading process. Since then, several studies have been proposed to improve the discriminating ability (i.e., monotonic ranking of spreaders) of the core decomposition (e.g., [2, 15, 23]). In this line of research, the network is assumed to be *undirected*, and the empirical findings on the spreading process refer to standard epidemic models (e.g., SIR or SIS).

An alternative line of research corresponds to the widely studied *influence maximization* (IM) [18] problem: given a directed network, a (stochastic) diffusion model, and a budget on the number s of seeds (i.e., early-adopters or initial influencers), IM asks to find a s-sized seed-set S that maximizes the *influence spread* over the network, i.e., the expected number of nodes that are activated, starting from S, at the end of the diffusion process. The main distinction between finding a good seed-set and estimating the spreading potential of nodes in isolation, is that the former problem requires to take into account the cumulative effect of the influence spread. In fact, different nodes may exert influence on largely overlapping portions of the network, so that their cumulative spread would be wrongly estimated by just considering the sum of their spreading potential.

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

Example 1.1. Let us consider the example graph in Fig. 1. Suppose we are required to select one seed of the propagation process (i.e., $s = 1$). It can be noted that node v_1 has a strategic location as it can reach all nodes in the graph. This is clearly an ideal situation for an IM which, depending on the setting of influence probabilities (here omitted for simplicity) and the diffusion model adopted, will likely select v_1 as seed. By contrast, most of the centrality measures will fail at capturing the spreading ability of that node in the network. In fact, none among out-degree, directed closeness and betweenness, and

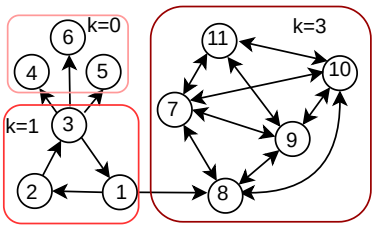

Figure 1: Core decomposition over a directed network. Cores are determined according to the nodes' *out-degree*.

PageRank is able to rank v_1 at the top. Also, considering the outcomes of out-degree-based core decomposition of the example graph, any node in the inner-most core (i.e., core with $k = 3$), would be preferred as seed to any other node with lower core-index, including v_1, despite no nodes in the inner-most core can propagate outwards, thus they cannot be an optimal choice for an IM solution.

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

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

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

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

Finally, we provide evidence that a major limitation that prevents classic decomposition algorithms to find the most influential spreaders, is their inability to incorporate higher-order degree of nodes. Our analysis shows that *distance-based generalization of core decomposition* [9] provides a more informative characterization of how important nodes are in terms of their reachability, thus providing an effective approach to the identification of good spreaders.

2 BACKGROUND AND RELATED WORK

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

Influence propagation. The analysis of social contagion, i.e., the spread of new practices, beliefs, technologies and products through a population, driven by *social influence*, is a very central theme in social sciences, and it has also attracted a lot of interest in the data science community [6]. Such phenomenon develops in two main subjects: the structure of the network and the actions or communications of the users over the network. Researchers have studied the role played by the network topology [33] and by several of its macroscopic characteristics, such as the level of homophily [34] and the modular structure of the network [4, 25], as well as node-level characteristics, such as their centrality, or their capacity of spanning structural holes, thus bridging communities and facilitating, or blocking, the spread of information. Other researchers have considered the social network and the log of past user-activities jointly, and studied important problems such as learning the parameters of the propagation model, i.e., the strength of influence along each edge [16, 28], or how to distinguish real social influence from "homophily" [1, 8, 12, 13]. Finally, a wide literature exists on the analysis of social influence in specific domains: for instance, studying person-to-person recommendation for purchasing books and videos [20, 22], telecommunications services [17], or studying information cascades driven by social influence in Twitter [3, 27].

Fueled by the seminal work by Kempe et al. [18], most of the attention has been devoted to exploiting social influence for "word-of-mouth" driven viral marketing applications: this is the case of the stochastic optimization problem known as influence maximization (IM). Given a social network, where each edge (u, v) is associated with a weight (or probability) $p_{u,v}$ representing the strength of influence that u exerts over v, IM requires to select the set of initial users that maximizes the *expected spread*, i.e., number of users in the social network that gets "infected", according to an assumed underlying *diffusion model*. IM is NP-hard under most standard diffusion models, such as *Independent Cascade* (IC) and *Linear Threshold* (LT) models, however, the simple greedy algorithm provides $(1-1/e)$ approximation guarantee, provided that the diffusion model is monotone and submodular (like in the cases of IC and LT). Since the expected spread cannot efficiently be evaluated exactly, most of the effort have been devoted to address this scalability issue by reducing the number of needed Monte Carlo estimations [21]. Alternatively, proxy-based methods have been developed to avoid running Monte Carlo simulations, by estimating the influence spread of the seed

set through a reduced diffusion context; although, without ensuring theoretical approximation guarantee.

A significant study that overcomes the efficiency bottleneck of the simulation based methods, while preserving the theoretical approximation guarantee, is proposed in [10], which introduces the *Reverse Influence Sampling* (RIS) framework for IM. The key idea is that the expected spread can be estimated by taking into account a number of pre-computed sketches, i.e., realizations drawn from the distribution induced by influence graph according to the diffusion model. This breakthrough result paved the way for a variety of sketch-based algorithms. Tang et al. in [31] are the first to design a practically efficient solution, TIM/TIM+, whose improvement over RIS consists in keeping the same theoretical complexity as [10] with significantly fewer sketches, bounded by the influence of the unknown optimal set (OPT). More recent RIS-based IMM [30] and SSA [26] algorithms share the common motif of estimating OPT with a fewer number of sketches. IMM improves over TIM/TIM+ through a martingale analysis, while SSA takes an orthogonal perspective, as the number of sketches needed by the algorithm is determined at runtime via an iterative approach. The TIM/TIM+, IMM, and SSA methods will be used in our evaluation (§4–6).

