PAPER • OPEN ACCESS

Influence maximization under limited network information: seeding high-degree neighbors

View the article online for updates and enhancements.

Influence maximization under limited network information: seeding high-degree neighbors

Jiamin Ou[1,*] ⓘ, Vincent Buskens[1,2], Arnout van de Rijt[3] and Debabrata Panja[2,4] ⓘ

1 Department of Sociology, Utrecht University, Padualaan 14, 3584 CC Utrecht, The Netherlands
2 Centre for Complex Systems Studies, Utrecht University, Minnaertgebouw, Leuvenlaan 4, 3584 CE Utrecht, The Netherlands
3 Department of Political and Social Sciences, European University Institute (EUI), Via dei Roccettini, 9, 50014 San Domenico di Fiesole (FI), Italy
4 Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands
* Author to whom any correspondence should be addressed.

E-mail: j.ou@uu.nl

Abstract

The diffusion of information, norms, and practices across a social network can be initiated by compelling a small number of seed individuals to adopt first. Strategies proposed in previous work either assume full network information or a large degree of control over what information is collected. However, privacy settings on the Internet and high non-response in surveys often severely limit available connectivity information. Here we propose a seeding strategy for scenarios with limited network information: Only the degrees and connections of some random nodes are known. This new strategy is a modification of 'random neighbor sampling' (or 'one-hop') and seeds the highest-degree neighbors of randomly selected nodes. Simulating a fractional threshold model, we find that this new strategy excels in networks with heavy tailed degree distributions such as scale-free networks and large online social networks. It outperforms the conventional one-hop strategy even though the latter can seed 50% more nodes, and other seeding possibilities including pure high-degree seeding and clustered seeding.

1. Introduction

Networks provide conduits for the propagation of information [1], social norms [2] and pathogens [3]. For enhancing the efficacy of propagation in a human interaction network, as well as for effectively inhibiting propagation, it is often useful to locate a core set of nodes—i.e. *seeds*—whose triggering can maximize the expected total number of nodes that is eventually activated, or, once immunized, would reduce propagation the most [4, 5]. Public health interventions pertaining to vaccination [6], dissemination of health-relevant behaviours [7, 8], and viral marketing that target a small set of influencers [9] rely on this principle.

Strategies proposed in prior work either assume full network information [4, 10–15] or a large degree of control over what network information is collected [16–22]. However, privacy settings on the Internet and high non-response in surveys often severely limit available connectivity information. Strategies that require complete or specific network information can then not be practically implemented. This calls for alternative strategies that can also be employed in more barren scenarios commonly confronted by researchers and practitioners, where the majority of nodes and ties is unknown and where control over what information is available is lacking. We thus ask: How can one effectively seed a contagion when information on the network is limited?

In the present paper we study the effectiveness of various seeding strategies under conditions of limited network information, namely when only the degrees and connections of some random nodes are known. We do so by simulating a fractional threshold model of diffusion on a range of both empirical and synthetic networks. All seeding strategies we simulate use only the limited network information available. For each strategy we assess the ultimate fraction of activated nodes, averaged across simulation runs.

We find that one seeding strategy stands out: Seeding high-degree neighbors. This strategy builds on the friendship paradox that the neighbors of a randomly selected node have higher expected degrees than that node itself [6, 7, 23–26]. It samples a subset of nodes, selects one random neighbor of each node, then uses the highest-degree nodes among the set of neighbors as seeds for the contagion process. We find that this strategy by-and-large (with some noteworthy exceptions) achieves greater overall propagation than other strategies, including random neighbor seeding, high-degree seeding and clustered seeding. We demonstrate that the comparative efficacy of the strategy is maintained even when its competitors are advantaged through modest increases in the number of seeds they can draw on. The success of the strategy may be explained by the fact that it achieves a good balance between its seeds being influential while at the same time being scattered broadly.

The paper is organized as follows. In section 1.1 we review the existing theoretical approaches and current practices in seeding with imperfect knowledge of the network. In section 1.2 we formally introduce the high-degree neighbor seeding strategy, which we refer to in short as one-hopHD, as well as other seeding strategies working under similar limits to network information. In section 2 we introduce methods and data concerning the underlying networks, contagion models, and the settings for simulations that are used to evaluate the seeding strategies. In section 3.1 we discuss the seeding outcomes of one-hopHD and other seeding strategies in three typical synthetic networks and its implications. In sections 3.2–3.4 we analyze the simulation results in 10 real-world social networks, with varying sizes from 100 to 7600, and demonstrate that the one-hopHD strategy outperforms other seeding strategies with similar network information, especially for networks with heavy tailed degree distributions such as scale-free networks and large online social networks. We end with a conclusion section 4.

1.1. Related work: existing theoretical approaches and current practices

Theoretical studies have shown that identifying the core set of nodes under full network information is NP-hard, meaning that a solution is not guaranteed within polynomial time. Hence, approximation algorithms or heuristics have been developed to provide a solution that is not necessarily the best, but still are reasonably effective [4, 10–15]. These generally assume full knowledge of the network, that is, the complete network structure in terms of which node is connected to which others (i.e. the *edge list*) is known. For example, a recent study demonstrated that a new topological measure for identifying the central individuals best suited for spreading behaviors that requires peer reinforcement outperformed other strategies [15]. It calculates the set of all possible paths connecting every two nodes, and thus requires that the complete edge list is available [15]. As on-the-ground seeding experiences are accumulating, progressively more studies recognize that such full network information is often very expensive, if not impossible, to collect [27–30]. Surveys typically exhibit high non-response rates, so that information about social networks is only available for a minority of respondents. The rising concern of data privacy makes complete network information even more difficult to collect, both online and offline. This renders seeding without full network information a question of rising concerns and practical significance.

Thus far, related theoretical studies on seeding strategies with limited network information have appeared under a variety of names, e.g. *seeding with imperfect network information*, *seeding with partial network information*, *seeding in a partially observable network* or *influence maximization problem for unknown networks*. Based on the type of missing network information and the scalability of the envisaged algorithms, these strategies can be put into two broad categories.

The first category concerns approximation algorithms or heuristics developed for full network information to networks with partially disclosed structure [16–22]. The central idea is to efficiently disclose part of the network structure, i.e. the edge list of some nodes (or 'subnetwork') so that the established algorithms or heuristics for the full network can be used and the seeding outcomes with full network information can be approximated [21, 22]. Inspired by degree-based heuristic algorithm, Mihara *et al* [16, 17] used a snowball sampling strategy to reveal the network structure and provide seeds. It starts from a small set of nodes, each of which will disclose their degrees and list of connections. Then the node with highest expected degree from the given information is probed again to reveal its degree and list of connections. Such a process is repeated a number of times and the highest degree nodes after all the probing is used as seeds. Similarly, Eckles *et al* [20] chose a set of random nodes and asked them to reveal their connections with a probability. The revealed connections will go through the same process again to reveal more of the network. It results in a partly revealed network that can be used as the input for the algorithms developed for full network information [10, 14]. While letting go of the assumption of full network information, these studies still assume that there is large degree of control over what network information is collected, e.g. through snowball sampling [16, 17] or probabilistic random walks [20], and pose nontrivial demands on available network information [20–22]. This renders it impossible to use them for seeding in the scenarios where limits on available network information are predetermined.

The second category dives into stochastic seeding with more realistic assumptions of network information, which is relevant to our work in this paper. Public health interventions, marketing campaigns and other real-life problems typically have access to very limited network information such as degrees (i.e. how many connections one has) and random connections of some nodes. Under the constraint of such limited information, a few seeding strategies are used. Guided by the intuition that individuals having many connections are likely to be influential, one can choose the highest degree individuals among those whose degrees are known [7, 8]. Alternatively, one can use a random seeding strategy to blindly seed as many individuals as possible [31].

Within the second category, one approach that has received substantial attention in both theoretical studies and real world practices is the one-hop strategy [7, 24, 25, 32]. Depending on the context, it has been named 'acquaintance vaccination' in epidemiology [6], 'nomination' [7] or 'social referents' [23]. It is inspired by the phenomena discovered by Feld [24] that individuals are likely to have fewer friends than their friends do, on average. Therefore, by asking a random node to disclose randomly one of its connections, one can reach a node with a higher degree in most cases. Such a phenomena has been verified in general networks with various degree distributions [32]. Compared to random seeding, one-hop does require some network information, but only at a minimum level. In practice, it is easily achieved by asking random individuals to nominate one of their friends. This strategy has received a great deal of attention in the vaccination and immunization studies [32–34], and is one of the few seeding strategies used to promote public health interventions such as clean water, nutrients [7], maternal and child health [26] and combat school bullying [23]. In a field experiment involving 5773 households in rural Honduras, Kim *et al* 2015 [7] found that interventions targeting friends of randomly selected individuals were more effective than random seeding, resulting in higher adoption rates of nutritional interventions and water purification. Using a similar 'friendship nomination' technique, another experiment to promote behaviour change in maternal and child health is undergoing for 30 000 people, 176 villages in Western Honduras [26]. Paluck *et al* 2016 [23] used a variant of one-hop strategy to reduce bullying in schools. Followed by friends nominations, they targeted the students who had been nominated most times to combat school bullying.

The current study explores a modification of the one-hop strategy [7, 24–26, 32–34]. Specifically, we investigate how the one-hop strategy would perform if we identify higher degree nodes in a better way. In essence, the one-hop strategy adds one extra step to random seeding, which traces back from random nodes to each of their random connections. It has two underlying principles. The primary principle is to target higher degree nodes by random nomination. The second one, which is more hidden, is that it can distribute seeds broadly throughout the network because of its random selection of initial nodes, though to a lesser extent than random seeding. This study investigates how the one-hop strategy performs if its first principle is enhanced. We formalize and evaluate a variant of one-hop seeding strategy (one-hopHD), i.e. high degree seeding of random connections, which is in line with the idea behind the empirical experiment in Paluck *et al* 2016 [23].

Interestingly, a recent study argues that network information is not really valuable for seeding, because a small increase in the number of random seeds can catch up with the optimal outcomes with full network information anyway [31]. However, this observation applies to the independent cascade model studied in [31], but not to the fractional threshold model we study (details in section 2.3). In the independent cascade model, contagions are 'simple', that is, people need only one exposure from their connections to pick up the concerned attribute (i.e. adopt), such as the spread of knowledge, easily-convincing rumors and pathogens. For many social phenomena, the contagion is arguably 'complex' [35, 36]: People need multiple exposures from their connections before they adopt. Complex contagion theory argues that the diffusion of many social norms are complex, for which individuals require connection with multiple activated peers (i.e. 'social reinforcement') before they become activated [37–39]. Complex contagion applies to most behaviours that are risky, costly or need to fight against habits such as rioting, healthier diet, more regular exercise, and the purchase of new products. In numerical simulations, the idea of complex contagion can be realized by the fractional threshold model and other variants. In the fractional threshold model studied in this paper, contagions are complex as they require that a threshold number of network neighbors adopt first before a focal individual adopts. Existing studies suggest that information on degree and network neighbors of even a small number of nodes can be useful for seeding as they provide better cascade outcomes than random seeding when contagion is complex [22, 23]. However, it is not clear that if random seeding could easily reproduce the cascade outcomes of other seeding strategies by just having a few more seeds, as the case in simple contagion [31]. In other words, the value of limited network information and how to leverage them for seeding under complex contagion is still unclear.

