Online Myopic Network Covering

K. Avrachenkov[1], P. Basu[2], G. Neglia[1], B. Ribeiro[3]*, and D. Towsley[3]

UMass Technical Report UM-CS-2012-034

[1] INRIA
06902 Sophia Antipolis
France
{konstantin.avratchenkov,
giovanni.neglia}@inria.fr

[2]Raytheon BBN Technologies
Cambridge, MA
pbasu@bbn.com

[3] Department of Computer Science
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003
{ribeiro,towsley}@cs.umass.edu

ABSTRACT

Efficient marketing or awareness-raising campaigns seek to recruit n influential individuals – where n is the campaign budget – that are able to cover a large target audience through their social connections. So far most of the related literature on maximizing this network cover assumes that the social network topology is known. Even in such a case the optimal solution is NP-hard. In practice, however, the network topology is generally unknown and needs to be discovered on-the-fly. In this work we consider an unknown topology where recruited individuals disclose their social connections (a feature known as *one-hop lookahead*). The goal of this work is to provide an efficient greedy online algorithm that recruits individuals as to maximize the size of target audience covered by the campaign.

We propose a new greedy online algorithm, Maximum Expected d-Excess Degree (MEED), and provide, to the best of our knowledge, the first detailed theoretical analysis of the cover size of a variety of well known network sampling algorithms on finite networks. Our proposed algorithm greedily maximizes the expected size of the cover. For a class of random power law networks we show that MEED simplifies into a straightforward procedure, which we denote MOD (Maximum Observed Degree). We substantiate our analytical results with extensive simulations and show that MOD significantly outperforms all analyzed myopic algorithms. We note that performance may be further improved if the node degree distribution is known or can be estimated online during the campaign.

1. INTRODUCTION

This paper addresses the need to efficiently select n individuals in a network such that they cover, through their neighbors, the largest possible fraction of the network. Online social networks have generated much attention as a breeding ground for new forms of social studies, social mobilization, and online campaigns. Recruiting individuals from a population – for instance, recruiting volunteers to get their friends to vote in an election – is no easy task. The recruitment of each individual comes at a cost in time, money, and social capital; and the total budget is often small with respect to the total population. Moreover, recruitment is frequently targeted towards a subpopulation – say, individuals that will likely vote for a given candidate – that may be a relatively small fraction of the whole population. Most works on network cover, e.g. [12, 16, 29], either consider the social network topology to be known in advance or assume the capability of a *two-hop lookahead* (where the identity of all nodes within a two-hop neighborhood of a recruited node are known), which is often not the case in the wild.

In this work we look at the cover problem when the network topology is unknown. Following previous literature, we assume that any individual in the network can be recruited – but in our case recruitments mostly happen through friends recruiting friends. This link-tracing technique has been long used by social scientists to sample hard-to-reach subpopulations [13, 15, 18]. The homophily often present in social networks – the tendency for similar individuals to be friends [30] – enables the likely effective recruitment of individuals that are either in the target subpopulation or know many unrecruited individuals in the target subpopulation. This is achieved simply by asking each recruited individual to refer other target individuals.

The recent 2012 U.S. presidential election presents a real-life example of an application of link-tracing recruitments to maximize the network cover of a target subpopulation. A candidate's Facebook app asked its subscribers to send get-out-to-vote reminders to their like-minded friends in swing states [19]. Thus, the effectiveness of a subscriber is measured by the amount of its friends that live in swing states. Moreover, these messages also raised awareness of the app itself, allowing it to spread through the target subpopulation of interest (see also Bond et al. [7] for a description of a get-out-to-vote Facebook app experiment in the 2010 U.S. elections).

Problem Formulation

We formulate the target subpopulation cover problem as a maximum *connected* cover (MCC) problem on an unknown connected graph $G = (V, E)$ (we also refer to G as a network), where V is the set of target individuals and E the set of individuals' mutual connections. We assume all graph parameters are unknown. Our analysis can be easily extended to a disconnected network by considering each connected component separately.

Our main goal is to design efficient online greedy algorithms to solve the following problem on G: let n be a given campaign budget; we want to determine a group of n individuals to be recruited in order to maximize the size of the covered subset, i.e. of the set including the recruited nodes

[*] Corresponding author: ribeiro@cs.umass.edu

and their neighbors. Our only initially available information is a single node sampled from the population. Later we can acquire the neighborhood of any recruited node. It follows that the recruited nodes form a connected subgraph.

More formally, let $\mathcal{B}(t)$ be the set of t known target individuals after t recruitments (also denoted the t-th step). Let $\mathcal{B}(0)$ be the set containing the initially known individual[1] Let $\mathcal{N}(\mathcal{B}(t))$ be a function that returns the set of unrecruited neighbors of $\mathcal{B}(t) \subseteq V$; Fig. 1 illustrates $\mathcal{B}(t)$ and $\mathcal{N}(\mathcal{B}(t))$. The online algorithm proceeds as follows: at step t, $0 < t \leq n$, the algorithm recruits node $v \in \mathcal{N}(\mathcal{B}(t-1))$ and performs the update $\mathcal{B}(t) = \mathcal{B}(t-1) \cup \{v\}$. The ability to obtaining the identities of the neighbors of recruited nodes, $\mathcal{N}(\mathcal{B}(t))$, $\forall t$, is known as *one-hop lookahead* in the graph sampling literature [1, 31]. The objective of the online algorithm is to try to maximize the size of the network cover set $\overline{\mathcal{W}}(t) = \mathcal{B}(t) \cup \mathcal{N}(\mathcal{B}(t))$, for $t = 1, \ldots, n$, without having a priori access to topology information. We refer to the problem of online covering an unknown network in the presence of one-hop lookahead as the *online myopic network covering* problem.