Graph decomposition. *Cores* in a graph were first studied in [29] for characterizing tightly-knit groups in social networks. Since then, core decomposition has been used as a tool for several applications related to the understanding of mesoscale structural characteristics of a network, but also to capture the centrality or influential status of nodes. Among its advantages, core decomposition for an input graph is unique, and hence well-defined, and it can be computed efficiently in linear time w.r.t. the number of edges in the graph.

As mentioned in the Introduction, [19] is one of the earliest studies exploring relations between the spread of influence in undirected networks and core decomposition. The study shows that, under the SIR epidemic model, nodes with the best spreading potential are likely not those with the highest degree or betweenness centrality, but are in the most internal core of the network.

Following the lead of [19], in [24] a similar analysis is carried out in terms of *truss* decomposition [32]. Nodes selected within internal regions of a network according to the truss decomposition, tend to produce infections that are significantly more viral in the early steps of propagation as opposed to the one obtained started from the most internal cores, though this advantage becomes less evident as the propagation approaches to its termination.

The *k-peak* decomposition proposed in [15] aims to find robust decomposition when a network has distinct and independent regions of different edges density. Following the same setting as [19], it is shown that when the initial spreaders are chosen among those with the highest k-peak number, the size of the information cascade may be up to 50% greater than the size based on k-core decomposition. In [2], the *coreness centrality* is defined on top of the classic core decomposition, by aggregating the core-index of all neighbors of a given node. Again under the SIR model and for undirected and unweighted networks, this method has shown to produce better rankings than those based on k-core decomposition.

A recent study [9] extends the k-core decomposition to account for a neighbor-distance threshold h. Differently from [2], the proposed notion of (k, h)-core redefines the coreness property based on a higher-order degree of nodes, i.e., the core-index of a node is function of the number of nodes reachable up to a given distance h.

The notion of k-core decomposition was also extended to probabilistic graphs [7]. The (k, η)-core is defined as a subgraph such that each of its nodes has at least degree k with confidence at least η. Notably, in an IM evaluation scenario, where the edge probabilities are assumed to be influence propagation probabilities, the greedy algorithm could in principle exploit the computation of (k, η)-cores in order to locate the seeds in the inner most η-cores. Another bivariate-core notion is proposed in [14], where the (k, l)-D-core is defined to account for nodes with in-degree at least k and out-degree at least l. The significance of this approach was mainly assessed over collaboration networks where, unlike social influence-driven networks, both the inward and outward connectivities of nodes might be explicitly parametrized.

3 DECOMPOSITION OF DIRECTED GRAPHS

In this section, we present the graph decomposition methods examined in our study. One notable point is that, since these methods are *originally conceived for undirected networks* (cf. §2), we first need to revise their definitions in order to make these methods amenable to support an IM task, which requires a directed network as input context of influence propagation. Also, *our choice of decomposition methods was guided by two main factors*: (i) they are able to scale to large networks; (ii) they can be meaningfully extended to directed networks; and (iii) they are representatives of the most widely used decomposition strategies and variants.

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

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

It is easy to show that the well-known $O(|E|)$ algorithm in [5] can straightforwardly be adapted to a directed network: nodes are ordered by increasing out-degree, then nodes u with lowest out-degree are iteratively removed from the graph and each incoming neighbor of u decreases its out-degree, and the process continues until no node remains. The core-index of a node is the out-degree at the moment of its removal.

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

there is a non-empty k-contour; it is straightforward to note that $K^P(G) = K^C(G)$, for any graph G.

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

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

Once the support of each edge is computed, we apply the algorithm proposed in [32] to obtain the decomposition. However, since the k-truss decomposition is defined with respect to the edges of the graph, we eventually assign a score to each node that is equal to the average truss-index of the node's outgoing edges. Also, we denote with $K^T(G)$ the highest of the node truss-indexes.

Neighbor-coreness aggregation. Adapting from [2], each node v is assigned with a *neighbor-coreness* score given by $C_{nc}(v) = \sum_{u \in N^{out}(v)} c(u)$, where $c(u)$ denotes the core-index assigned to node u and $N^{out}(v)$ is the set of v's out-neighbors. We also denote with $K^{NC}(G)$ the maximum neighbor-coreness score.

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

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

To compute the (h, k)-cores, we adapted Algorithm 1 in [9] by specializing the notion of h-neighborhood for out-neighbors.

Example 3.1. Let us consider again the example shown in Fig. 1, to check whether the various graph-decomposition algorithms are able to assign v_1 with the highest score. We have already observed that this is not the case when using the k-core decomposition (cf. Example 1.1). Similar outcome holds also for the k-peak decomposition — two distinct contours are found, with v_1 having peak-number 0 along with nodes $v_2, \ldots, v_6$, and the remaining nodes with peak-number 3 — the k-truss decomposition and the neighbor-coreness aggregation method — which assign the highest score to nodes $v_7, \ldots, v_{11}$. By contrast, the distance generalized core decomposition is able to detect, for $h = 2$, three cores, where the inner-most one does contain node v_1 (along with $v_7, \ldots, v_{11}$).