Under complex contagion and the constraint of limited network information, we explore a number of seeding strategies, finding a particular variant of the one-hop strategy promising: If we add an extra step to the one-hop strategy, that is, we choose the highest degree nodes from random connections (hereafter called 'one-hopHD'), we obtain the best seeding outcome compared to other seeding strategies using similar levels of network information. We find this to be the case across a broad range of empirical and synthetic social

networks. The added value of this new strategy remains largely unchanged when the number of seeds by the random strategy or one-hop strategy increases significantly. This new strategy is in line with the idea of an anti-bullying campaign used in US [23]: students in 56 schools first nominated a few of their friends, and among those who had been nominated, the top 10% who received the most nominations were chosen as seeds to take a public stance against conflicts in their schools, resulting in a 30% drop of reported conflict levels. The core of such an anti-bullying strategy is high-degree seeding of random connections. To the best of our knowledge, this has not been formally proposed as a strategy, neither has it been addressed and evaluated despite a promising empirical evidence.

The effectiveness of the proposed one-hopHD strategy is evaluated through comparison to other seeding strategies with similar assumptions on the availability of information on the network, i.e. only the degrees and connections of some random nodes. We use four strategies for comparison. The first three are the above-mentioned high-degree seeding, conventional one-hop strategy and random seeding. Specifically, high-degree seeding assumes only the degrees of a small percentage of nodes are given and chooses the ones with highest degrees as seeds. The fourth strategy is clustered seeding, based upon limited network information on degrees and connections. Since complex contagion requires multiple exposure to overcome the threshold, clustered seeding initiates the diffusion process by activating a group of directly connected nodes. In this way, other nodes connected to this group of seeds are likely to receive multiple exposure and become activated as well [40, 41]. This seeding strategy is realized by activating nodes that are directly connected to the most central node and iterating the same process until the number of seeds have been reached [15]. Therefore, the fourth strategy will activate the node with highest degree, from nodes with given degrees, and its direct connections. More details of the strategies used for comparison can be found in section 1.2.

1.2. The proposed seeding strategy, reference strategies and their required network information

To formally define the one-hopHD strategy, we first introduce some notation. Let $G = (V, E)$ be the underlying network, in which $V = \{v_1, v_2, \ldots, v_N\}$ is the set of N nodes, and $E = \{e_1, e_2, \ldots, e_M\}$ is the set of all M edges. The set of seeds is defined as $S = \{d_1, d_2, \ldots, d_s\}$, which is a subset of s nodes from V ($S \subset V$).

In the one-hop strategy, one first draws s random nodes from V to form a subset $R = \{r_1, r_2, \ldots, r_s\}$. Each (random) node discloses one of their connections, resulting in a set of random connections $RN = \{rn_1, rn_2, \ldots, rn_s\}$ which are directly used as seeds ($S \equiv RN$). The proposed one-hopHD strategy adds one more step to the conventional one-hop strategy, which is to seed the highest degree nodes from the set of these random connections ($S \subset RN$). To make this possible, the number of random connections RN must be larger than s. Considering the possible network information that can be attained by an affordable network survey in both online and offline environments, following the empirical experiment by Paluck *et al* 2016 [23], here we assume maximum 15% of individuals in the network will respond to the survey and disclose their connections. Moreover, instead of revealing a few of their connections as in the experiment [23], we assume that the respondents will provide only one connection, i.e. $R = RN = $ 15% of N. For all the nodes in RN, we further assume their degrees will be known by another round of survey. Then, from all the nodes in RN, the top s highest degree nodes will be chosen as seeds.

The network information required by the one-hopHD is accessible in many real-world settings. First, this approach relies on whom the random connections of 15% $\times N$ random nodes are, and the degrees of these random connections. Note that the 15% $\times N$ random nodes do not need to disclose all their neighbors. Instead, they are asked to nominate *only one* of their friends in the network, similar to the well-established one-hop strategy. In this way, this new strategy circumvents the thorny issue of collecting complete contact list of the respondents, which can be difficult as the response rate can be extremely low and unreliable due to the recall error. Second, it demands minimum control over what network information is collected, as the first sample on those people who nominated their neighbors can be completely random, and only degrees are asked for the nominated friends in the second round. This is in stark contrast to snowball sampling and other network probing techniques that require deliberate control to collect the information along certain paths (usually much longer than 2 path lengths) within the network.

To evaluate the effectiveness of one-hopHD, other seeding strategies with similar constraints on the accessibility of the network information are studied and used as references. The referenced strategies have access to only the degrees and connections of some nodes and have no or minimum control on which nodes they are. Under these two constraints, there are still some differences regarding the network information required by the referenced strategies. The differences, and the justifications of the comparison to one-hopHD, are explained in the following paragraphs. We argue that one-hopHD would be an effective strategy if it outperforms other referenced strategies as specified below,

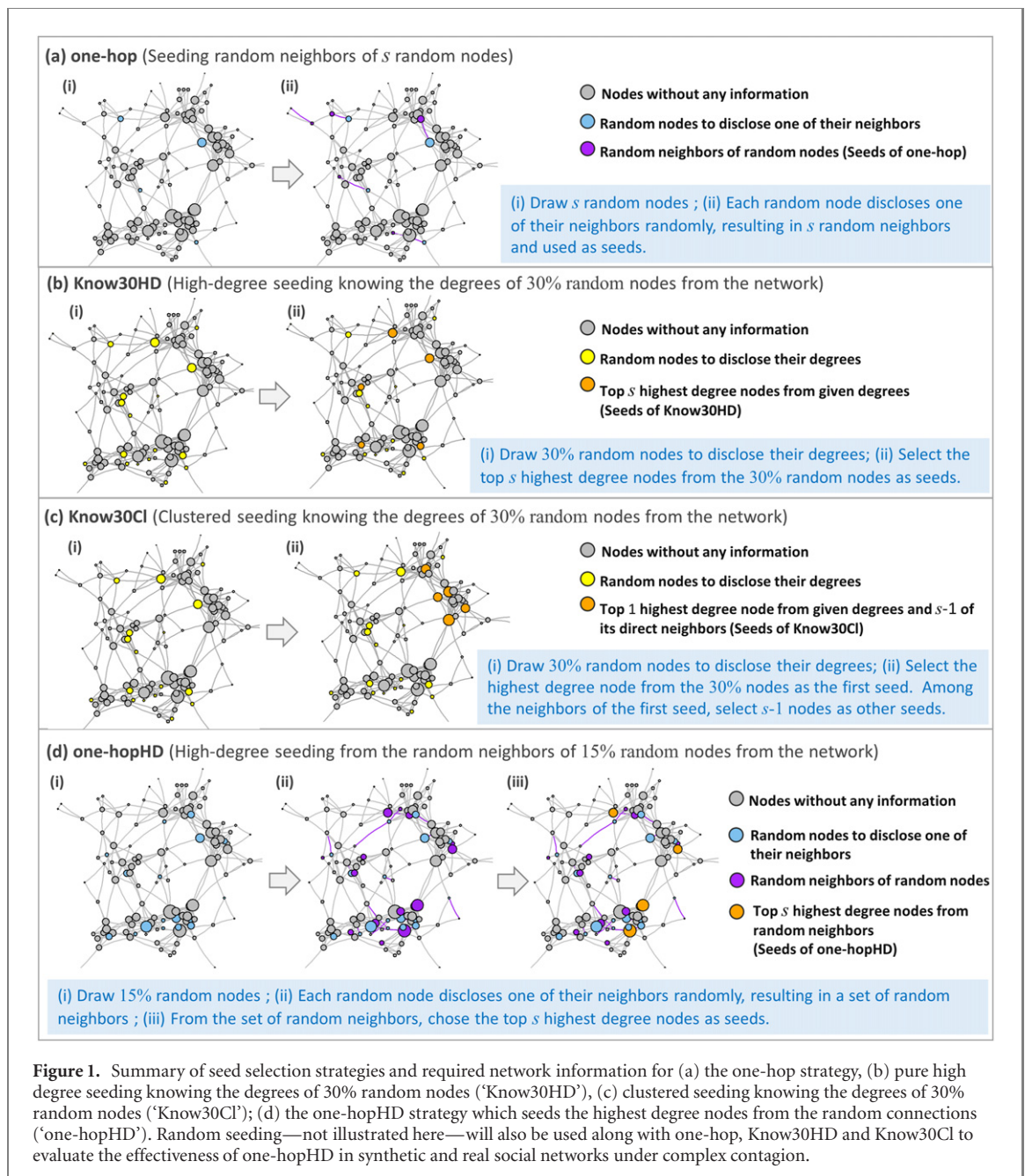(a) The one-hopHD strategy is more efficient than random seeding with an identical seed set size ('Random').

Figure 1. Summary of seed selection strategies and required network information for (a) the one-hop strategy, (b) pure high degree seeding knowing the degrees of 30% random nodes ('Know30HD'), (c) clustered seeding knowing the degrees of 30% random nodes ('Know30Cl'); (d) the one-hopHD strategy which seeds the highest degree nodes from the random connections ('one-hopHD'). Random seeding—not illustrated here—will also be used along with one-hop, Know30HD and Know30Cl to evaluate the effectiveness of one-hopHD in synthetic and real social networks under complex contagion.

(b) The one-hopHD strategy is more efficient than random seeding even if the latter can seed twice as many ('Random2').

(c) The one-hopHD strategy is more efficient than the conventional one-hop strategy with an identical seed set size (figure 1(a), labeled 'one-hop').

(d) The one-hopHD strategy is more efficient than the conventional one-hop strategy even when the latter can seed 50% more ('one-hop1.5').

(e) The one-hopHD strategy is more efficient than pure high-degree seeding with an identical seed set size. Specifically, the pure high-degree seeding is based upon partial network information, i.e. knowing the degrees of 30% random nodes. It then chooses the top *s* highest degree nodes among the 30% nodes with given degree as seeds (figure 1(b), labeled 'Know30HD').

(f) The one-hopHD strategy is more efficient than clustered seeding with an identical seed set size. Specifically, the clustered seeding is based upon partial network information, i.e. knowing the degrees of 30% random nodes. It then chooses the top highest degree node among the 30% nodes whose degrees are known as the first seed. Among the connections of the first seed, $s - 1$ nodes will be chosen randomly to complete the seed set [15] (figure 1(c), labeled 'Know30Cl').

The strategy Random requires no network information. It is used here to reflect the seeding outcome that can be achieved without any knowledge of the network structure. While we expect that the seeding outcome of one-hopHD will be better than the strategy Random if both of them start from the same number of seeds, we want to test if the advantage provided by one-hopHD seeding can be easily compensated by seeding more nodes using the strategy Random, as observed in a previous study in the context of simple contagion [31]. If the one-hopHD could provide better seeding outcomes even though the random strategy can seed twice as many (strategy Random 2), it will suggest that limited network information (as those used by one-hopHD) are critical to diffuse complex contagion and the same result cannot be achieved by just having more seeds.

In comparison to one-hopHD, the conventional one-hop strategy demands only the random connections of random nodes (See figure 1(a)), thus demanding less information than the former. It is expected that one-hopHD could outperform the conventional one-hop if they seed the same amount of nodes. To test the superiority of one-hopHD, we will also allow the conventional one-hop strategy to have more seeds (50% increase), similarly to the test against strategy Random. The conventional one-hop strategy with 50% more seeds will be called 'one-hop1.5'.

One-hopHD selects the higher degree nodes from the random connections (see figure 1(d)). One can argue that seeds can be simply selected from the nodes of higher degrees, without searching for the random connections. We therefore compare one-hopHD to pure high-degree seeding: one-hopHD requires the random connections of 15% of the nodes and the degrees of 15% of the nodes, in comparison, we assume that Know30HD has access to the degrees of 30% of the nodes (the strategy 'Know30HD') (see figure 1(b)). In other words, Know30HD has more access to the degree information as a compensation to the fact that it does not require any information of direct connections. Such compensation is considered fair as the difficulty or cost to ask a respondent, to provide one name among the contacts (the random connection used by one-hopHD) or to provide the total number of her/his contacts (the degree information used by Know30HD), are similar. By comparing one-hopHD with Know30HD, we can thus test if the combination of searching random connections and high-degree seeding is more efficient than pure high-degree seeding.