Contributions

We make the following contributions:

(1) We thoroughly evaluate – analytically and through extensive simulations on social network datasets[2] – the performance of several known network sampling algorithms. **(1.1)** We investigate the cover sizes of Breadth-First Search (BFS) and Depth-First Search (DFS). We observe a consistent large variance in the cover sizes found by BFS and that BFS tends to underperform in comparison to a greedy oracle scheme that recruits at every step the node with the largest number of uncovered neighbors. We partially blame network homophily for the lack of performance from BFS. DFS, which at first sight should improve upon BFS in circumventing the above homophily problem, performs even worse. Using random networks, we show why DFS finds a small cover after $t \ll N$ recruited nodes. **(1.2)** A Random Walk (RW), more precisely RW *without replacement* (RWnr), where nodes revisited by the walker are not counted towards the recruitment budget, is shown to consistently outperform (sometimes significantly) BFS in our simulations. **(1.3)** We propose a new online algorithm inspired by the Susceptible-Infected (SI) epidemic model but observe that RWnr is consistently more efficient than SI.

(2) Our work is, to the best of our knowledge, the first to provide an analytical characterization of the sizes of $\overline{\mathcal{W}}(t)$ (the cover) as a function of t (recruited nodes), for RWnr and the SI epidemic algorithms on *finite networks*. As recently acknowledged in [10], this was a challenging open problem. Moreover, we establish an interesting connection between cover through RWs and the coupon subset collection problem. We validate our theoretical results through simulations.

(3) We propose a new online algorithm (MEED, Maximum Expected d-Excess Degree) that greedily maximizes the expected size of the cover. For a broad class of power law networks, MEED simplifies into a straight forward heuristic, which we denote Maximum Observed Degree (MOD). We substantiate our analytical results with simulations. Extensive simulations on a variety of social network datasets show that MOD consistently outperforms (sometimes significantly) all other analyzed algorithms. Performance can be further improved if the node degree distribution is known or can be estimated online during the campaign.

Outline

The reminder of this work is organized as follows. Sec. 2 presents the notation and background used throughout this work. Sec. 3 discusses optimal solutions and approximations in connection to the connected minimum dominating set. Sec. 4 presents the datasets used in this work and our simulation setup. Sec. 5 provides an analysis of the effectiveness of Breadth-First-Search (BFS) and Depth-First-Search (DFS). Sec. 6 provides a deep analysis of the effectiveness of two types of random walks and compare them to BFS. Sec. 7 proposes a sampling algorithm inspired by Susceptible-Infected (SI) epidemic models. We also provide an analytical solution describing the cover size of SI as a function of t. An important feature of our analysis is our ability to model finite graphs, which is key to understanding the effectiveness of large campaigns in respect to the size of the target population. Sec. 8 proposes MEED and MOD as a simple approximation of MEED. Sec. 8 also provides theoretical and simulation results, the latter comparing MOD against the other algorithms. And, finally, Sec. 9 summarizes our contributions and reviews the related work.

2. NOTATION & BACKGROUND

We consider an unknown connected network $G = (V, E)$ with $N = |V|$ nodes, $M = |E|$ edges, and degree distribution $\{p_k\}_{k=1,\ldots N-1}$. We assume all graph parameters are unknown to us. Denote $\mathcal{N}_a(v)$ the set of neighbors of node $v \in V$, irrespectively of their recruitment status, and $k_v = |\mathcal{N}_a(v)|$ is the degree of v. For each step $t = 1, \ldots, n$, where $n \in \{1, \ldots, N-1\}$ is the campaign budget, we classify the nodes in V into three disjoint sets. The set $\mathcal{B}(t)$ denotes the recruited nodes at step t; these are the black nodes in Fig. 1. Unrecruited neighbors of recruited nodes are denoted *observed nodes* and form the set $\mathcal{N}(\mathcal{B}(t)) = \cup_{v \in \mathcal{B}(t)} \mathcal{N}_a(v) - \mathcal{B}(t)$ (gray nodes in Fig. 1). We say a node $v \in V$ is *covered* at step t if $v \in \overline{\mathcal{W}}(t)$, where $\mathcal{W}(t) = V - \mathcal{B}(t) \cup \mathcal{N}(\mathcal{B}(t))$ is the set of all *uncovered* nodes (white nodes in Fig. 1) and $\overline{\mathcal{W}}(t) = V - \mathcal{W}(t)$ its complement. Note that at time t we are unaware of the existence of nodes in $\mathcal{W}(t)$.

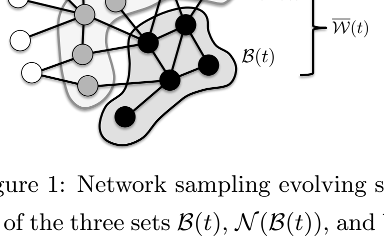


Figure 1: Network sampling evolving sets.

The sizes of the three sets $\mathcal{B}(t)$, $\mathcal{N}(\mathcal{B}(t))$, and $\mathcal{W}(t)$ are denoted $B(t)$, $\mathsf{N}(\mathcal{B}(t))$ and $W(t)$, respectively. Clearly $B(t) +$

[1] The analysis can be extended to consider many initially known individuals. Note that the task of finding the initial set of nodes in the target subpopulation is a problem on its own [1, 28].

[2] See Sec. 9 for some limited analysis of other "non-social" networks.

Variable	Description
N	no. of nodes
M	no. of edges
$\mathcal{N}_a(v)$	set of neighbors of $v \in V$
$\langle x \rangle$	average value of quantity x
p_k	fraction of nodes with degree k
$\mathcal{B}(t)$, $(B(t))$	set (number) of sampled nodes at step t
$\mathcal{N}(A)$, $(\mathsf{N}(A))$	set (no.) of unrecruited neighs of $A \subseteq V$
$\mathcal{W}(t)$, $(W(t))$	set (no.) of uncovered nodes at step t

Table 1: Notation Table

$\mathsf{N}(\mathcal{B}(t)) + W(t) = N$ at any step $1 \leq t \leq n$. Finally, for the sake of simplicity, we allow a slight abuse of notation, denoting by $\langle . \rangle$ both the empirical mean and the expected value. The exact interpretation of $\langle x \rangle$ will then depend on the nature of the quantity x. We use the convention that $\langle k \rangle$ denotes the average degree. Table 1 summarizes the notation used throughout the paper.