Table 1: Summary of evaluation network data.

network	#nodes	#edges	avg. in-deg.	avg. path len.	dens.	diam.	#sources	#sinks
DBLP - DB	317K	1M	3.31	7.89	$1.05e^{-5}$	31	127K	12K
Epinions - Ep	116K	722K	6.2	4.79	$5.3e^{-5}$	16	28K	43K
Nethept - Net	15K	62K	4.1	5.83	$2.7e^{-4}$	5	0	0
Twitter - Tw	21K	227K	10.38	6.28	$4.7e^{-4}$	32	3K	3K
Instagram - Ig	17K	617K	35.25	4.24	$2e^{-3}$	15	0	0
FriendFeed - FF	493K	19M	38.85	3.82	$7.8e^{-5}$	32	42K	292K

4 EVALUATION METHODOLOGY

We used 6 real-world online social network datasets, whose properties are summarized in Table 1. Our choice of these network data is justified as they can be regarded as benchmarks in IM or graph-decomposition studies. In particular, *Epinions, DBLP, Nethept* networks were used in the original works proposing the three IM methods under examination (i.e., TIM/TIM+, IMM, and SSA); the *Twitter* dataset was used in [7] to assess the significance of uncertain graph decomposition for IM; *Instagram* and *FriendFeed* were studied in [11] for targeted IM in a user engagement context.

We considered the two most commonly used diffusion models in IM, namely Independent Cascade (IC) and Linear Theshold (LT) models [18]. Due to space limits, the results presented in the remainder of this paper are only based on IC. The experimental results using LT — which can be found in the *Supplemental Material* available online — are consistent with the findings for IC, reported in this paper.

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

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

Results are organized into two main sections: first, we focus on those methods that rely only on first-order node-degree information (§5), then we evaluate the impact of the adoption of a distance-aware generalization of the core-decomposition (§6).

5 DEGREE-BASED CORES

We investigate where the most influential nodes selected by state-of-the-art IM algorithms — TIM/TIM+ [31], IMM [30], and SSA [26] (cf. §2) — are located in the network w.r.t. different graph-decomposition

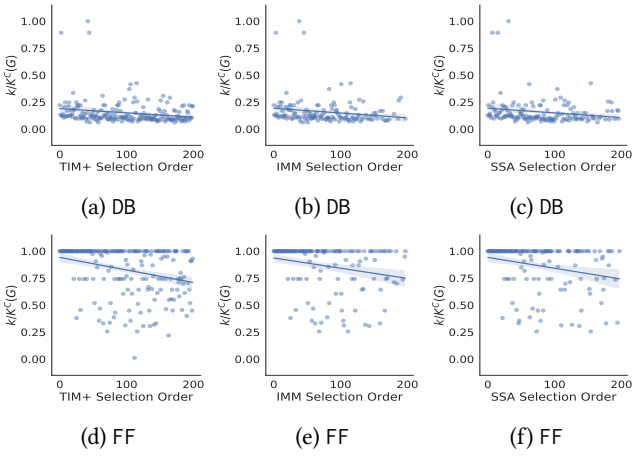

Figure 2: Normalized core-index ($k/K^C(G)$) of the first 200 seeds computed by (a,d) TIM+, (b,e) IMM, and (c,f) SSA, under the IC model.

methods (§5.1). Prompted by the results obtained in this early step of evaluation, we will delve into the features that could be used as proxies for identifying a "good" subnetwork for locating IM-(near)optimal influential spreaders (§5.2).

5.1 Seed selection order

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

Figure 2 reports on the y-axis the normalized core index (i.e., the core index of the node divided by the degeneracy of graph) for the first 200 seeds — computed by TIM+, IMM, and SSA, respectively — ordered on the x-axis according to their selection order, i.e., the iteration corresponding to the insertion of a node into the seed set. For this analysis, we report only results corresponding to two datasets; nonetheless, these results are representative of a general scenario encompassing all remaining networks. We refer the reader to the *Supplemental Material* associated with this paper.

One remark that stands out is that the three IM methods exhibit a very consistent behavior, which seems to depend mostly on the network. This is not really surprising, since all such algorithms share the state-of-the-art RIS-based approach in their algorithmic scheme (cf. §2). While on dataset DB most of the seeds, with few notable exceptions among the first seeds, are in peripheral cores (the majority of the seeds have core-index between the 5-th and 25-th percentiles), for FF the situation is slightly different: many seeds are selected in high cores, although a good portion of seeds are identified in lower cores. What is common to both datasets (and to the others not reported in Fig. 2) is that, as hinted by the regression line in each plot, as the selection progresses the various IM methods are more likely to identify the seeds among those with lower core index, i.e., in the periphery of the network. This can be explained with the fact that our evaluation methods work under the IC and LT diffusion models, whose activation functions are monotone and submodular: after the earlier stages of seed selection, the IM methods would start exploring the periphery of the network,