Last but not least, the strategy 'Know30Cl' is used to represent one important seeding principle when it comes to complex contagion: clustered seeding. In complex contagion, seeds should be connected instead of scattered in order to provide the multiple exposures ('social reinforcement') to adjacent nodes [40, 41]. The strategy Know30Cl follows the same clustered seeding procedure as in other studies [15], but assumes that only part of the network can be known. As shown in figure 1(c), this strategy first uses the degrees of 30% of the nodes and chooses the one node with highest degree, similar to Know30HD. But it differs from the latter, by requiring the highest-degree node to disclose all its neighbors. Then, this node with highest degree and its $s - 1$ neighbors will be chosen as seeds. In comparison to one-hopHD and other referenced strategies, Know30Cl is the most demanding one on network information. The information required by Know30Cl imposes that the seeds are not only high-degree but also directly connected to each other. By contrast, one-hopHD does not guarantee the seeds are connected. By comparing one-hopHD to Know30Cl, we can test if one-hopHD can overcome the social reinforcement required by the complex contagion.

In summary, the network information required by the strategy Random is the least, followed by the strategy of conventional one-hop. One-hopHD and Know30 use similar amount of network information.

The strategy Know30Cl has the highest demand on network information.

2. Method and data

2.1. Networks for simulation and their properties

First we use three typical synthetic networks: an Erdős–Rényi (ER) a random network, small-world (SW) network and a scale-free (SF) network for studying the performance of seeding strategies under simplified network structures. Albeit not realistic, they capture some properties of real-world social networks. Testing on these synthetic networks can therefore shed light on the performances of seeding strategies when some of the real-world network properties are pushed to the limit (which is hard to test using real-world networks alone). Specifically, the ER random network is characterized by short path lengths, low clustering and moderate degree variances following a normal distribution. It is generated by the Erdős–Rényi–Gilbert model [42], in which every two nodes are connected with the same probability (we use 0.1). The SW network reproduces the short path length and high clustering of real-world networks, but with a relatively homogeneous degree distribution. The SW network is built from a regular lattice, followed by a rewiring probability (we use 0.025) to create shortcuts to mimic the small-world effect [43]. The SF network differs from the ER and SW networks with a highly heterogeneous degree distribution that follow a power-law decay, extremely low clustering coefficient and long path lengths. It is generated from the Barabási–Albert (BA) model [44, 45] with linear preferential attachment. All the synthetic networks are of the same size ($N = 100$).

Table 1. Details of the 10 social networks from real world.

Network name	Description	Size (N)	Avg. shortest path length (D)[a]	Clustering coefficient (CC)	Modularity[b] (M)
High school	Connections and friendship relations between students in a high school [46]	128	4.0 (2.4)	0.48	0.74
Primary school	Connections between the children and teachers in a primary school [47]	236	1.8 (1.3)	0.43	0.34
Infect	Daily connection networks at a Science Gallery [48]	221	3.5 (2.2)	0.30	0.83
Facebook	Ego-based network of Facebook users [49]	4039	3.7 (2.5)	0.52	0.81
Bitcoin	Bitcoin alpha web of trust network [52, 53]	3782	3.5 (2.2)	0.08	0.42
API	Social network of LastFM users from Asia [54]	7623	5.2 (3.5)	0.18	0.77
Wiki	Wikipedia who-votes-on-whom network [50, 51]	7066	3.5 (1.9)	0.13	0.37
P2P	Gnutella peer to peer network [55, 56]	6301	4.6 (3.2)	0.02	0.34
CA	Collaboration network from the arXiv covering authors from general relativity and quantum Cosmology category [55]	4158	6.0 (4.5)	0.60	0.79
Email	Email network about all incoming and outgoing email between members from a large European research institution [55, 57]	1006	2.6 (1.5)	0.27	0.37

[a]Notes: number in the parenthesis indicates the average path length between the top 10% highest degree nodes.
[b]Modularity is calculated by the algorithm of community structure via short random walks (cluster walktrap) using the igraph package in *R*.

We also use ten real-world social networks of various sizes (ranging from 100 to 7600 nodes) and properties. Among them, 3 are offline social networks with less than 250 nodes, from the SocioPatterns dataset (https://sociopatterns.org). They include a friendship network between high school students [46], a connection network between the children and teachers in a primary school [47], and a daily connection network collected from a Science Gallery [48]. The other seven networks are online social networks containing 1000 to 7600 nodes, such as those of Facebook [49], Wikipedia [50, 51], Bitcoin [52, 53] and others. Most of these networks have short path lengths, but differ in other network properties such as clustering coefficient and modularity. Descriptions of the networks and their properties can be found in table 1.

2.2. The threshold model of network diffusion

We employ the fractional threshold model [58, 59] to study the diffusion in the network. In this model, each node has a state value v, which is either 0 (*inactive*) or 1 (*active*). All nodes except the seeds start from the inactive state, and once activated, they remain so in the following time steps. At each time step, every node checks the weighted state value of their connections, which is the weighted percentage of activate connections. Assuming that all connections (node j) of node i have the same weight for node i, the weighted state value of its connections w_i at time t is

$$w_i(t) = \frac{\text{Number of activated connections of node } i \text{ at time } t}{\text{Total number of connections of node } i}, \tag{1}$$

Table 2. Simulation procedures for a given network, seed set size and threshold.

Simulation procedures

1. Choose an underlying network.

2. Prepare the seed sets for a given size of *s*.

If *seeding strategy* = 'Random',
       (i) Randomly select *s* nodes to form a set of random nodes as *R*;
       (ii) Record the nodes in *R* as seeds;
       (iii) Repeat (i) and (ii) for 10 000 times to generate a seed matrix with a dimension of 10 000 $\times$ *s*.
If *seeding strategy* = 'one-hop',
       (i) Randomly select *s* nodes to form a set of random nodes as *R*;
       (ii) Each node in *R* discloses one of their connections to form a set of random connections as RN;
       (iii) Record the nodes in RN as seeds;
       (iv) Repeat (i)–(iii) for 10 000 times to generate a seed matrix with a dimension of 10 000 $\times$ *s*. If *seeding strategy* = 'Know30HD',
       (i) Randomly select 30% $\times$ *N* nodes to form a set of random nodes as *R*;
       (ii) Each node in *R* discloses its degree;
       (iii) Select the top *s* highest degree nodes from *R* as seeds;
       (iv) Repeat (i)–(iii) for 10 000 times to generate a seed matrix with a dimension of 10 000 $\times$ *s*.
If *seeding strategy* = 'Know30Cl',
       (i) Randomly select 30% $\times$ *N* nodes to form a set of random nodes as *R*;
       (ii) Each node in *R* discloses its degree;
       (iii) Select the highest degree node from *R* as first seed d_1;
       (iv) Seed d_1 randomly discloses $s-1$ of its connections and used as remaining seeds: $d_2, \ldots, d_s$;
       (v) Repeat (i)–(iv) for 10 000 times to generate a seed matrix with a dimension of 10 000 $\times$ *s*.
If *seeding strategy* = 'one-hopHD',
       (i) Randomly select 15% $\times$ *N* nodes to form a set of random nodes as *R*;
       (ii) Each node in *R* discloses one of their connections to form a set of random connections as RN;
       (iii) Each node in RN discloses its degree;
       (iv) Select the top *s* highest degree nodes from RN as seeds;
       (v) Repeat (i)–(iv) for 10 000 times to generate a seed matrix with a dimension of 10 000 $\times$ *s*.

3. Diffusion model for a given threshold θ.

    (i) Each node is assigned with a threshold θ and a starting state value of 0;
    (ii) Change the state values of nodes to 1 according to the *k*th row of the seed matrix, in time step t_0; for every following time step, check the weighted state value of each node,s connections (w_i) and if $w_i \geqslant \theta$, change the state value of node *i* to 1; stop if the number of activated nodes at the current time step *t* is equal to the previous time step $t-1$; record the final activated number of nodes;
    (iii) Change the *k* in (ii) from 1 to 10 000 and so that all the combinations in the seed matrix are tried out to kick off the diffusion process. It therefore produces 10 000 possible final activated fractions AF_t under this current combination of network,seeding strategy, seed set size *s* and threshold θ.

Node *i* will be activated at time $t+1$ only if w_i is equal to or exceeds the predefined threshold θ_i, i.e.

$$v_i(t+1) = \begin{cases} 1, & w_i(t) \geqslant \theta_i \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

All the nodes in the network will adopt the same fractional threshold [58, 59], but we will vary the value of fractional threshold for the whole population to capture different contagiousness (more details in next section) [15]. In our simulations, the total number of activated nodes $\mathrm{AN}(t)$ at every (discrete) time step *t*, is recorded. The diffusion process is stopped when the number of activated nodes at the (final) time step t_f is equal to the previous time step $t_\mathrm{f}-1$. The fraction of activated nodes at time t_f is then the *final activated fraction* $\mathrm{AF}(t_\mathrm{f})$, given by

$$\mathrm{AF}(t_\mathrm{f}) = \frac{1}{N}\sum_{i=1}^{N} v_i(t_\mathrm{f}), \tag{3}$$

and is used as the index for evaluating the effectiveness of seeding strategies.

2.3. Simulation setup

We choose the synthetic and real networks introduced in section 2.1, and apply the simulation procedures described in table 2. Three further conditions need to be specified before starting the simulations. The first one is the number of seeds, *s*. Typical values of 3% $\times$ *N*, 5% $\times$ *N* and 7% $\times$ *N* will be used. In most cases, the seed set sizes of different seeding strategies will be the same. Exceptions are Random2 and one-hop1.5, in which the seed set sizes will be larger than the one-hopHD.

The second condition is of course the seeding strategy. According to the given seeding strategies, seeds will be chosen accordingly in line with the procedures in table 2. Note that each seeding strategy has its own randomness in seed selection; for instance, for a given network and a given *s*, the seeds provided by the random seeding strategy is different for each run. This is also the case for one-hop, Know30HD, Know30Cl and one-hopHD, which all start from a set of random nodes. To capture the randomness of each strategy, the seed selection process is performed 10 000 times for a given network and *s*. In other words, under each combination of network, seed set size and seeding strategy, a seed matrix with a dimension of 10 000 × *s* is generated and used as inputs for the diffusion model. Details in the generation of seed matrix can be found in table 2.