Our analysis makes extensive use of the configuration random graph model [39]. This is a defined as a uniform probability distribution over the ensemble of the graphs where nodes have a given degree distribution $\{p_k\}_{k=1,\ldots N-1}$. A configuration model sample can be generated as follows. The degree k_i is attributed to each node i according to the selected degree distribution. Each vertex i can then be thought of having k_i stubs attached to it that are the ends of edges-to-be. By connecting randomly selected stubs' pairs the graph sample is generated.

The configuration model is widely used in the complex network literature [5, 40] also for the simplicity of the analysis. Moreover, as we will soon see, the formulas we derive to predict the value of $\mathsf{N}(\mathcal{B}(t))$ considering the configuration model match the results of our simulation for actual topologies remarkably well.

3. NETWORK COVERS, ORACLES, & APPROXIMATE SOLUTIONS

The problem we study is closely related to the well-studied Maximum coverage [36] and Minimum Connected Dominating Set (MCDS) problems. The maximum coverage problem can be described in our setting as selecting at most n nodes such that the union of the nodes they cover has maximal size. The maximum coverage problem is NP-hard, and cannot be approximated within $1 - 1/e + o(1)$, where e is the Euler constant. A simple submodular function greedy algorithm, however, is able to find a $1 - 1/e$ approximation [36]. Our problem setting, however, requires the recruited nodes to be connected to each other. In the connected setting, above mentioned greedy algorithm is similar to a greedy algorithm used to solve the MCDS, described as follows.

Given a graph $G = (V, E)$ with N nodes, $DS \subseteq V$ is a dominating set if $\forall_{v \in V} : v \in DS$ or $v \in V - DS$. Thus, if all nodes in DS are recruited as *dominators*, it may be possible to reach all nodes in the network through these *dominators*. The MDS problem is to find the set DS with the minimum cardinality. MCDS imposes an additional restriction that the subgraph induced by the vertices in DS has to be connected.

Our goal is not to cover *all* of G; instead, we seek to cover as much of G as possible with recruitment budget n. However, since MCDS is very closely related to our problem, key results and techniques from the MCDS literature can provide crucial insights into the role of lookahead in network coverage especially about worst-case performance guarantees when compared to the optimal solution. In a situation where complete network knowledge is available[3], solving MCDS is NP-hard [12]. However, there exist well-known linear approximation preserving reductions (L-reductions [21]) from the SetCover problem to MCDS [16] that yield a guaranteed approximation factor of $O(\ln n)$.

Definition 3.1 (Observed degree). *The number of recruited neighbors of a node.*

Definition 3.2 (d-excess degree). *A node with degree k and observed degree $d \leq k$ has excess degree $k - d$.*

If there is limited "lookahead", say, *two-hop* information of the neighborhood of each recruited node, the natural algorithm is to greedily recruit nodes that have the maximum number of uncovered neighbors, i.e., with the maximum *excess degree*. Guha and Khuller [16] implemented this greedy algorithm by building *growing* a tree T in an online fashion, starting from a single node. Initially all nodes are unrecruited (white). At each step, a vertex $v \in T$ with the largest excess degree is recruited (colored black) and edges are added to T which exist between v and all its neighbors that are not in T (these unrecruited neighbors are colored gray). The algorithm stops when all nodes are colored either gray or black, and the connected dominating set (CDS) is the set of non-leaf nodes in T. They showed that the above algorithm has a guaranteed approximation ratio of $O(\Delta)$, where Δ is the maximum degree of the network. We refer to the aforementioned algorithm as "Oracle" as it requires two-hop lookahead in order to compute the excess degree of nodes in $\mathcal{N}(\mathcal{B}(t)))$, a capability often missing in real online social networks. An example of an implementation of Guha and Khuller's Oracle can be found in Maiya and Berger-Wolf [29] (denoted *Expansion Sampling* in their work).

Interestingly, Guha and Khuller also showed that the approximation factor can be significantly improved when *three-hop lookahead* is exploited in a modified greedy step: recruit a *pair* of adjacent vertices (i.e. mark them black) and compare the yield in the number of gray nodes acquired in the neighborhood of this pair; at each step greedily select a pair of vertices or a single vertex that maximizes this yield. This modified greedy step surprisingly yields an approximation ratio of $O(\ln \Delta)$ instead of $O(\Delta)$. This additional lookahead, however, is not be available in several practical settings that we are interested in studying in this paper.

In practice, with *one-hop lookahead*, only *observed degree* information is available at nodes in $\mathcal{B}(t)$. This results in our MEED algorithm, which uses Guha and Khuller's Oracle approach using the expected excess degree in place of the true excess degree. MEED, however, requires the degree distribution of the network as side information. In the absence of degree distribution information, we show that for some random power law networks, a natural myopic online greedy algorithm of recruiting the node with the maximum observed degree approximates MEED – this is our MOD algorithm and also Maiya and Berger-Wolf's SEC scheme [29]. Expected value analysis as well as simulations in Sec. 8 show that MOD is a good heuristic when operating on realistic

[3] This could arise in "intelligence gathering" applications where analysts have pieced together the topology of an adversary network and now want to recruit the best (connected) set of influencers that will *cover* it.

social network such as those obeying a power law degree distribution.

We note that due the online nature of the different algorithms presented above, it is easy to apply the maximum budget criteria of MCNC and stop whenever the allocated budget n has been consumed. While the theoretical approximation guarantees may not strictly apply (unless $n = N$), we believe that in practice they still hold for the networks studied in this paper.

4. DATASETS & SIMULATION SETUP

We use the Enron email dataset as the running example throughout this work. The Enron email dataset contains data from a subpopulation of about 150 users, mostly senior management of Enron. This data was made public by the Federal Energy Regulatory Commission during its investigation of Enron. In total there are 36,692 nodes (unique email users) with average degree of 10.02 and average clustering coefficient of 0.5. The high clustering coefficient suggest great homophily in this network. The email corpus description can be found in Klimmt and Yang [25]. The version of the email graph we use can be downloaded from the Stanford's SNAP repository [37].