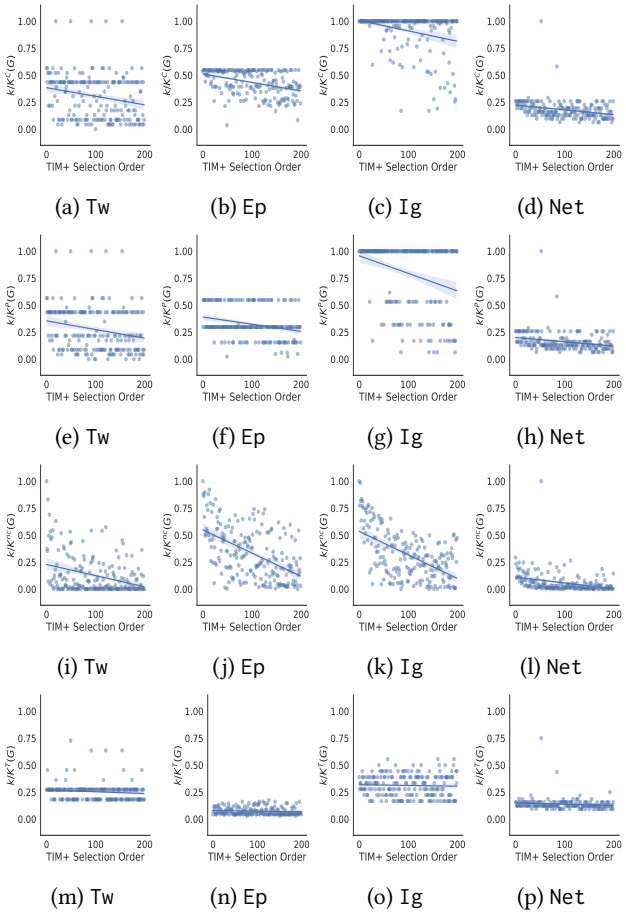

Figure 3: From top to bottom, normalized core-index ($k/K^C(G)$), peak-number ($k/K^P(G)$), neighbor-coreness ($k/K^{NC}(G)$), and truss-index ($k/K^T(G)$) of the first 200 seeds computed by TIM+, under the IC model.

since therein it will likely reside the nodes whose marginal gain is potentially less affected by the earliest selections.

This is in contrast with the findings in [19, 24], according to which the most influential nodes should reside in the inner-most core of the network. This difference is due to the fact that those works consider a SIR propagation model, whereas we use IC/LT, and on the fact that they focus on the spreading potential of nodes in isolation, while our analysis considers the cumulative expected spread of the seed set of the IM problem.

Results drawn from the previous analysis were confirmed by analogous evaluation extended to the other graph decomposition methods and networks. As shown in Fig. 3, in most networks (e.g., Tw, Ep, Net), the majority of the seeds are located in subnetworks that correspond to mid/low values of each particular decomposition method. One exception is represented by Ig, where most seeds are located in the inner subnetworks, provided that k-core or k-peak decomposition is used. Among the various decomposition techniques, it can be noted that neighbor-coreness provides high-variance, hence poorly meaningful results for our analysis. This is explained since neighbor-coreness was originally conceived as a

proxy solution for ranking nodes w.r.t. their individual influence, rather than for achieving coarser-grain graph-decompositions; this also prompted us to ignore it in the remainder of our study. Another interesting remark regards the k-truss decomposition. In fact, identifying the seeds within the inner-most subnetworks induced by this method appears to be a disadvantageous choice for our purposes, as most of the seeds are located within the outer subnetworks (i.e., those containing nodes with lower truss-index values).

The above results, coupled with the ones discussed in the previous section (§5.1), provide evidence that allocating seeds in the inner-most regions of a network may turn out to be a poorly effective strategy for IM. This contingency may be ascribed to the fact that concentrating the selection of nodes within the same subnetwork induced by a graph-decomposition technique, would prevent us to exploit the submodularity of the activation function of the IM methods. Intuitively, it may happen that the propagation remains trapped inside the densest regions of a network, and consequently it will not be able to involve other parts of the network; this particularly holds for the k-truss decomposition, which considers the number of triangles a particular node is involved in.

Notably, our findings totally fit the LT model as well. Due to space limitations of this paper, results corresponding to LT can be found in the online-available *Supplemental Material*.

5.2 Characterization of the Cores/Contours

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

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

Core/Contour distribution. Figure 4 shows how nodes are distributed over the different cores of the network. If we compare these results in light of the previous findings (§5.1), we observe that a lower skewness in the distribution would correspond to the identification of seeds within the inner cores. In fact, the distributions for FF and Ig, exhibit a much lower skewness (i.e., 3.2 and 2.3, resp.) as compared to the one corresponding to the other networks, however with the exception of Tw. Albeit the skewness could serve as a moderately good indicator of how effective it will be to allocate seeds within the inner cores of a network, we need to further investigate the characteristics of the cores.

As regards the k-peak decomposition (results shown in the online-available *Supplemental Material*), we observe it tends to favor skewer distributions than the core-decomposition ones. In particular, although $K^C(G) = K^P(G)$, the number of distinct contours in the evaluation networks is found to be consistently smaller than the number of distinct cores in the network. This implies that the k-peak decomposition may provide a coarser view on a network structure, where most nodes are concentrated in the subnetworks

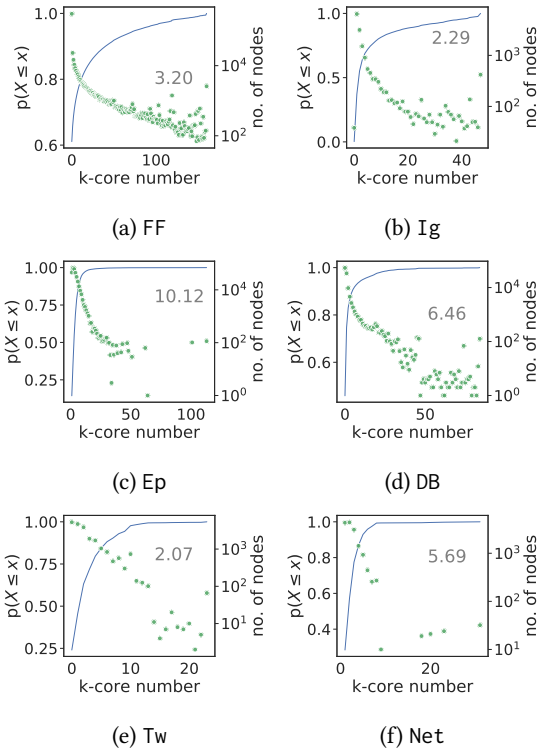

Figure 4: Distribution of nodes over the cores of the network. Each plot shows, for every core-index k (x-axis), the number of nodes with core-index at most k on the rightmost y-axis, and the cumulative distribution of core-index on the leftmost y-axis. Also, the skewness of the distribution is reported inside each plot.