The third condition is to define the threshold of the diffusion model. For a given threshold θ, each node in the network is assigned with the same threshold value. We test a wide range of threshold values ranging from 0.15 to 0.6, with an increment of 0.05, for each network and seeding strategy. Once the above three conditions are defined, the diffusion process will be simulated.
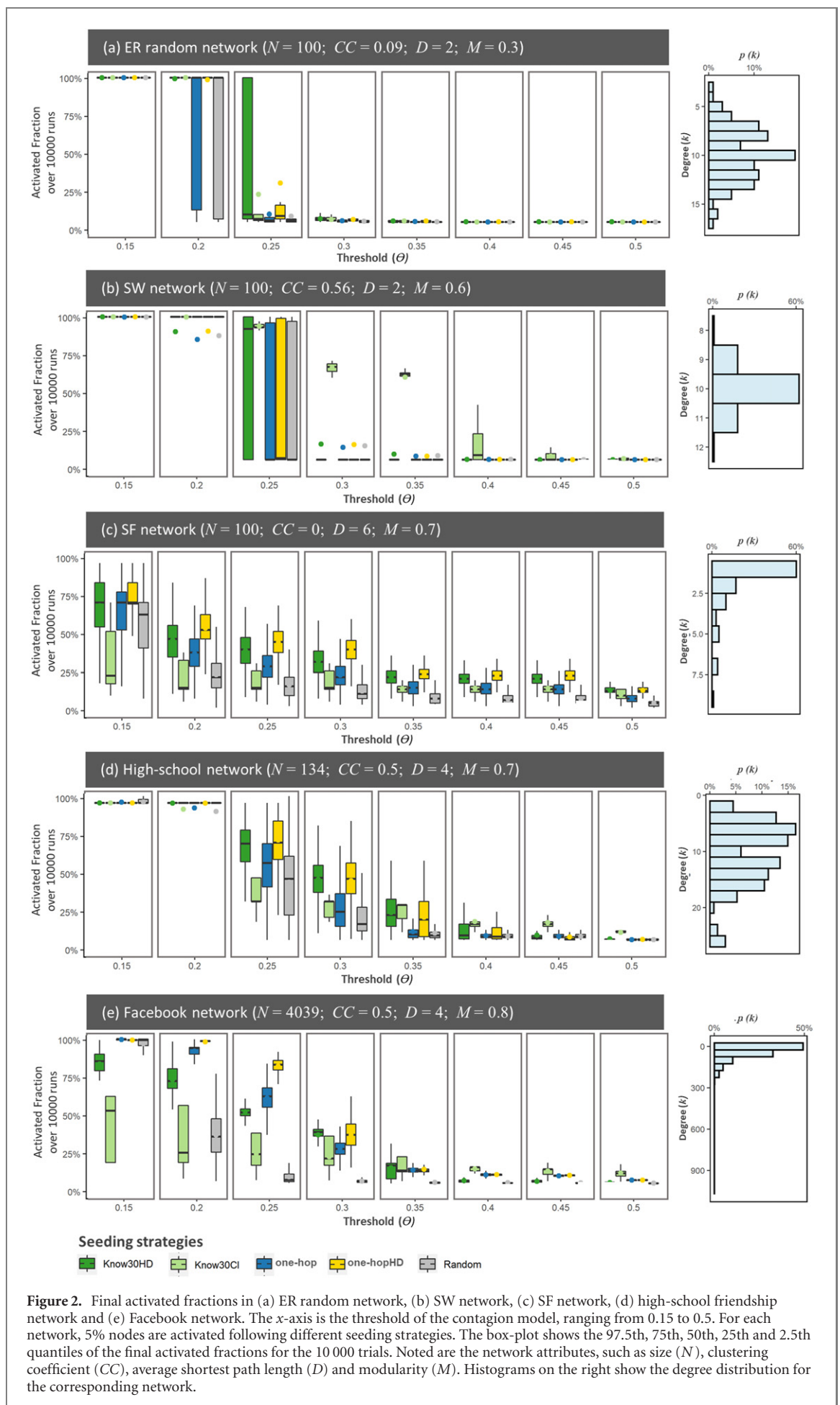
3. Results

3.1. Seeding outcomes for synthetic networks and implications for real-world networks

In figure 2 we display the simulation outcomes for five seeding strategies: one-hop, one-hopHD, Know30HD, Know30Cl and random, for three types of synthetic networks (ER, SW and SF), as well as those for two real-world networks (high school and Facebook; cf. table 1). The figure shows the activated fractions of agents under thresholds ranging from 0.15 to 0.5. Lower threshold values lead to higher activation percentages, as it must: e.g. an agent with threshold $\theta = 0.2$ and with 10 connections in a network will become activated if two or more connections were activated, in contrast to an agent with $\theta = 0.5$ and with 10 connections requiring at least five activated connections before following suit. The seed set size in figure 2 is 5%, i.e. each seeding strategy can activate 5% of the nodes to kick off the activation diffusion process. As described in table 2, for every specific threshold and seed set size and any given network, each seeding strategy has been simulated 10 000 times. The variability in the final activated fractions, due to randomness in the activation diffusion process over these 10 000 runs are captured by the error bars in figure 2.

Each synthetic network represents a unique network architecture. ER networks describe a theoretical world when human connections are made all by random chances. The networks are characterized by short path lengths but low clustering without community structures. The degree distribution, i.e. the probability distribution $P(k)$ of degree k in the networks, follows a normal distribution. The latter attribute means that the percentage of nodes having degrees much higher or lower than the mean degree k_{mean} decreases sharply with $|k - k_{mean}|$. In such (homogeneous) networks, most nodes have similar roles in accelerating or slowing down the activation diffusion process. This explains why the differences between seeding strategies are only noticeable in a narrow band of thresholds. In the example shown in figure 2(a), differences between seeding strategies are only observed when thresholds are 0.2 and 0.25. For $\theta = 0.2$, most seeding strategies except one-hop and random seeding can trigger adoption in the whole network by seeding only 5% nodes, in all the 10 000 random trials. The one-hop and random seeding strategies are the least efficient ones. For both strategies, if they can only seed 5% of the nodes, the chances for their seeds to trigger adoption in the whole network is less than 2.5%. When threshold increases to 0.25, the difficulty to pass down diffusion increases. In this case, Know30HD is the only strategy that still has some chance to reach activation fractions that are larger than 50%, while other strategies end up with final activation fractions lower than 25% in nearly all the trails. In such a synthetic network wherein every connection is made by random chance, the only significant distinguishing factor among the nodes is their degrees. Know30HD turns out to be the most efficient seeding strategy, followed by one-hopHD.

SW networks are known for their reproduction of the 'small-world' properties observed in real social networks, i.e. short path length and high clustering. Compared to ER networks, SW network has community structures with closed triads (cliques) linking nearest nodes and a few shortcuts connecting different communities [43]. Degree variation in SW networks is even lower than that in ER networks. Rewiring from a regular lattice, most nodes in SW networks used here have exactly the same degree of 10. Only the few nodes rewired to create new connections have degrees that are either 9 or 11, as shown in the degree distribution in figure 2(b), which approximates a delta distribution (each node has a degree of exactly *k*). As degree differences between nodes in SW networks are not present, Know30HD is no longer the most efficient strategy. Instead, Know30Cl stands out as the most efficient seeding strategy even for high thresholds. This is as expected since clustered seeding is designed for complex contagion under the small-world nature [40, 41]. Under the standardized community structure in SW networks, seeding a cluster of nodes is essential to overcome the thresholds of nodes connected to the same cluster [40]. Moreover, since there are clustered structures present in SW networks, connections among different clusters close to each other are plentiful [40, 43]. Such connections create channels for contagion passing from one cluster to another, and consequently, throughout the entire network.

Figure 2. Final activated fractions in (a) ER random network, (b) SW network, (c) SF network, (d) high-school friendship network and (e) Facebook network. The *x*-axis is the threshold of the contagion model, ranging from 0.15 to 0.5. For each network, 5% nodes are activated following different seeding strategies. The box-plot shows the 97.5th, 75th, 50th, 25th and 2.5th quantiles of the final activated fractions for the 10 000 trials. Noted are the network attributes, such as size (*N*), clustering coefficient (*CC*), average shortest path length (*D*) and modularity (*M*). Histograms on the right show the degree distribution for the corresponding network.

In such a network with close-knit community structure and low degree variances, one-hopHD strategy is not efficient.

SF networks mimic heavy tailed degree distribution observed often in real-world networks. The degree distribution of SF networks follows a power-law (see right panel of figure 2(c)). It results in a hub-dominated network architecture: hubs are those (few) nodes that are highly connected while most nodes having degrees that are much less than the mean. In such a synthetic network dominated by hubs, one-hopHD stands out as the most efficient seeding strategy, as shown in figure 2(c). The strategy used by one-hopHD, i.e. nomination by random nodes and then high-degree seeding, turns out to be more efficient than pure high-degree seeding by Know30HD, though the latter has access to more degree information. Clustered seeding is no longer applicable in such a network structure, which ends up with final activated fractions that are even lower than one-hop and random seeding.

These results in the synthetic networks demonstrate that seeding outcomes of the five strategies depend strongly on the architecture of the concerned network: while community structure (or high clustering) makes clustered seeding the most suitable strategy, when heavy tailed degree distribution is taken into account, one-hopHD is the most efficient one. However, real-world social networks are neither pure SW nor pure SF networks but a mixture of both heavy tailed degree distribution and community structures. For a comparison with the theoretical results from the synthetic networks, we therefore use two further real-world networks.

The first one is a high school friendship network representing a small-size ($N = 134$) social network that is very similar to the synthetic SW networks in terms of high clustering and low path length (figure 2(d)). Nevertheless, the approximately delta degree distribution for SW networks does not hold for this network, as shown as the log-normal degree distribution in the right panel of figure 2(d). Under such a mixture of high clustering and moderate degree variances, Know30HD and one-hopHD are the two most effective strategies. Both of them provide higher final activated fractions than Know30Cl in most cases. When threshold goes up to 0.35, Know30Cl outperforms these two but the margin is relatively small. The second network is a large-size ($N = 4039$) Facebook network. It has the same clustering coefficient ($CC = 0.5$) and average shortest path lengths ($D = 4$) as the high school network, but its degree distribution follows a power law. In other words, it inherits the heavy tailed degree distribution from the SF network, while maintaining high clustering and low path lengths that is similar to SW network. With high variations in node degrees, one-hopHD stands out as the most effective seeding option, leading to final an activated fraction that is much higher than those of clustered seeding (Know30Cl) and pure high-degree seeding (Know30HD). It seems that under the co-existence of heavy tailed degree distribution, high clustering and short path length, one-hopHD is a promising strategy and its performance stands out even more when the degree variance increases.

In the following subsections, we further validate the above observations for more real-world networks.

3.2. Comparison between the outcomes of one-hopHD and Know30HD strategies for real-world networks

In this subsection, we compare the seeding outcomes by one-hopHD and Know30HD in ten real-world networks and analyze if the combination of searching from random connections and then high-degree seeding is more efficient than pure high-degree seeding.

The differences between one-hopHD and Know30HD seeding outcomes in ten networks are shown in figures 3(a)–(j), for a wide range of threshold values and seed set sizes. The x-axis is the threshold value, ranging from 0.15 to 0.6. The y-axis is the seed set size within the range of 3% to 10%. The color gradient is the absolute difference between the final activated fractions (again, averaged over 10 000 runs) achieved by one-hopHD and Know30HD, in which gradients from yellow to red indicates that one-hopHD leads to higher activated fraction than Know30HD and vice versa for gradients from green to blue. Each panel in figures 3(a)–(j) corresponds to the results in a specific network, under every possible combination for seed set size and threshold. The coordinate of every color pixel within each panel corresponds to the specific seed set size and threshold value. The highlighted point in figure 3(j), for example, shows that the final activated fraction achieved by one-hopHD is larger than that of Know30HD in the API network when $\theta = 0.4$ and 5% nodes are used as seeds.