We also make use of other social network datasets – all of them, except Flickr, are available online at SNAP [37]. We now describe our datasets making use of N, $\langle k \rangle$, and c to denote the number of nodes, the average degree, and the clustering coefficient, respectively: Epinions ($N = 75{,}877$, $\langle k \rangle = 10.7$, $c = 0.26$) and Slashdot ($N = 82{,}168$, $\langle k \rangle = 0.1$, $c = 0.23$) online social networks, Wiki-talk ($N = 2{,}394{,}385$, $\langle k \rangle = 3.9$, $c = 0.2$) Wikipedia user-to-user discussion graph, EmailEU ($N = 265{,}214$, $\langle k \rangle = 2.8$, $c = 0.28$) the network email communication between users of a large European research institution, Youtube ($N = 1{,}134{,}890$, $\langle k \rangle = 5.3$, $c = 0.17$) friendship network of youtube.com users, and finally, Flickr dataset, a snapshot of an online photosharing network with $N = 1{,}715{,}255$ nodes and $\langle k \rangle = 12.2$, collected in Mislove et al. [33].

We also contrast our social network results with the results on three non-social networks. These networks can also be found online at SNAP [37]. Gnutella ($N = 62{,}561$, $\langle k \rangle = 4.7$, $c = 0.01$) a collection of merged P2P client snapshots collected in Ripeanu et al. [43], HepTh ($N = 27{,}770$, $\langle k \rangle = 25.4$, $c = 0.3$) a paper citation graph, and Amazon ($N = 334{,}863$, $\langle k \rangle = 5.5$, $c = 0.4$) the network of co-purchased products on the amazon.com website.

Simulation setup.

Unless stated otherwise our metrics consist of averages over 1,000 simulation runs. We use colored shadows in our plots to show the value of standard deviation plotted around the average. The shadow serves two proposes. First, its vertical width multiplied by $1.96/\sqrt{1000}$ gives approximately the 95% confidence intervals of our averages. Second, its value measures the variability between independent runs, by which we compare how consistently good (or bad) an algorithm performs. In our simulations $\mathcal{B}(0)$ includes a single node recruited uniformly at random from V. The order in which neighbors of a node appear on its list of neighbors is randomized from run to run to avoid arbitrary biases that may arise from the choice of node IDs in the dataset.

5. BFS & DFS ALGORITHMS

We begin our study by comparing the performance of two different approaches derived from two basic graph traversal algorithms: **Breadth-First Search (BFS)** and **Depth-First Search (DFS)**. BFS is chosen because it is widely used in network sampling [11, 26, 33, 35]. In these algorithms nodes in the $\mathcal{N}(\mathcal{B}(t))$ are recruited according to the time they were first observed (a node is observed when one of its neighbors is recruited). If we consider that nodes are put in a queue when they are first observed and then removed when they are recruited, then BFS employs a First In First Out discipline for the queue, recruiting the first observed node in $\mathcal{N}(\mathcal{B}(t-1))$, while DFS employs a Last In First Out discipline, recruiting the last observed node in $\mathcal{N}(\mathcal{B}(t-1))$. At each step a new, previously unrecruited node, is recruited such that at step $t = N - 1$ all nodes are recruited, i.e., $\mathsf{N}(\mathcal{B}(N-1)) = 0$.

Fig. 2a shows the average cover size $\langle \overline{W}(t) \rangle$ of BFS and DFS as a function of t on the Enron email network (recall that we average over 1,000 simulation runs). We find similar results on all of our social network datasets, see Figs. 5a-f. The simulations show while both BFS and DFS achieve the full coverage for $t \approx N$, BFS significantly outperforms DFS for all other values of t. To understand this difference, we qualitatively analyze the step in which a given node $v \in V$ with degree k_v is recruited.

Because both BFS and DFS follow edges to recruit nodes, the probability that v is first observed in $\mathcal{N}(\mathcal{B}(t))$ at step t is approximately $(\gamma_v/N)(1 - \gamma_v/N)^{t-1}$, where $\gamma_v = k_v/\langle k \rangle$ (this simple formula should be a good approximation in a configuration model where nodes are recruited independently by both algorithms; it also assumes $t \ll N$). Thus, large degree nodes tend to be observed earlier in the process than small degree nodes. As a FIFO policy recruits the earliest observed nodes from $\mathcal{N}(\mathcal{B}(t))$, BFS tends to recruit large degree nodes first, on the other hand, a LIFO policy recruits the latest observed nodes from $\mathcal{N}(\mathcal{B}(t))$, hence **DFS tends to recruit small degree nodes first.** This DFS result contradicts previous results in the literature [29], which we revisit in Sec. 9.

Note that Fig. 2a shows larger standard deviations for BFS than for DFS (although in some social networks the relative difference may be small). This is because the cover size of a non-neglegible fraction of the BFS runs deviates from the average. This instability is due to the strong dependence of the BFS cover size on the initial node $\mathcal{B}(0) = \{i\}$. As BFS explores the network in "waves" (expanding rings from i), the initial node selection may significantly impact BFS's cover size. Moreover, **we expect BFS to perform poorly on networks with a large degree of homophily** (as seen at the end of Sec. 6 in a simple regular lattice example). Homophily is the tendency of individuals to connect to similar individuals [30], thus creating patches of clustered nodes in the network. This means that if $v \in V$ is connected to $u \in V$ and $z \in V$, then u and z are more likely to be connected than random chance would allow. In addition, if v is not connected to $w \in V$ then u and z are more likely than random not to be connected to w. In such scenario it pays not to recruit both u and z together, as their neighbors significantly overlap with higher probability than random chance alone would allow.

DFS clearly avoids the above homophily problem by traversing the graph in depth first order. To increase the cover set size we only need to modify the LIFO recruitment policy without resorting to BFS's FIFO policy. In what follows (Sec. 6) we explore the use of Random Walks (RWs). As we see in the next section, a RW share commonalities with DFS in that it also traverses the graph from the last recruited node (however, a RW may try to recruit a node more than once). But, different from DFS, it allows recruitment of observed nodes in $\mathcal{N}(\mathcal{B}(t))$ irrespective of when the node was observed. One drawback of RWs – that fortunately can be easily mitigated via caching – is the possibility of recruiting already recruited nodes.