with lower peak number, thus hindering the ability of this technique to discriminate the influence-spread potential of nodes.

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

Figure 5 shows the fraction of edge-set that belongs to each of the two classes, based on core-indexes of their sources — very similar behaviors were also found in results corresponding to peak-numbers (shown in the online-available *Supplemental Material*). We recognize three types of characteristics in the inward percentage-bars, as the normalized core-index increases: (i) a roughly decreasing trend, for FF and Ig, (ii) a roughly constant trend, for DB, and (iii) a roughly bimodal decreasing trend, for the remaining networks. For the former group, while the inward percentage remains much higher than the outward one until mid-high regimes in the x-axis, this gap tends to become small for the highest cores, showing that nodes in the inner-most core (i.e., rightmost side of a plot) also have a good connectivity towards the periphery of the network. Quite differently from FF and Ig, Ep and Tw show a roughly bimodal decreasing behavior, which appears to have a break-point around half

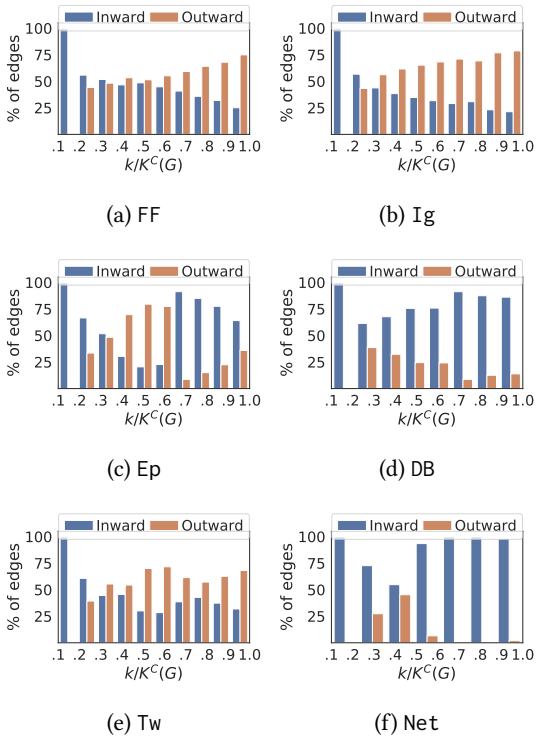

Figure 5: Percentage of inward and outward edges vs. normalized core-index $k/K^C(G)$. The i-th percentage bar ($i = 1..9$) corresponds to edges such that the source node has normalized core-index in $(x_i, x_{i+1}]$, upon a segmentation of the x-axis values into ten intervals $(x_1, x_2], \ldots, (x_9, x_{10}]$.

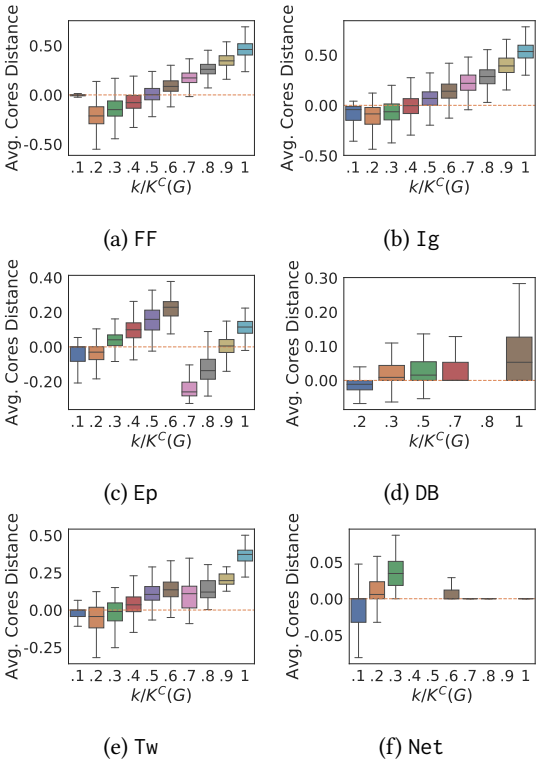

Figure 6: Distribution of the node's average normalized core distance vs. normalized core-index $k/K^C(G)$. For each core-index k, the corresponding boxplot represents the distribution of the average normalized core distances computed for each node having core-index k.

of the degeneracy. Notably, this corresponds to the core where most of the seeds are actually found according to the results discussed earlier in this section (§5.1). However, while the second decreasing trend ends up with a 60% inward edges for the first and second inner-most cores in Ep, a further interesting scenario occurs in Tw. Here, the nodes within the inner-most core are mostly connected to each other, since a considerably high fraction of edges (above 80%) are inward. In DB, the inward edges are the large majority, regardless of the core-indexes of their nodes, which might be ascribed to a relatively high percentage of source nodes.