The networks used for comparison here cover offline and online social networks with various sizes and characteristics. The first three networks in figures 3(a)–(c) are offline social networks of a primary school, a high school and an exhibition, respectively. Each of them includes less than 250 nodes and has the common characteristics of a small-world network as being highly clustered and having short path length. The degree distribution follows a log-normal distribution (e.g. first two subpanels in figure 3(k)), indicating moderate degree variances. The observations we made in figure 2(d) hold true for these networks with small-world like characteristics and moderate degree variances. For example, recall that figure 2(d) corresponds to a specific case of the high school network, with seed set size of 5%, for which one-hopHD and Know30HD are more effective than the other seeding options. Among these two, Know30HD leads to slightly better seeding outcomes than one-hopHD, as shown in the green areas in figures 3(a)–(c). In other words, for small-world like networks
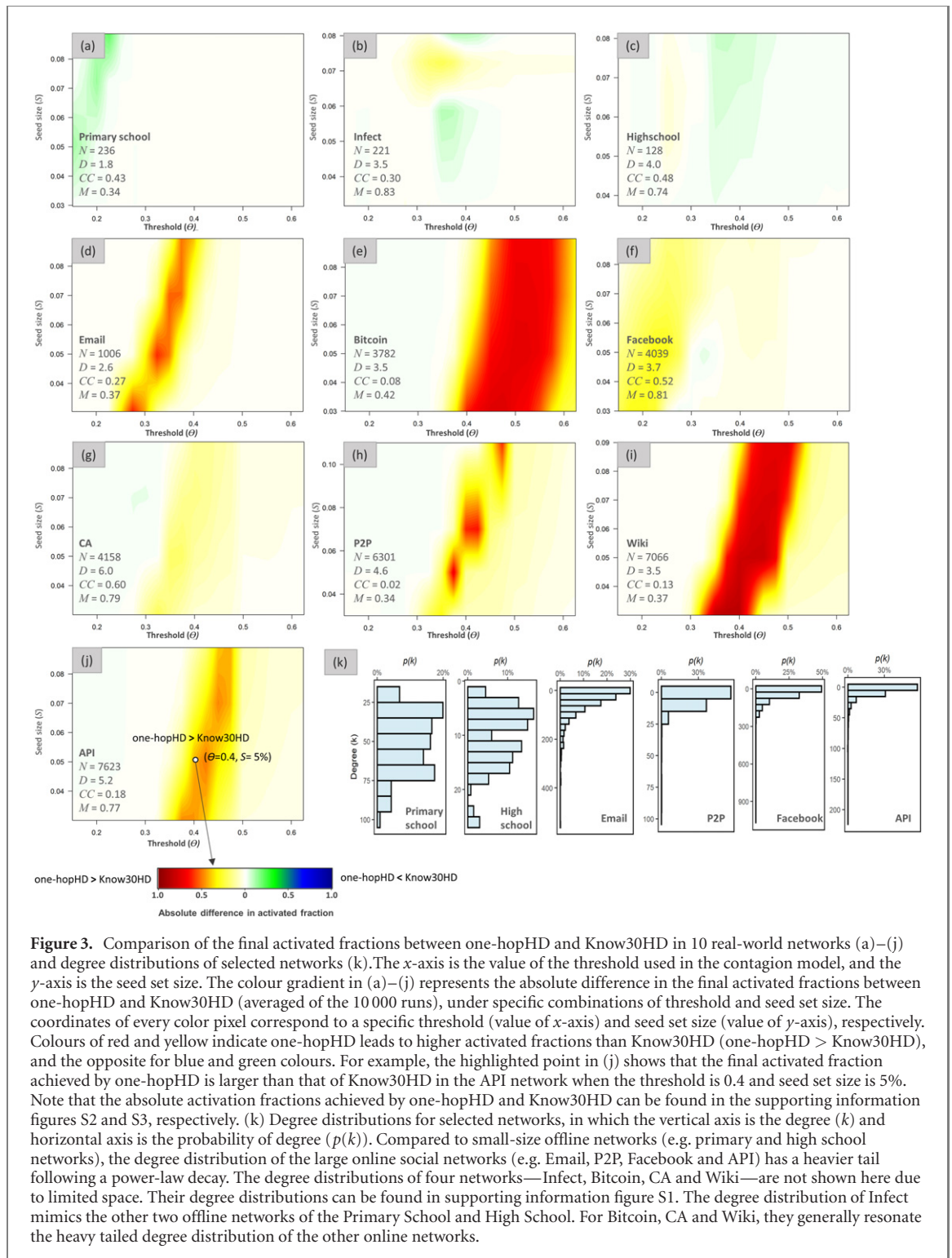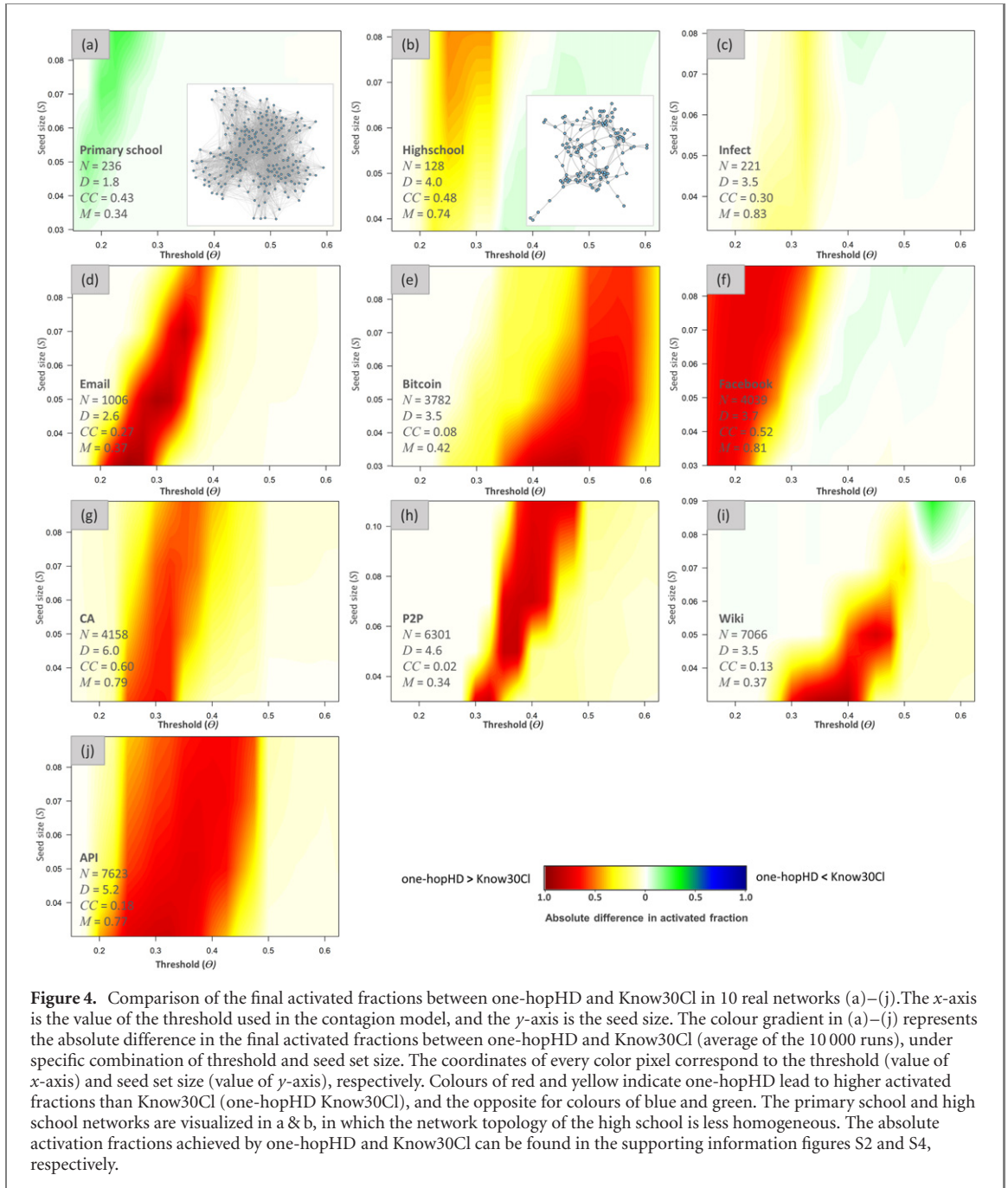
Figure 3. Comparison of the final activated fractions between one-hopHD and Know30HD in 10 real-world networks (a)−(j) and degree distributions of selected networks (k).The *x*-axis is the value of the threshold used in the contagion model, and the *y*-axis is the seed set size. The colour gradient in (a)−(j) represents the absolute difference in the final activated fractions between one-hopHD and Know30HD (averaged of the 10 000 runs), under specific combinations of threshold and seed set size. The coordinates of every color pixel correspond to a specific threshold (value of *x*-axis) and seed set size (value of *y*-axis), respectively. Colours of red and yellow indicate one-hopHD leads to higher activated fractions than Know30HD (one-hopHD > Know30HD), and the opposite for blue and green colours. For example, the highlighted point in (j) shows that the final activated fraction achieved by one-hopHD is larger than that of Know30HD in the API network when the threshold is 0.4 and seed set size is 5%. Note that the absolute activation fractions achieved by one-hopHD and Know30HD can be found in the supporting information figures S2 and S3, respectively. (k) Degree distributions for selected networks, in which the vertical axis is the degree (*k*) and horizontal axis is the probability of degree (*p*(*k*)). Compared to small-size offline networks (e.g. primary and high school networks), the degree distribution of the large online social networks (e.g. Email, P2P, Facebook and API) has a heavier tail following a power-law decay. The degree distributions of four networks—Infect, Bitcoin, CA and Wiki—are not shown here due to limited space. Their degree distributions can be found in supporting information figure S1. The degree distribution of Infect mimics the other two offline networks of the Primary School and High School. For Bitcoin, CA and Wiki, they generally resonate the heavy tailed degree distribution of the other online networks.

with moderate degree variances, the proposed one-hopHD strategy is not necessarily the best way of seeding. Its seeding outcomes can get easily caught up by pure high-degree seeding as long as the latter has access to the degree information of more nodes.

The other seven networks from figures 3(d)−(j) are online social networks for email communication, Bitcoin trading, Facebook friendship, academic author collaboration (CA), p2p, Wiki voting and API, respectively, including from 1000 to 7600 nodes. Compared to the offline networks, such online networks have similar or even shorter path lengths and much more skewed degree distributions following a power-law decay (e.g. last four subpanels of figure 3(k)). In these online networks, one-hopHD outperforms Know30HD in nearly every combination of seed set size and threshold considered here (figures 3(d)−(j)), even though Know30HD has the degree information of more nodes. In some cases, the absolute difference is as large as 1 (red areas in figures 3(d)−(j)), indicating that one-hopHD can lead to adoption in the whole network while

Figure 4. Comparison of the final activated fractions between one-hopHD and Know30Cl in 10 real networks (a)−(j).The *x*-axis is the value of the threshold used in the contagion model, and the *y*-axis is the seed size. The colour gradient in (a)−(j) represents the absolute difference in the final activated fractions between one-hopHD and Know30Cl (average of the 10 000 runs), under specific combination of threshold and seed set size. The coordinates of every color pixel correspond to the threshold (value of *x*-axis) and seed set size (value of *y*-axis), respectively. Colours of red and yellow indicate one-hopHD lead to higher activated fractions than Know30Cl (one-hopHD Know30Cl), and the opposite for colours of blue and green. The primary school and high school networks are visualized in a & b, in which the network topology of the high school is less homogeneous. The absolute activation fractions achieved by one-hopHD and Know30Cl can be found in the supporting information figures S2 and S4, respectively.

Know30HD fails to trigger any significant diffusion. It suggests that, for such large networks with heavy tailed degree distribution and good connectivity (i.e. relatively short path lengths), one-hopHD is much more efficient than Know30HD and the premium cannot be caught up by pure high-degree seeding even though the latter one has access to more degree information.

Moreover, the efficiency of one-hopHD in (figures 3(d)−(j)) is independent of the clustering coefficients. The networks in (figures 3(d)−(j)) have clustering coefficients varying from the lowest of 0.08 in the Bitcoin network (figure 3(e)) to the highest of 0.6 in the CA network (figure 3(g)). The dominant factors deciding the relative performance between one-hopHD and Know30HD seem to be the degree variance and connectivity of the network. As demonstrated in previous studies [60, 61], networks with heavy tailed degree distribution such as those with a power-law decay end up in a system that most nodes have very few links while a few nodes (or hubs) have unproportionally large numbers of links. These few hubs can reach the majority of the network within one step. In turn, searching the connections from a random node is very likely to reach these hubs [60]. As a result, the tactic used by one-hopHD, i.e. searching from random connections before high-degree seeding, is very efficient to reach the highly connected nodes. This is the key why one-hopHD is efficient in such inhomogeneous networks with heavy tailed degree distribution.