6. RW ALGORITHMS

Now let us analyze the cover size of **Random Walks (RWs) with one-hop lookahead**. Increased attention has been paid to random walks as a tool for network sampling [13–15, 27, 42] mostly due to its good statistical properties. In the RW algorithm $\mathcal{N}(\mathcal{B}(t))$ is still the set of all observed unrecruited nodes at time t. However, in a RW the node to be recruited at step $t+1$ is a random neighbor of the node recruited at step t, regardless of the time that the node was observed or even if it was already recruited. We begin our analysis assuming that a node that is recruited again at step t needs to be paid (that is, time advances even if no new recruitments were performed). We refer to this traditional RW algorithm as **RW with replacement (RW)**, in which nodes already recruited can be recruited again. At the end of this section we extend our analysis to the case of **RW without replacement (RWnr)** where already recruited nodes are "cached" so that recruiting nodes from $\mathcal{B}(t)$ does not count towards the recruitment budget.

The cover size of RWs with one-hop lookahead has been the subject of previous work [31]. However, we feel that one needs to exercise caution when interpreting the results in Mihail et al. [31]. Mihail et al. shows that a RW with one-hop lookahead finds the majority of nodes in sublinear time in an infinite configuration model with heavy tailed power law degree distribution. As our approach demonstrates below, covering finite networks is patently different from covering infinite networks. In particular, we show that for any given finite sized network, the discovery rate is never superlinear (this linear growth rate, however, can be large). Our model also allows us to predict with high accuracy the expected number of covered nodes as a function of $0 < t \leq n$.

Let us first analyze the performance of RW. In RW, the expected cover size at step t is

$$\begin{aligned}\langle \overline{W}(t)\rangle &= N - \sum_{\forall v\in V} P[\text{node } v \text{ is uncovered}] \\ &= N - \mathbf{q}\sum_{\forall v\in V} {}_{\mathcal{N}_a(v)}\mathbf{P}^t\underline{1},\end{aligned} \tag{1}$$

where $\mathbf{q}$ is a vector with the initial distribution of the random walk, $\underline{1}$ is a column vector of ones, and ${}_{\mathcal{N}_a(v)}\mathbf{P}$ is a taboo transition probability matrix defined by

$$[{}_{\mathcal{N}_a(v)}\mathbf{P}]_{ij} = \begin{cases} p_{ij}, & \text{if } i,j \notin \mathcal{N}(v), \\ 0, & \text{otherwise,}\end{cases}$$

with $\mathcal{N}_a(v)$ denoting the neighbourhood set of node v, including node v.

The above formula (1) requires complete topology knowledge and does not allow simple analytical solution. However, consider the following approximation to a RW. Nodes are recruited with replacement in i.i.d. fashion according to the stationary distribution of the random walk. Then, the expected cover size at step t would be given by

$$\begin{aligned}\langle \overline{W}(t)\rangle &= N - \sum_{\forall v\in V} P[\text{node } v \text{ is uncovered}] \\ &= N - \sum_{\forall v\in V} (1-\alpha_v)^t,\end{aligned} \tag{2}$$

where
$$\alpha_v = \frac{1}{2M}\left(k_v + \sum_{j\in\mathcal{N}_a(v)} k_j\right).$$

The above can be interpreted as a particular case of the coupon subset collection problem [2, 32]. Each step t, $t = 1,\ldots,N-1$, we draw a subset of "coupons", a subset of newly observed nodes in our terminology. We can observe a node either by sampling it directly (this corresponds to the term $k_v/(2M)$) or by sampling one of its neighbors (this corresponds to the term $\sum_{j\in\mathcal{N}_a(v)} k_j/(2M)$). The value of $k_v + \sum_{j\in\mathcal{N}_a(v)} k_j$ is known as the second neighbor degree [1]. In Appendix A we use matrix perturbation theory to show that (2) well approximates (1) for fast mixing RWs (see [4, 42] for fast mixing RW techniques).

Applying the Taylor series expansion and the fact that $2M = \langle k\rangle N$, we can write – for small t – the following approximation

$$\langle \overline{W}(t)\rangle \approx \frac{t}{\langle k\rangle N}\sum_{\forall v\in V}\left(k_v + \sum_{j\in\mathcal{N}_a(v)} k_j\right).$$

Next, we note that $\sum_{\forall v\in V}\sum_{j\in\mathcal{N}_a(v)} k_j = \sum_{\forall v\in V} k_v^2$, which yields

$$\langle \overline{W}(t)\rangle \approx t\frac{\langle k^2\rangle + \langle k\rangle}{\langle k\rangle},$$

where $\langle k^2\rangle$ is the empirical second moment . Thus, the cover process has approximately linear large growth rate in the initial phase, if the second moment $\langle k^2\rangle$ is large with respect to the average $\langle k\rangle$. Next, let us treat (2) as a continuous function of $t > 0$. The second derivative with respect to t is

$$\frac{d^2}{dt^2}\langle \overline{W}(t)\rangle = -\sum_{\forall v}(1-\alpha_v)^t \log^2(1-\alpha_v) < 0,$$

which is a concave function and, consequently, for any fixed network, the growth rate of the cover of RW with one-hop lookahead is at most linear.

Next we analyze the **RWnr** algorithm. Here, when the walker comes across an already recruited node, it does not recruit it again but rather randomly proceeds to one of its already cached neighbors. We define the algorithm evolution such that at each step we recruit a new node (that is, "virtual re-recruitments" do not advance t). RWnr closely follows the derivations for RW, with the only difference that now some recruitments cannot be "wasted" on already recruited nodes. The key observation here is that we need to keep track of the edges belonging to recruited nodes. We suggest the following approximation. Let $Z(t)$ be the number of edges not yet discovered at time t, then we approximate (details can be found in Appendix B)

$$\langle \overline{W}(t)\rangle \approx N - \sum_{v\in V}\prod_{i=0}^{t-1}\left(1 - \frac{1}{2\langle Z(i)\rangle}\left(k_v + \sum_{j\in\mathcal{N}_a(v)} k_j\right)\right), \tag{3}$$

where

$$\langle Z(t)\rangle \approx \sum_{(u,v)\in E} \prod_{i=0}^{t-1} \left(1 - \frac{k_u + k_v}{2\langle Z(i)\rangle}\right).$$

From (3) we note that RWnr and privileges nodes with large second neighbor degrees (just like RW does).