Pairwise core distances. We consider here a more robust measure than the inward/outward property of edges, which accounts for the difference of core-index values of two linked nodes. Given any edge (u, v), we define the *pairwise normalized core distance* as $dist(u, v) = (k_u - k_v)/K^C(G)$, with k_u and k_v the core-index assigned to u and v, respectively. Upon this, for each node u we compute the average normalized core distance over its out-neighbors. A positive value means that u is mostly connected with nodes belonging to outer cores, and the greater the value, the more u's out-neighbors can be considered as peripheral w.r.t. the u's location.

Figure 6 shows the boxplot distributions of average normalized core distance w.r.t. the normalized core-index values. The analysis of such plots allows us to integrate and enrich the results observed in Fig. 5. Considering first Ig and FF, where most of the seeds have the maximum core-index (§Figs. 2–3), we observe a clearly increasing trend of the nodes' average normalized core distance. With corresponding boxplot median around 0.5, nodes within the highest core-index show to be well connected with nodes located in mid-level outer cores. A different situation is observed on Tw, Ep, and DB, where the maximum average normalized core distance mostly remains below 0.4, 0.3, and 0.1, respectively. Remarkably, in Ep (Fig. 6(c)), where most seeds have mid/low core-index (§Fig. 3), we observe again a breakpoint in the distribution around half of the degeneracy, where the peak of average normalized core occurs, while the second increasing trend almost remains below positive values in the y-axis, with the inner-most boxplot having very low median (around 0.1). Also, on DB (Fig. 6(d)), the values of range of each boxplot (always below 0.1) indicate that the edges tend to connect nodes that have very close core-index, which is also consistent with the fact that nearly all seeds are not located within the inner-most core (§Fig. 3).

5.3 Discussion

In this first stage of evaluation, we have learned that searching for influential spreaders within the inner subnetworks (based on any particular decomposition method) does not ensure to find the best seeds for an IM problem. Indeed, it should not be surprising that topological properties of the networks take a crucial role in

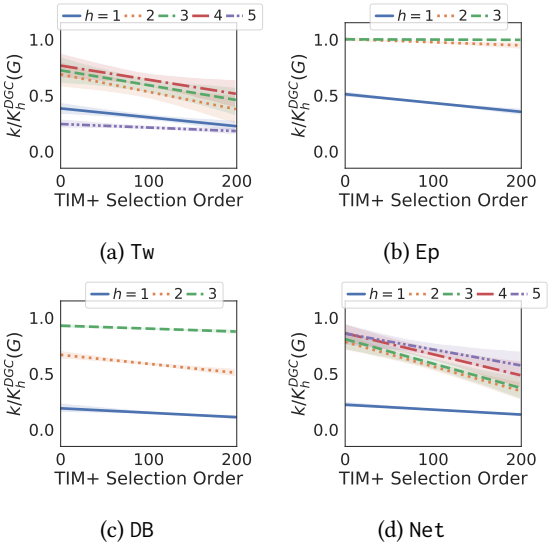

Figure 7: Linear regression of the normalized distance-generalized core-index ($k/K_h^{DGC}(G)$) of the first 200 seeds computed by TIM+, under the IC model.

determining whether or not nodes in the inner-most cores have the best influence-spreading potential. In fact, Ig and FF, where most of the seeds were found in the inner-most core, are also the networks that exhibit a significantly higher average in-degree and a network density that is slightly higher than the other networks (§Table 1). The remaining networks, where seeds were mostly identified outside the inner-most cores, show a substantially sparser structure, as indicated by their values of average path length, density, and diameter.

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

6 HIGHER-ORDER CORES

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

Results are organized into three parts. In the first part, we replicate the same setting adopted in the early step of the previous evaluation (cf. §5), in order to assess the relation of (k,h)-core decomposition with the outcomes of an IM algorithm (§6.1). Next, we assess the sensitivity of the decomposition w.r.t. the value of the neighbor-distance threshold h (§6.2). Finally, we also investigate the individual influential-spreading potential of nodes, and put this in relation with the decomposition outcomes (§6.4).

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

Table 2: Maximum (k,h)-core-index (leftmost) and number of distinct (k,h)-cores (rightmost), for varying h.

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

6.1 Seed selection order

Analogously to the analysis presented in the first phase of Stage 1 (cf. §5.1), we first investigated the relations between the (k,h)-core-index values and the selection order of the discovered seeds.

Looking at the plots in Fig. 7, it stands out that *a significant fraction of seeds is now found to be located in the inner-most (k,h)-core(s).* This is particularly evident in Ep and DB, where all top-200 seeds (i.e., not only the early-selected ones corresponding to a small budget s) are in the inner-most core or immediately outer one, with $h \in \{2,3\}$ and $h=3$, respectively. A further important finding is that while regression lines tend to rise up for higher h, with major gain from $h=1$ (i.e., equivalent to core decomposition) to $h=2$, this trend is not monotone in general. Indeed, it may happen that an overly high value of h (typically higher than 4) could lead to decreased performance, even worse than the corresponding core decomposition (as observed for Tw, where the regression line for $h=5$ lays on about 0.25).

6.2 Sensitivity to h

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

First, as reported in Table 2, we observe that the number of cores and the maximum core-index grow significantly as h increases — recall that $h=1$ corresponds to the classic k-core decomposition — which suggests how the (k,h)-core decomposition can enable a fine-grain micro/mesoscale structure analysis.