Figure 5. Diffusion process in the high school network driven by the seeds of Know30Cl and one-hopHD. Nodes in yellow indicate that they are activated, while inactivated nodes are in blue. At time t_0, Know30Cl always activates a cluster of seeds, which are direct connections of each other as highlighted in the red circle in (a) and (g). On contrast, the one-hopHD starts from partly connected nodes, as highlighted in two red circles in (d) and (i). When the threshold is not too high (e.g. being 0.3 in this example), the highly clustered seeds provided by Know30Cl turn out to be overly concentrated in one area and fail to reach other parts of the network, as shown by the diffusion process from (a) to (c). Seeds of one-hopHD, on the contrary, can reach the majority of the network (f) since they start from more than one clusters. When the threshold is increased to 0.4, seeds need to be highly clustered to activate any cascade in the network. Know30Cl can activate slightly more nodes that one-hopHD, referring to the panels of (h) and (j). But the margin provided by Know30Cl is very low since even clustered seeding would fail to activate significant percentage of nodes when threshold is high in an inhomogeneous network.

3.3. Comparison between the outcomes of one-hopHD and Know30Cl for realworld networks

An important seeding strategy for complex contagion is clustered seeding: seeding a cluster of nodes that are connected to same nodes so that the other nodes connected to this cluster can receive multiple exposures to overcome their thresholds [15, 40, 41]. In order to evaluate the effectiveness of one-hopHD under complex contagion, its seeding outcomes should therefore be compared with clustered seeding with limited network information. Specifically, clustered seeding strategy (Know30Cl) assumes that the degrees of 30% nodes are known ('known nodes'), and seeds are the highest degree node chosen from these known nodes, and its $s - 1$ direct connections. In contrast, one-hopHD picks the highest degree nodes out from the random connections and has no guarantee that the seeds are direct connections of each other. According to the justification for
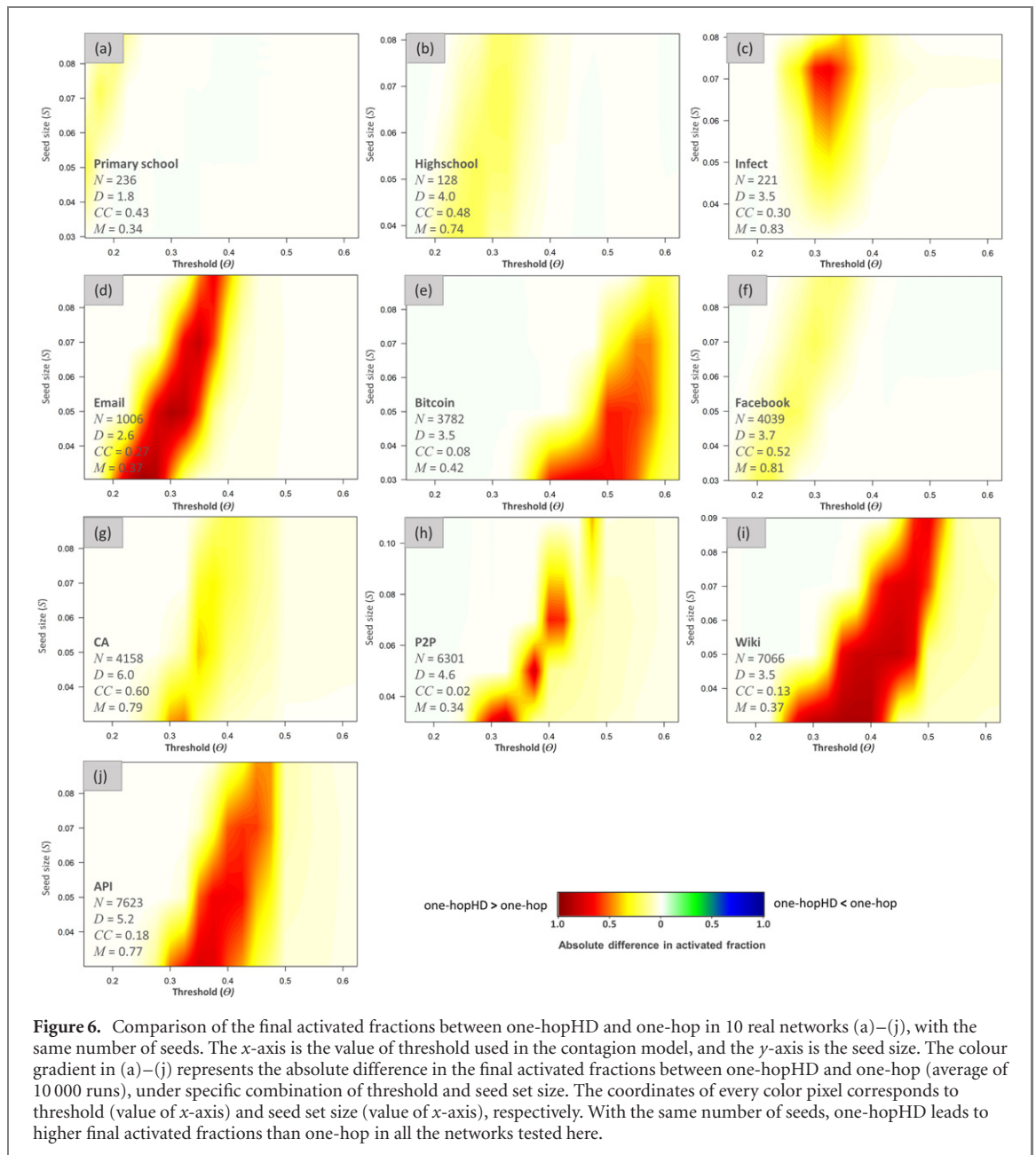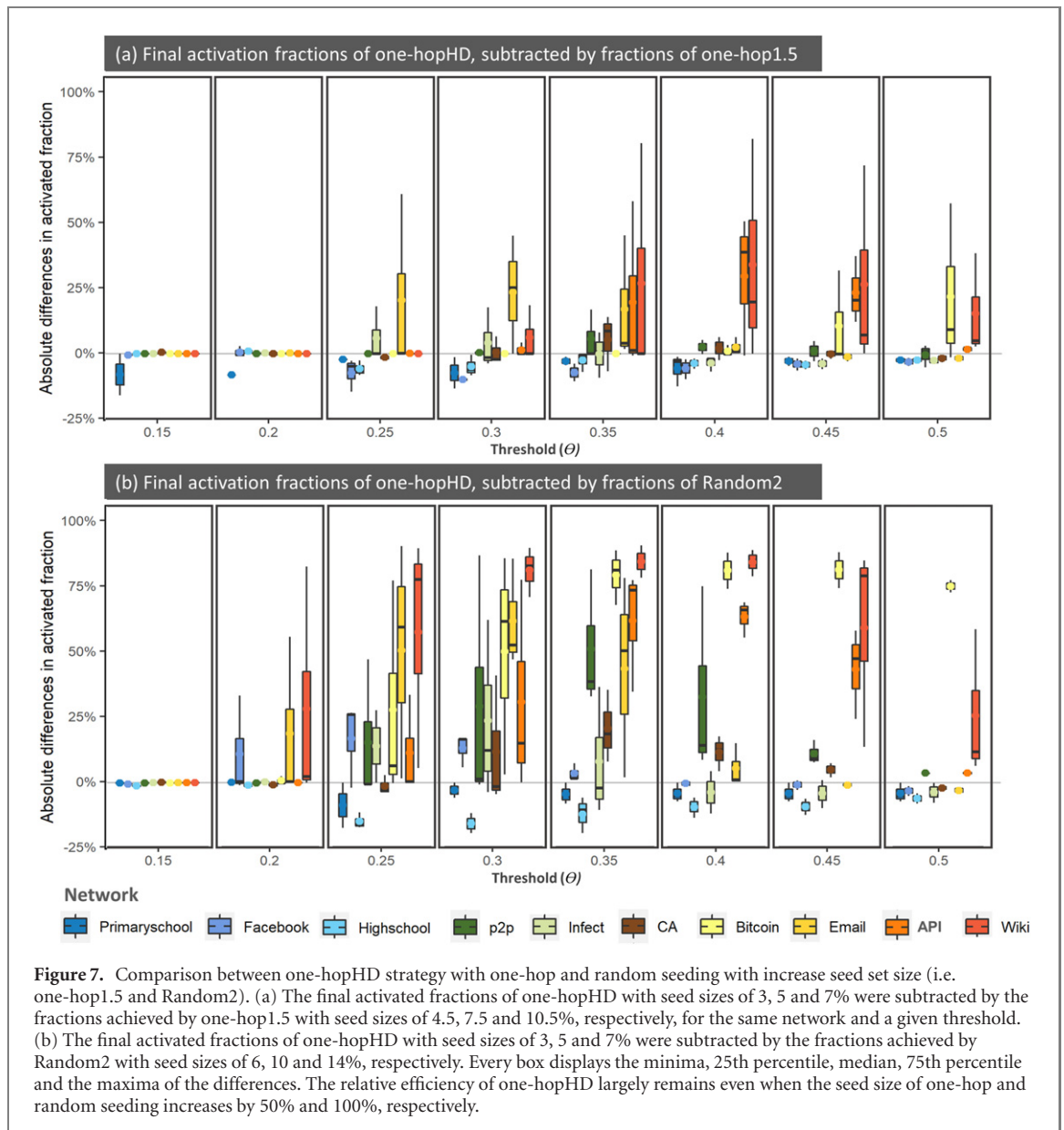
Figure 6. Comparison of the final activated fractions between one-hopHD and one-hop in 10 real networks (a)–(j), with the same number of seeds. The *x*-axis is the value of threshold used in the contagion model, and the *y*-axis is the seed size. The colour gradient in (a)–(j) represents the absolute difference in the final activated fractions between one-hopHD and one-hop (average of 10 000 runs), under specific combination of threshold and seed set size. The coordinates of every color pixel corresponds to threshold (value of *x*-axis) and seed set size (value of *x*-axis), respectively. With the same number of seeds, one-hopHD leads to higher final activated fractions than one-hop in all the networks tested here.

clustered seeding, one-hopHD would be suboptimal since seeds are scattered throughout the network and do not provide the required multiple exposures to other nodes adjacent to the seeds.

Simulation involving our ten real-world networks, however, suggests that one-hopHD outperforms the clustered seeding strategy (Know30Cl) in most cases, as shown in figures 4(a), (d) and (j), where we compare the final activated fractions by one-hopHD to those of Know30Cl. Specifically, with the exception of the primary school network (figure 4(a)) and high-threshold scenarios in a few networks, one-hopHD provides much higher final activated fractions than Know30Cl. This is especially true for large networks (figures 4(d)–(j)), where the absolute difference between the final activated fraction of one-hopHD and Know30Cl is as high as unity. These results indicate that, by activating the same number of seeds, one-hopHD successfully diffuses the contagion to the whole network while seeds of Know30Cl fail to activate substantial amounts of nodes, calling into question the justification for clustered seeding.

The reason why Know30Cl leads to worse seeding outcomes may lie in the fact that its effectiveness is mostly demonstrated in homogeneous networks such as synthetic SW networks (i.e. most nodes have similar degrees and there are abundant connections between every two clusters, meaning that the networks have low modularity values) [40, 41]. A homogeneous network ensures that seeding a cluster can activate the nodes adjacent to the seeds, and play the crucial role of passing the contagion on from one cluster to another by the redundant connections among clusters. This reasoning is bolstered by the observation that among all the real networks used for simulations in figure 4, the primary school network is the most homogeneous, and also

Figure 7. Comparison between one-hopHD strategy with one-hop and random seeding with increase seed set size (i.e. one-hop1.5 and Random2). (a) The final activated fractions of one-hopHD with seed sizes of 3, 5 and 7% were subtracted by the fractions achieved by one-hop1.5 with seed sizes of 4.5, 7.5 and 10.5%, respectively, for the same network and a given threshold. (b) The final activated fractions of one-hopHD with seed sizes of 3, 5 and 7% were subtracted by the fractions achieved by Random2 with seed sizes of 6, 10 and 14%, respectively. Every box displays the minima, 25th percentile, median, 75th percentile and the maxima of the differences. The relative efficiency of one-hopHD largely remains even when the seed size of one-hop and random seeding increases by 50% and 100%, respectively.

the only network for which Know30Cl performs better than one-hopHD. Most nodes in the primary school network have degrees around the mean value of 50, which, as shown in figure 4(a), results in a closely and evenly knitted network. Seeding a cluster of nodes in such a homogeneous network triggers large scale diffusion as predicted by [40, 41].