The plots in Fig. 2a show that RWnr significantly outperforms RW for large values of t. Note that for small values of t, the graphs in Fig. 2a shows that the performance of RW is similar to the performance of RWnr. This behavior is observed because for small values of t, $\langle Z(t)\rangle \approx \langle k\rangle N$, and thus the values of $\langle W(t)\rangle$ in eqs. (2) and (3) are similar.

The plots in Fig. 2a also show that RWnr significantly outperforms BFS. These results are consistent on all other datasets (refer to the plots in Figs. 5a-f) This curious fact can be explained by observation that BFS is affected by the graph homophily.

At first sight, however, it seems that both RW and RWnr should be impacted by homophily, as both have a tendency to "over-explore" a neighborhood, just as BFS. However, as we see next, this depends on the network and, in social networks, we believe that this is not likely the case. We contrast the performance of RW against BFS on two regular lattices. We choose these graphs because on regular lattices $\langle k^2\rangle/\langle k\rangle$ is small, leaving just homophily (for BFS) and the RW escape probability (the probability that a RW never revisits the same node or neighborhood again in an infinite network) as the primary factors in determining cover size for BFS and RW. Fig. 3 shows the performance of RW and BFS on regular 2D and 3D lattices. The plots show the cover size $\overline{W}(t)$ as a function of t where both lattices have approximately $N = 10^5$ nodes each. Observe that RW suffers from its tendency to return to the same nodes. On an infinite 2D regular lattice a RW returns to the same node with probability one while in a regular 3D lattice this probability is 0.34 [47]. On the other hand, BFS is affected by the clustering of the lattice, such that for most recruited nodes BFS in average covers approximately only one new node per step. A detailed analysis of these approximations can be found in Appendix C.

7. SI ALGORITHM

In this section we consider a different method, inspired by the **Susceptible-Infected (SI)** model in epidemiology: at step t recruit a node from $\mathcal{N}(\mathcal{B}(t))$ by randomly selecting one of the edges between $\mathcal{B}(t)$ and $\mathcal{N}(\mathcal{B}(t))$. Under SI node $v \in \mathcal{N}(\mathcal{B}(t))$ with $d(v,t)$ neighbors in $\mathcal{B}(t)$ is recruited at time t with probability $d(v,t)/\sum_{u\in\mathcal{N}(\mathcal{B}(t))} d(u,t)$. Our analysis of RW-based cover size relied on the fact that a node is always recruited with probability proportional to the degree of that node. This is no longer the case for SI-based cover sizes. Moreover, unlike the SI-related epidemic literature on infinite graphs [5], we will observe from our analysis that the probability that a particular node is recruited depends on t, the number of steps that have been executed.

Before we delve into the analysis of SI, we first show that we cannot ignore the impact of t on the degree distribution of nodes in $\overline{\mathcal{B}}(t)$, which, as we see next, becomes significantly different than the degree distribution across the whole network $\{p_k\}_{k=1,\ldots}$ as t gets larger.

7.1 The effect of t on the degree distribution of $\overline{\mathcal{B}}(t)$

In order to model the evolution of $B(t)$ and $\mathsf{N}(\mathcal{B}(t))$ we need to understand the impact of SI recruitment policy on the degree distribution of the nodes still left to recruit, $\overline{\mathcal{B}}(t)$. Fig. 4 shows the empirical Complimentary Cumulative Distribution Function ($P[K \geq k]$), denoted CCDF in the plots, of nodes' degrees in $\mathcal{B}(t)$ using SI over the Enron email network for $t \in \{1000,\ 2000{,}4000{,}8000{,}16000\}$ and against the CCDF of all the nodes in V. The empirical CCDF is averaged over 1 000 runs. Observe that even when t is reasonably small, e.g., $t = 1{,}000$, the tail of the CCDF of $\overline{\mathcal{B}}(t)$ is still significantly "lighter" than the tail of the CCDF of V. This is because large degree nodes are more likely to be recruited early to $\mathcal{B}(t)$; and as $\overline{\mathcal{B}}(t)$ is depleted of large degree nodes, the tail of the degree distribution of $\overline{\mathcal{B}}(t)$ gets "lighter". We use this property when analyzing the cover performance of SI.

7.2 Analysis of SI cover

We start our analysis by characterizing the evolution of the cover size as a function of t. Section 7.1 shows that the number of remaining nodes in $\overline{\mathcal{B}}(t)$ with degree $k = 1,\ldots,N-1$ is a function of t and k. Hence, our analysis divides the recruited nodes in $\mathcal{B}(t)$ into classes corresponding to different degrees. In particular, using mean field approximations we characterize $b_k(t)$, the fraction of nodes of degree k in G that are recruited by time t, $k = 1,\ldots,N-1$. As the number of nodes of degree k in $\overline{\mathcal{B}}(t)$ is given by $Np_k(1-b_k(t))$, the degree distribution of $\overline{\mathcal{B}}(t)$ can be approximated by $\{Cp_k(1-b_k(t))\}_{k=1,\ldots,N-1}$, where C is a normalizing constant. We now characterize the connections between nodes of various degrees between $\mathcal{B}(t)$ and $\overline{\mathcal{B}}(t)$. In the configuration model – described in Sec. 2 – the probability that a given node $u \in V$ of degree k is connected to a randomly chosen node $X \in V$ of degree $k_X = h$ is