In Fig. 8, we observe that, when $h>1$, the number of nodes in the subnetworks with lower (k,h)-core-index is significantly smaller than for $h=1$. This is clearly due since nodes tend to be more connected to each other as h increases. More interestingly, the inner-most generalized cores (i.e., tail of the distributions) are consistently more populated than for $h=1$. Nonetheless, as displayed in the insets of Fig. 8 for all networks, the inner-most generalized core covers a fraction of the whole node-set that is relatively small, yet meaningful for a seed-set selection task.

6.3 Discussion

We have unveiled that the best-influential spreaders can actually be located within one or very few inner-most core(s) of a network provided that a higher-order graph-decomposition method is used. The neighbor-distance threshold (i.e., h) plays a key role in the decomposition, since too large values of the parameter may in principle lead, at the cost of increased computational overhead, to few cores covering most nodes in the network, thus reducing the benefits of

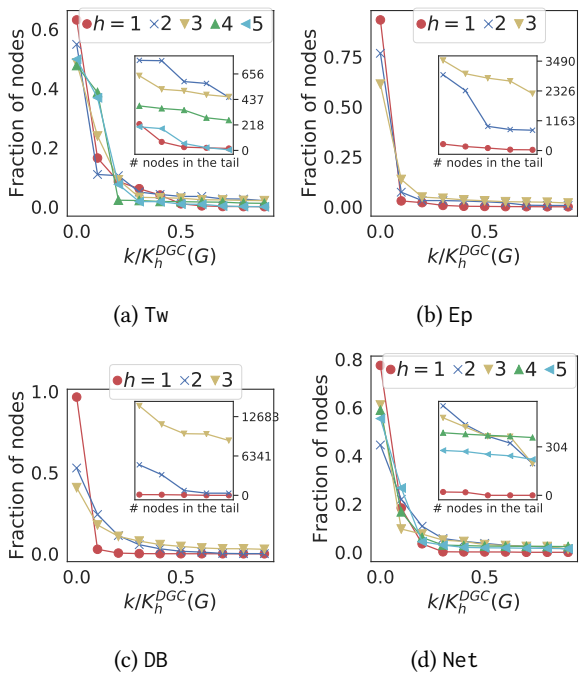

Figure 8: Fraction of nodes per normalized distance-generalized core-index $k/K_h^{DGC}(G)$, for varying h. Insets zoom in the tail of each distribution, showing the exact number of nodes in the last quartile of $k/K_h^{DGC}(G)$.

solving the identification of seeds within a small subnetwork; this would mostly happen when the chosen h approaches the average path length of the network, therefore more nodes fall into the same cores. However, in practice, $h \in \{2, 4\}$ turned out to be the most effective choice to concentrate the identification of a relatively large seed-set within the inner-most generalized core. As one rule-of-thumb, a proper setting h is the one leading to observe the tail in the distribution of generalized core-index as corresponding to a fraction of nodes comparable with the budget for the seed-set to be discovered. Nonetheless, it emerges an interesting opportunity for a theoretical investigation of relations between h and structural characteristics of the network, which we leave as future work.

6.4 Individual influence-spreading ability

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

Figure 9 shows that a high (h, k)-core-index is in general a more reliable indicator of the influence a node can individually produce. In fact, in many cases, nodes having higher (h, k)-core-index exhibit higher influence potential. By contrast, such nodes are not necessarily those with the highest core-index according to k-core decomposition. Also, it should be noted the inner cores detected by

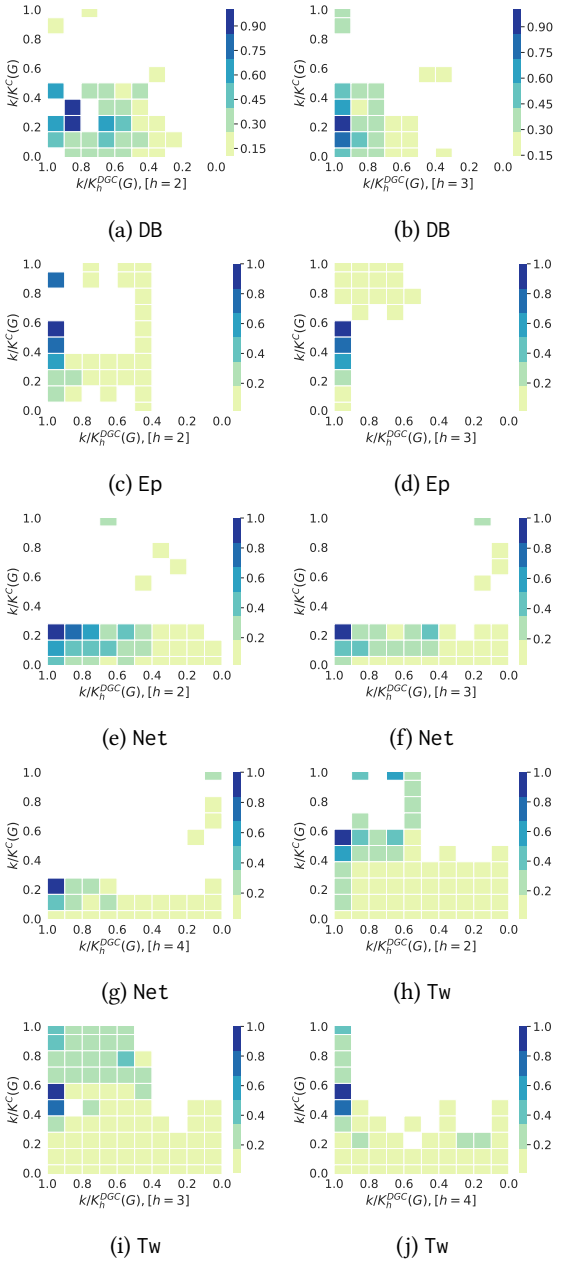

Figure 9: Average spread of nodes w.r.t. selected combinations of k-core-index (y-axis) and (h, k)-core-index for a particular choice of h (x-axis). The expected spread of each node is computed by considering the node as a singleton seed-set. Darker colors correspond to higher normalized spread.

k-core decomposition ($h = 1$) are very different from the ones corresponding to higher values of h. In fact, many nodes with low/mid core-index turn out to have a very high (h, k)-core-index.