For other real-world networks, specifically with less homogeneous architecture, clustered seeding may not work well since the seeds in a clustered seeding strategy are overly concentrated in the same cluster, combined with a situation that there may not be enough connections between clusters to pass the contagion on. Let us demonstrate this concept using the high school network as an example. As for network architecture, the high school network has a similar size and clustering coefficient but a less homogeneous architecture in comparison to the primary school network, highlighted by a more skewed degree distribution and a higher value of modularity. Moreover, the modularity value in the high school network doubles that of the primary school network.

The seeding outcomes for Know30Cl and one-hopHD strategies for the high school network are shown in figure 5. At time t_0, the two strategies activate the same number of seeds (5% of the nodes). The seeds selected by Know30Cl are always direct connections of each other, as highlighted in the red circle in figure 5(a). In contrast, seeds selected by one-hopHD are not necessarily directly connected, but are sometimes partly connected as shown in figure 5(d). When the threshold is not too high (such as being 0.3), the partly connected seeds provided by one-hopHD turn out to be efficient enough to activate 72% of nodes in the network, as shown in figures 5(d)–(f). Under the same threshold value, seeds of Know30Cl are a group of directly connected nodes. Their activation reaches and activates their common connections within the close vicinity fast, within the first

few time steps (figures 5(a) and (b)). As the diffusion process goes on, however, they fail to reach the other parts of the network where the connections are sparse (figure 5(c)). The clustering of seeds becomes necessary again only when the threshed is high. For example, when the threshold goes up to 0.4 in figures 5(g)–(j), the partly connected seeds provided by one-hopHD fail to overcome the thresholds of any nearby nodes (figures 5(i) and (j)), while seeds of Know30Cl can still activate the nodes directly connected to the seeds (figures 5(g) and (h)). The difference between two strategies is however quite limited when the threshold is high, as shown in figure 5(h).

The above example shows that the key for the success of one-hopHD under complex contagion is to have seeds that are partly connected. Though one-hopHD does not deliberately use clustering as a criteria of seed selection, the selected seeds are often partly connected due to the connectedness of high degree nodes in most networks. For the ten real-world networks analyzed here, despite large variations of network sizes ranging from 100 to 7600, the average shortest path lengths between the top 10% highest degree nodes are between 1.3 to 4.5 (table 1). The academic author collaboration network (CA) has the highest value of 4.5, for a network size of $N = 4158$. For this network with the highest path length between highest degree nodes, one-hopHD still leads to higher final activated fractions than Know30Cl, though its relative efficiency (figure 4(g)) is less significant than what we observed in other networks (figures 4(d)–(j)). The connectivity of high-degree nodes can be partly explained by the core-periphery network structure resulting from homophily [62, 63]. Homophily refers to the phenomena that similar individuals in a social network tend to know each other and act in a similar manner [64]. Such individual choices result in the emergence of core-periphery structure in the network level: a dense 'core' of tightly connected nodes and a sparse 'periphery' of nodes loosely connected to the core and among each other [63, 65]. The short path lengths between high-degree nodes we discussed above is a reflection of the dense-core network pattern. Since homophily and the resulting core-periphery structure are ubiquitous in many real-world social networks, the success of one-hopHD can be extended to a wider range of networks in addition to those in this study.

3.4. Comparison between the outcomes of one-hopHD, one-hop and random seeding for real-world networks

Compared to one-hopHD, both the conventional one-hop and random seeding require less network information. If all these strategies are used to activate the same number of seeds, as expected, one-hopHD would lead to better seeding outcomes. Indeed, as shown in figure 6, one-hopHD leads to higher activated fraction than one-hop strategy in all the networks tested here, under various settings of threshold and seed set size. Results for the comparison between one-hopHD and random seeding are similar, which can be found in the supporting information figure S5.

One critique of the one-hop strategy is the generality of its efficacy, which varies a lot from one network to another, depending on the architecture of the network and the setting of contagion models such as seed set size and threshold value [8, 66]. Such variabilities in the performance are to be expected, given that the seeding strategy is applied stochastically, starting from a random draw of nodes from the network. Seeds provided by one-hop are characterized by the use of high-degree nodes (normally higher than the mean degree). They are also likely to be distributed broadly throughout the network, though to a lesser extent than for the case when seeding is random. The one-hopHD stresses the first characteristic even more by selecting the high-degree nodes from random connections. In turn, the second characteristic is compromised as seeds would be more concentrated towards the 'center' of the network. The center here refers to the innermost core of the network where most efficient spreaders locate [67]. Such a trade-off between high degree and randomness turns out to be successful in increasing the final activated fraction in a wide range of network structures, seed set sizes and threshold value (figure 6). Compared to the one-hop strategy, one-hopHD provides better seeding outcomes across different networks, seed set sizes and thresholds.

Recognizing that one-hop and random seeding require less network information than one-hopHD, we test whether the premium yielded by one-hopHD can be easily caught up by one-hop or random if the latter can seed more. Specifically, the number of seeds by one-hop and random are increased by 50% ('one-hop1.5') and 100% ('Random2'), respectively. Results are shown in figure 7. In figure 7(a), the final activated fractions achieved by one-hop by seeding 4.5, 7.5 and 10.5% nodes were compared to the results achieved by one-hopHD by seeding 3, 5 and 7% nodes, respectively. Each box plot shows the variations of differences between two seeding strategies for the same network and a given threshold. For example, the first box in figure 7(a) shows the differences between one-hopHD and one-hop1.5 for the primary school network when the threshold is 0.15. When one-hop can seed 50% more nodes, its final activated fraction can indeed catch up with one-hopHD in a few networks such as the primary school network, Facebook network and high school network. For the other networks, the final activated fractions by one-hop1.5 still lag behind one-hopHD, especially for the networks of Bitcoin, Email, API and Wiki. In short, while one-hop can catch up the seeding outcomes of one-hopHD in some cases simply by increasing the number of seeds, the number of seeds need to be increased

by at least 50% and it does not apply to every network. In fact, for half the networks tested here, a 50% increase of the seed number of one-hop still fails to reproduce the seeding outcomes achieved by one-hopHD.

Regarding random seeding that requires no network information at all, its number of seeds are doubled and compared to one-hopHD. In figure 7(b), random strategy seeds 6, 10 and 14% nodes in each network and the final activated fractions were compared to those achieved by one-hopHD which can only seed 3, 5, and 7% nodes in the same network. Even though random seeding doubles the number of seeds, its seeding outcomes still largely falls behind those of one-hopHD in most networks. The absolute differences between one-hopHD and Random2 are as high as 50% to 80% in half of the cases. The premium yielded by one-hopHD in terms of the final activated fraction largely remains even though random strategy can seed twice as many.

The above results suggest that the conclusion found for simple contagion [31], i.e. a small increase of the random seed set size can match the seeding outcomes with network information, is not applicable to complex contagion. Our results here suggest that network information, even if it is very limited, *does* matter when the diffusion process is governed by complex contagion. With the help of some degree and connection information, seeding strategies such as one-hopHD can lead to significantly higher final activated fractions that cannot be matched by blindly seeding more.

4. Conclusion

This study has introduced the one-hopHD seeding strategy, a variant of the conventional one-hop strategy, for effectively seeding contagions in networks with very limited information on nodes and edges. The conventional one-hop strategy selects *high-degree* nodes in a *random* fashion. The first characteristic attempts to select potentially influential nodes and the second one hopes to have the seeds distributed broadly throughout the network. The one-hopHD strategy proposed here enhances the first characteristic by selecting the high-degree nodes from the random connections provided by the conventional one-hop strategy. While such a strategy compromises the randomness of conventional one-hop seeds, it reaches a better balance between reaching influential nodes and scattering seeds broadly. It outperforms alternative seeding strategies from the literature across a wide range of networks and parameter settings.

The comparative performance of the one-hopHD in various networks is summarized in table 3. The efficiency of one-hopHD especially stands out in networks with heavy tailed degree distributions such as synthetic SF networks and the large online social networks. Heavy tailed degree distribution in a network has two implications. The first one is that there exist some highly connected nodes (e.g. hubs) within the network. The second one is that most nodes have very few links, but are connected to the hubs. For such networks, the connections nominated by random nodes (i.e. the first step of one-hopHD) are more often highly connected nodes than the random nodes themselves. This chance is much higher when the extra step of explicitly screening for high degree nodes among nominated neighbors is taken, as in the one-hopHD strategy. Therefore, for networks with heavy tailed degree distribution, the combination of searching from random connections and high-degree seeding is much more efficient than pure high-degree seeding of random nodes. We have demonstrated this by comparing the results of one-hopHD to Know30HD in both the synthetic and real networks. Even though the Know30HD strategy has access to more degree information, i.e. 30% nodes compared to 15% for one-hopHD, its seeding outcomes largely lag behind those of one-hopHD in the synthetic scale-free networks and large online social networks with degree distribution approximating a power-law decay. Another factor contributing to the success of one-hopHD is the ubiquitous core-periphery network structure. It explains the connectivity of high-degree nodes and why the seeds selected by one-hopHD are often partly connected. Combined with the random nature inherent in the random connection nomination process, seeds for one-hopHD are partly connected but are not overly concentrated. These are desirable properties for overcoming thresholds in complex contagion processes, as well as for wide diffusion of activation to different parts of the network when the architecture of the network is not homogeneous. This is in fact the reason why one-hopHD outperforms the clustering seeding strategy with more degree information (Know30Cl) for a wide range of social networks.

While it is true that one-hopHD requires more network information than the conventional one-hop strategy, in most practical situations the extra information can be secured through reasonable efforts. Importantly, with the help of this extra information, one-hopHD can lead to similar or much better seeding outcomes than the conventional one-hop strategy even though the latter has the option to seed much more. This implies that improvements achieved by the one-hopHD seeding strategy cannot be matched by simply having more seeds. This also holds true when we compare the seeding outcomes of one-hopHD to random seeding even if the latter can seed twice as many. Our results highlight that limited network information to inform seeding under complex contagion and demonstrate that one-hopHD strategy is an efficient way to ensure of limited information on the involved network.

Table 3. Summary of network properties and seeding outcomes.

Networks	Path lengths	Clustering	Degree distribution	Seeding strategies comparison
Synthetic ER network	Short	Zero	Normal distribution	Know30HD > one-hopHD > Know30Cl > one-hop > random
Synthetic SW network	Short	High	Approx. delta distribution	Know30Cl > one-hopHD, Know30HD > one-hop > random
Synthetic SF network	Long	Zero	Power-law distribution	one-hopHD > Know30HD > one-hop > Know30Cl > random
Small offline networks: Primary school, high-school, exhibition networks	Short	High	Log-normal distribution	Know30HD > one-hopHD, Know30Cl > one-hop > random
Large online networks: Email, Bitcoin, Facebook, CA, P2P, Wiki and API	Short	Low to high	Power-law distribution	one-hopHD > one-hop > know30HD > Know30Cl > random

Another limitation of this study is that the efficiency of one-hopHD is not compared to the best possible outcomes achieved by influence-maximization algorithms under full network information. The calculation of such theoretical optima is complex and approximation cumbersome, but a potential direction for future research. We therefore do not know how far below the upper bound on performance we still are and whether yet other strategies could greatly improve on one-hopHD. These limitations not withstanding, the present paper suggests the possibility of greater efficacy of network interventions in applications ranging from public health to the diffusion of sustainable behaviors.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/JiaminOu/High-degree-seeding-of-random-neighbors-in-unknown-graph.