$$p_{kh} = 1 - \left(1 - \frac{h}{2M}\right)^k.$$

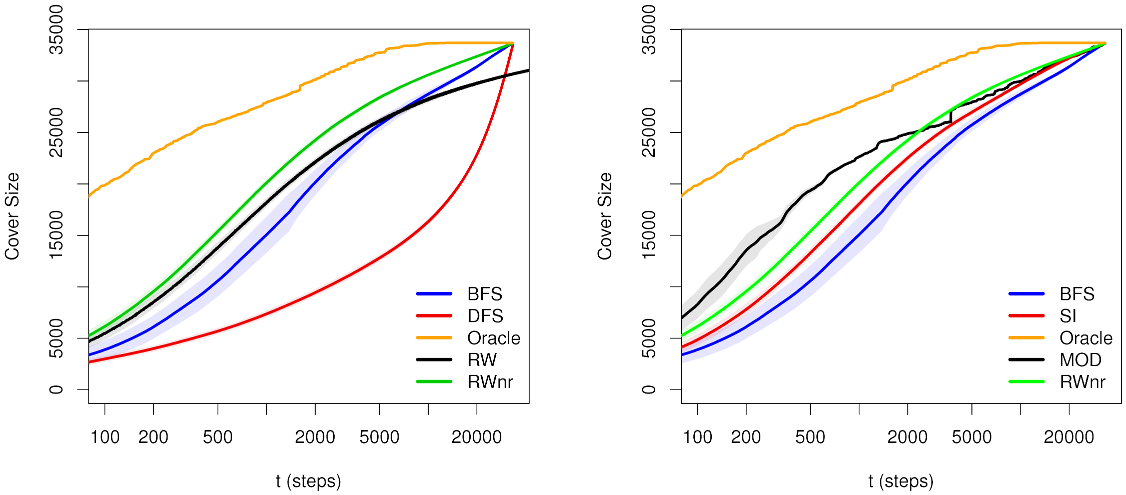
The probability that $X \in \mathcal{B}(t)$ given $k_X = h$ is $P[X \in \mathcal{B}(t)|k_X = h] = b_h(t)$, from which we approximate the probability that an infected node u of degree k has an infected neighbor of degree h by

$$P[X \in \mathcal{B}(t)\,,\, X \in \mathcal{N}_a(u)|k_X = h, u \in \overline{\mathcal{B}}(t), k_u = k] \approx b_h(t)p_{kh}.$$

Note that the above approximations assumes that $P[X \in \mathcal{B}(t)]$ does not depend $u \in \overline{\mathcal{B}}(t)$ (a reasonable approximation if X is randomly chosen from a large population of Np_h nodes in the graph). The probability that no degree h node connected to u has been recruited at time t can be approximated by $(1 - b_h(t)p_{kh})^{Np_h}$. If we condition on the event $u \in \overline{\mathcal{B}}(t)$, then the probability that u has at least one recruited neighbor is $\prod_h(1 - b_h(t)p_{kh})^{Np_h}$. Take notice that unconditioning the above on $P[u \in \overline{\mathcal{B}}(t)] = (1 - b_k(t))$ yields $\Upsilon_k(t)$, the probability that at time t a randomly chosen node of $\overline{\mathcal{B}}(t)$ of degree k is in $\mathcal{N}(\mathcal{B}(t))$, given by

$$\Upsilon_k(t) = (1 - b_k(t))\left(1 - \prod_h (1 - b_h(t)p_{kh})^{Np_h}\right).$$

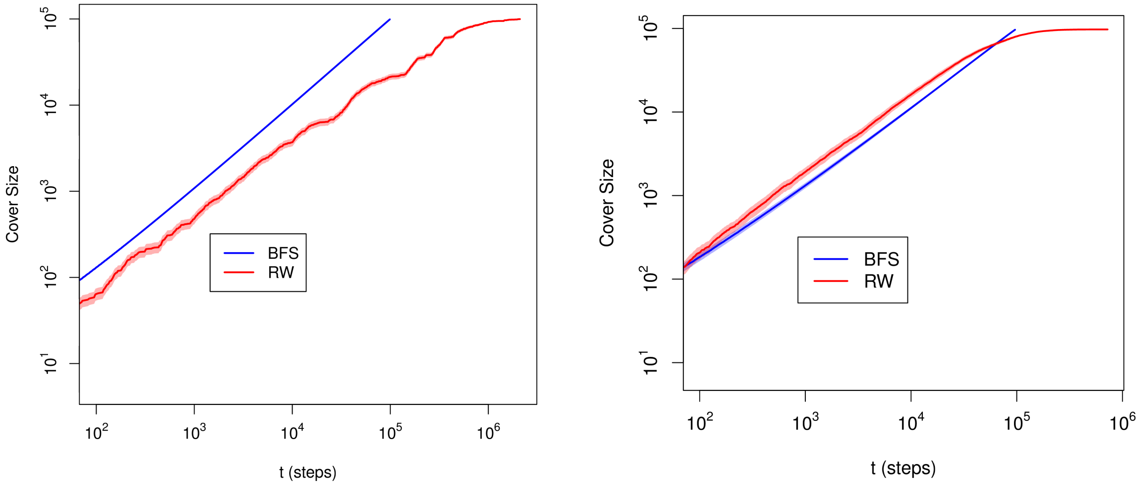
Finally, the expected number of observed nodes at time $t+1$,



(a)

(b)

Figure 2: **(Enron Network)** Empirical average cover size $\langle\overline{W}(t)\rangle$ as a function of $t \in \{1, N-1\}$. Fig. 2a compares Oracle, RW, RWnr, BFS, DFS and Fig. 2b compares Oracle, RWnr, SI, BFS, MOD. Shadows show double the standard deviation of 1,000 simulations; x-axis in log-scale.



(a) 2D Lattice

(b) 3D Lattice

Figure 3: $\langle\overline{W}(t)\rangle$ of RW against BFS on a 2D and 3D lattices. Simulations average over 1,000 runs, standard deviation shown as a colored shadow. Plots in log-log scale.