To sum up, for an appropriate value of h, nodes in the inner-most cores are always the ones having the highest influential-spread potential, either as singletons and as groups (§6.1). This outstanding

result clearly highlights the opportunity of exploiting a distance-aware core decomposition for effectively solving top-influencer identification problems that, while not being necessarily under the IM framework, would avoid trapping into an under/over estimation of cumulative spread of a set of nodes that is a typical of any top-s search centrality-based heuristic approach.

7 CONCLUSIONS

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

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

Supplemental material: source codes, preprocessed data used in the evaluation, as well as additional experimental results can be found at: *http://people.dimes.unical.it/andreatagarelli/cores4im/*.

REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2008.

[2] J. Bae and S. Kim. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A: Statistical Mechanics and its Applications*, 395:549–559, 2014.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *Proc. Int. Conf. on Web Search and Web Data Mining (WSDM)*, 2011.

[4] N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM)*, pages 33–42, 2013.

[5] V. Batagelj and M. Zaversnik. An O(m) algorithm for cores decomposition of networks. *CoRR*, cs.DS/0310049, 2003.

[6] F. Bonchi. Influence propagation in social networks: A data mining perspective. *IEEE Intell. Informatics Bull.*, 12(1):8–16, 2011.

[7] F. Bonchi, F. Gullo, A. Kaltenbrunner, and Y. Volkovich. Core decomposition of uncertain graphs. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 1316–1325, 2014.

[8] F. Bonchi, F. Gullo, B. Mishra, and D. Ramazzotti. Probabilistic causal analysis of social influence. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1003–1012, 2018.

[9] F. Bonchi, A. Khan, and L. Severini. Distance-generalized core decomposition. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 1006–1023, 2019.

[10] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Maximizing social influence in nearly optimal time. In *Proc. ACM-SIAM Symp. on Discrete Algorithms (SODA)*, pages 946–957, 2014.

[11] A. Caliò, R. Interdonato, C. Pulice, and A. Tagarelli. Topology-driven Diversity for Targeted Influence Maximization with Application to User Engagement in Social Networks. *IEEE Trans. Knowl. Data Eng.*, 30(12):2421–2434, 2018.

[12] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2008.

[13] T. L. Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proc. Int. Conf. on World Wide Web (WWW)*, 2010.

[14] C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. D-cores: measuring collaboration of directed graphs based on degeneracy. *Knowl. Inf. Syst.*, 35(2):311–343, 2013.

[15] P. Govindan, C. Wang, C. Xu, H. Duan, and S. Soundarajan. The k-peak decomposition: Mapping the global structure of graphs. In *Proc. ACM WebConf*, 2017.

[16] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In *Proc. ACM Int. Conf. on Web Search and Data Mining (WSDM)*, 2010.

[17] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 21(2):256–276, 2006.

[18] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 137–146, 2003.

[19] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.

[20] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *Trans. Web (TWEB)*, 1(1), 2007.

[21] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 420–429, 2007.

[22] J. Leskovec, A. Singh, and J. M. Kleinberg. Patterns of influence in a recommendation network. In *Proc. Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)*, 2006.

[23] Q. Ma and J. Ma. Identifying and ranking influential spreaders in complex networks with consideration of spreading probability. *Physica A: Statistical Mechanics and its Applications*, 465:312–330, 2017.

[24] F. Malliaros, M.-E. Rossi, and M. Vazirgiannis. Locating influential nodes in complex networks. *Scientific Reports*, 6, 2016.

[25] Y. Mehmood, N. Barbieri, F. Bonchi, and A. Ukkonen. CSI: community-level social influence analysis. In *Proc. European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 48–63, 2013.

[26] H. T. Nguyen, M. T. Thai, and T. N. Dinh. Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 695–710, 2016.

[27] D. M. Romero, B. Meeder, and J. M. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. Int. Conf. on World Wide Web (WWW)*, 2011.

[28] K. Saito, R. Nakano, and M. Kimura. Prediction of information diffusion probabilities for independent cascade model. In *Proc. Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES)*, 2008.

[29] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269–287, 1983.

[30] Y. Tang, Y. Shi, and X. Xiao. Influence maximization in near-linear time: A martingale approach. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 1539–1554, 2015.

[31] Y. Tang, X. Xiao, and Y. Shi. Influence Maximization: Near-optimal Time Complexity Meets Practical Efficiency. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 75–86, 2014.

[32] J. Wang and J. Cheng. Truss decomposition in massive networks. *Proc. VLDB Endow.*, 5(9):812–823, 2012.

[33] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini. The role of information diffusion in the evolution of social networks. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD)*, pages 356–364, 2013.

[34] Y. Yuan, A. Alabdulkareem, et al. An interpretable approach for social network formation among heterogeneous agents. *Nature Communications*, 9(1):4704, 2018.