Author contributions

All the authors conceived the mathematical principles. JO performed the numerical simulation and application to the synthetic and real networks, with help from DP. JO wrote the first draft of the manuscript. All authors reviewed the final text.

Conflict of interests

The authors declare no competing interests

ORCID iDs

Jiamin Ou https://orcid.org/0000-0001-5040-4941
Debabrata Panja https://orcid.org/0000-0003-2141-9735

References

[1] Vosoughi S, Roy D and Aral S 2018 The spread of true and false news online *Science* **359** 1146–51

[2] Christakis N A and Fowler J H 2007 The spread of obesity in a large social network over 32 years *N Engl J. Med.* **357** 370–9

[3] Gardy J L *et al* 2011 Whole-genome sequencing and social-network analysis of a tuberculosis outbreak *N Engl J. Med.* **364** 730–9

[4] Morone F and Makse H A 2015 Influence maximization in complex networks through optimal percolation *Nature* **524** 65–8

[5] Valente T W 2012 Network interventions *Science* **337** 49–53

[6] Ball F and Sirl D 2013 Acquaintance vaccination in an epidemic on a random graph with specified degree distribution *J. Appl. Probab.* **50** 1147–68

[7] Kim D A, Hwong A R, Stafford D, Hughes D A, O'Malley A, Fowler J H and Christakis N A 2015 Social network targeting to maximise population behaviour change: a cluster randomised controlled trial *Lancet* **386** 145–53

[8] Chin A, Eckles D and Ugander J 2021 Evaluating stochastic seeding strategies in networks *Manage. Sci.* **68** 1–63

[9] Katona Z, Zubcsek P and Sarvary M 2011 Network effects and personal influences: the diffusion of an online social network *J. Market. Res.* **48** 425–43

[10] Kempe D, Kleinberg J and Tardos É 2015 Maximizing the spread of influence through a social network *Theory Comput.* **11** 105–47

[11] Zhu J, Liu Y and Yin X 2017 A new structure-hole-based algorithm for influence maximization in large online social networks *IEEE Access* **5** 23405–12

[12] Cohen R, Havlin S and Ben-Avraham D 2003 Efficient immunization strategies for computer networks and populations *Phys. Rev. Lett.* **91** 247901

[13] Li J, Yue K, Zhang D and Liu W 2016 Robust influence blocking maximization in social networks *J. Comput. Res. Dev.* **53** 601–10

[14] Goyal A, Lu W and Lakshmanan V S 2011 CELF++: Optimizing the greedy algorithm for influence maximization in social networks *Proc. 20th Int. Conf. Companion on World Wide Web* pp 47–8

[15] Guilbeault D and Centola D 2021 Topological measures for identifying and predicting the spread of complex contagions *Nat. Commun.* **12** 4430

[16] Mihara S, Tsugawa S and Ohsaki H 2017 On the effectiveness of random jumps in an influence maximization algorithm for unknown graphs *2017 Int. Conf. Information Networking* pp 395–400

[17] Mihara S, Tsugawa S and Ohsaki H 2015 Influence maximization problem for unknown social networks *Proc. 2015 IEEE/ACM Int. Conf. Advances in Social Networks Analysis and Mining* pp 1539–46

[18] Wilder B, Immorlica N, Rice E and Tambe M 2017 Influence maximization with an unknown network by exploiting community structure *Proc. 3rd Int. Workshop on Social Influence Analysis* pp 1–7

[19] Wilder B, Immorlica N, Rice E and Tambe M 2018 Maximizing influence in an unknown social network *32nd AAAI Conf. Artificial Intelligence* pp 4743–50

[20] Eckles D, Esfandiari H, Mossel E and Amin Rahimian M 2022 Seeding with costly network information *Oper. Res.* **70** 2318–48

[21] Stein S, Eshghi S, Maghsudi S, Tassiulas L, Bellamy R K E and Jennings N R 2017 Heuristic algorithms for influence maximization in partially observable social networks *Proc. 3rd Int. Workshop on Social Influence Analysis* pp 20–32

[22] Erkol Ş and Yücel G 2017 Influence maximization based on partial network structure information: a comparative analysis on seed selection heuristics *Int. J. Mod. Phys.* C **28** 1750122

[23] Paluck E, Shepherd H and Aronow P M 2016 Changing climates of conflict: a social network experiment in 56 schools *Proc. Natl Acad. Sci. USA* **113** 566–71

[24] Feld S L 1991 Why your friends have more friends than you do *Am. J. Sociol.* **96** 1464–77

[25] Chami G F, Ahnert S E, Kabatereine N B and Tukahebwa E M 2017 Social network fragmentation and community health *Proc. Natl Acad. Sci. USA* **114** E7425–31

[26] Shakya H A, Stafford D, Hughes D, Keegan T, Negron R *et al* 2017 Exploiting social influence to magnify population-level behaviour change in maternal and child health: study protocol for a randomised controlled trial of network targeting algorithms in rural Honduras *BMJ Open* **7** e012996

[27] Hashemian M S, Stanley K G, Knowles D L, Calver J and Osgood N D 2012 Human network data collection in the wild: the epidemiological utility of micro-contact and location data *Proc. 2nd ACM SIGHIT Int. Health Informatics Symp., IHI '12* (New York, NY, USA) pp 255–64

[28] Pomare C, Long J C, Churruca K, Ellis L A and Braithwaite J 2019 Social network research in health care settings: design and data collection *Social Networks* **69** 14–21

[29] Carrasco J A, Hogan B, Wellman B and Miller E J 2008 Collecting social network data to study social activity-travel behavior: an egocentric approach *Environ. Planning* B **35** 961–80

[30] Ricken S T, Schuler R P, Grandhi S A and Jones Q 2010 Telluswho: Guided social network data collection *2010 43rd Hawaii Int. Conf. System Sciences* pp 1–10

[31] Akbarpour M, Malladi S and Amin S 2020 *Just a Few Seeds More: Value of Network Information for Diffusion* Available at SSRN: https://ssrn.com/abstract=3062830

[32] Lattanzi S and Singer Y 2015 The power of random neighbors in social networks *Proc. 8th ACM Int. Conf. Web Search and Data Mining* pp 77–86

[33] Peng S, Wang G, Zhou Y, Wan C, Wang C, Yu S and Niu J 2019 An immunization framework for social networks through big data based influence modeling *IEEE Trans. Depend. Secure Comput.* **16** 984–95

[34] Zhang J, Huang H, Luo Y, Fan Y and Yang Y 2018 Immunization-based redundancy elimination in mobile opportunistic networks-generated big data *Future Gener. Comput. Syst.* **79** 920–7

[35] Centola D and Macy M 2007 Complex contagions and the weakness of long ties *Am. J. Sociol.* **113** 702–34

[36] Centola D 2010 The spread of behavior in an online social network experiment *Science* **329** 1194–7

[37] Aral S, Muchnik L and Sundararajan A 2009 Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks *Proc. Natl Acad. Sci. USA* **106** 21544–9

[38] Mønsted B, Sapieżyński P, Ferrara E and Lehmann S 2017 Evidence of complex contagion of information in social media: an experiment using twitter bots *PLoS One* **12** 1–12

[39] Bakshy E, Rosenn I, Marlow C and Adamic L 2012 The role of social networks in information diffusion *Proc. 21st Int. Conf. World Wide Web* (New York, NY, USA) pp 519–28

[40] Centola D 2018 *How Behavior Spreads* (Princeton, NJ: Princeton University Press)

[41] Centola D and Macy M 2007 Complex contagions and the weakness of long ties *Am. J. Sociol.* **113** 702–34

[42] Fienberg S E 2012 A brief history of statistical models for network analysis and open challenges *J. Comput. Graph. Stati.* **21** 825–39

[43] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2

[44] Albert R and Barabási A-L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47–97

[45] Albert R, Jeong H and Barabási A-L 2000 Error and attack tolerance of complex networks *Nature* **406** 378–82

[46] Mastrandrea R, Fournet J and Barrat A 2015 Contact patterns in a high school: a comparison between data collected using wearable sensors, contact diaries and friendship surveys *PLoS One* **10** 1–26

[47] Stehlé J *et al* 2011 High-resolution measurements of face-to-face contact patterns in a primary school *PLoS One* **6** 1–13

[48] Isella L, Stehlé J, Barrat A, Cattuto C, Pinton J-F and Van den Broeck W 2011 What's in a crowd? Analysis of face-to-face behavioral networks *J. Theor. Biol.* **271** 166–80

[49] McAuley J and Leskovec J 2012 Learning to discover social circles in ego networks *Proc. 25th Int. Conf. Neural Information Processing Systems—volume 1, NIPS'12* (Red Hook, NY, USA) 1 pp 539–47

[50] Leskovec J, Huttenlocher D and Kleinberg J 2010 Signed networks in social media *Proc. SIGCHI Conf. Human Factors in Computing Systems* pp 1360–70

[51] Leskovec J, Huttenlocher D P and Kleinberg J M 2010 Predicting positive and negative links in online social networks *Proc. 19th Int. Conf. World Wide Web* pp 641–50

[52] Kumar S, Spezzano F, Subrahmanian V S and Faloutsos C 2016 Edge weight prediction in weighted signed networks *IEEE 16th Int. Conf. Data Mining* pp 221–30

[53] Kumar S, Hooi B, Makhija D, Kumar M, Faloutsos C and Subrahmanian V S 2018 Rev2: fraudulent user prediction in rating platforms *Proc. 11th ACM Int. Conf. Web Search and Data Mining* pp 333–41

[54] Rozemberczki B and Sarkar R 2020 Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models *Proc. 29th ACM Int. Conf. Information and Knowledge Management* pp 1325–34

[55] Leskovec J, Kleinberg J and Faloutsos C 2007 Graph evolution: densification and shrinking diameters *ACM Trans. Knowl. Discov. Data* **1** 2–es

[56] Ripeanu M, Foster I and Iamnitchi A 2002 Mapping the gnutella network: properties of large-scale peer-to-peer systems and implications for system design *IEEE Internet Comput. J.* **6** 10

[57] Yin H, Benson A R, Leskovec J and Gleich D F 2017 Local higher-order graph clustering *Proc. 23rd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* pp 555–64

[58] Morris S M 2000 *Rev. Econ. Stud.* **67** 57–78

[59] Berger E 2001 Dynamic monopolies of constant size *J. Comb. Theory* B **83** 191–200

[60] Crucitti P, Latora V, Marchiori M and Rapisarda A 2004 Error and attack tolerance of complex networks *Physica* A **340** 388–94

[61] Tsiotas D 2019 Detecting different topologies immanent in scale-free networks with the same degree distribution *Proc. Natl Acad. Sci. USA* **116** 6701–6

[62] Zhang X, Martin T and Newman M E J 2015 Identification of core-periphery structure in networks *Phys. Rev.* E **91** 032803

[63] Gallagher R J, Young J-G and Welles B F 2021 A clarified typology of core-periphery structure in networks *Sci. Adv.* **7** eabc9800

[64] McPherson M, Smith-Lovin L and Cook J M 2001 Birds of a feather: homophily in social networks *Annu. Rev. Sociol.* **27** 415–44

[65] Asikainen A, Iñiguez G, Ureña-Carrión J, Kaski K and Kivelä M 2020 Cumulative effects of triadic closure and homophily in social networks *Sci. Adv.* **6** 7310

[66] Kumar V and Sudhir K 2019 *Can Friends Seed More Buzz and Adoption?* Available at SSRN: https://ssrn.com/abstract=3395044

[67] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H and Makse H A 2010 Identification of influential spreaders in complex networks *Nat. Phys.* **6** 888–93