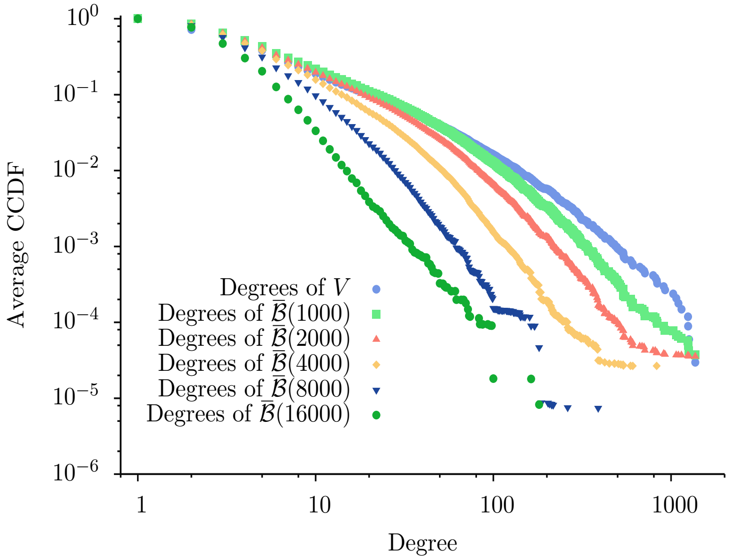


Figure 4: **(Enron Network)** Complementary Cumulative Distribution Function (CCDF) of $\overline{\mathcal{B}}(t)$ averaged over 1,000 runs, for $t \in \{1000, 2000, 4000, 8000, 16000\}$ recruited nodes in the Enron Network (Enron has approximately $N = 36{,}000$ nodes). As SI recruits nodes in the graph, the degree distribution of the remaining nodes in $\overline{\mathcal{B}}(t)$ suffers a dramatic change. Plot in log-log scale.

$\langle \mathsf{N}(\mathcal{B}(t+1)) \rangle$, is approximately

$$\langle \mathsf{N}(\mathcal{B}(t+1)) \rangle \approx \sum_k N p_k \Upsilon_k(t). \qquad (4)$$

Appendix D contrasts (4) against previous works, e.g. Pastor-Satorras and Vespignani [41].

Now we derive an equation for the dynamics of the number of sampled nodes $\mathcal{B}(t)$. When we sample a new edge (u, v) with $u \in \mathcal{B}(t)$ and $v \in \mathcal{N}(\mathcal{B}(t))$ from the frontier between $\mathcal{B}(t)$ and $\mathcal{N}(\mathcal{B}(t))$, the probability that v has degree k is proportional to $k p_k (1 - b_k(t))$. If we divide it by $p_k N$, we get the average increase in the fraction of sampled nodes of degree k, then:

$$b_k(t+1) = b_k(t) + \frac{k(1 - b_k(t))}{N \sum_h h p_h (1 - b_h(t))}. \qquad (5)$$

It is easy to check that

$$\langle B(t{+}1) \rangle = \sum_k N p_k b_k(t{+}1) = \sum_k N p_k b_k(t){+}1 = \langle B(t) \rangle{+}1,$$

we guarantee that at every step a new node is recruited.

We now contrast our mean field approximations against simulations. Fig. 6 shows $\langle \mathsf{N}(\mathcal{B}(t)) \rangle$ calculated according to (4) against the empirical value obtained from our simulations over two datasets, Enron and Gnutella. We plot the results in log-log scale to facilitate the comparison in respect to the relative error. Note that our approximation tracks the simulation results very well. Fig. 2b shows the SI cover size against the cover size of BFS and RWnr on the Enron email network. Note that the SI cover size is larger than the BFS cover size but smaller than that of RWnr. These results are consistent on all of the datasets analyzed, see Figs. 5a-f, (albeit sometimes the differences are small). The takeaway message from Fig. 2b and all the results from the plots in Figs. 5a-f is that there is much room for improvement. And while RWnr clearly outperforms all other methods in the Enron dataset, this gain all but disappears in other datasets (Epinions, Slashdot, and Flickr). In what follows we present an algorithm that consistently outperforms all of the algorithms studied so far.

8. EXPECTED EXCESS DEGREE MAXIMIZATION ALGORITHM

One lesson to take away from Guha and Khuller's Oracle is that knowing which node in $\mathcal{N}(\mathcal{B}(t))$ has the largest excess degree is crucial to achieving a good cover. While the excess degree of nodes in $\mathcal{N}(\mathcal{B}(t))$ is not available to us, we may still be able to estimate them from the available information. Let $d(v, t)$ be the *observed degree* of v at step t. Consider a large degree node $v \in \overline{\mathcal{B}}(t)$. We expect v to also have a large observed degree $d(v, t)$. However, depending on the degree distribution of the nodes in $\mathcal{N}(\mathcal{B}(t))$, there could be a "saturation point", where the observed degree of v, $d(v, t)$, is so large that the excess degree $k_v - d(v, t)$ is small. In this section we propose an algorithm, denoted **Maximum Expected Excess Degree (MEED)**, that at step $t > 0$ finds a node in $\mathcal{N}(\mathcal{B}(t))$ that has approximately the largest expected excess degree of all nodes in $\mathcal{N}(\mathcal{B}(t))$. In what follows we drop t from our notation for the sake of conciseness.

Let $\langle k | d(v) \rangle$ denote the expected degree of a node $v \in \mathcal{N}(\mathcal{B}(t))$ with observed degree $d(v)$. Using $\langle k | d(v) \rangle$ we propose the Maximum Expected d-Excess Degree (MEED) heuristic, which chooses the next recruited node $v^\star$ at step t as to maximize the expected excess degree of the recruited node. Another way to describe the MEED scheduler is through a partial order of the nodes in $\mathcal{N}(\mathcal{B}(t))$ in respect to their expected excess degree. For two nodes $u, v \in \mathcal{N}(\mathcal{B}(t))$ we say $u \geq_k v$ iff $\langle k - d(u) | d(u) \rangle \geq \langle k - d(v) | d(v) \rangle$. Thus, at step t the MEED algorithm recruits node $v^\star \in \mathcal{N}(\mathcal{B}(t))$ if $v^\star \geq_k v$, $\forall v \in \mathcal{N}(\mathcal{B}(t))$.

Next we obtain an approximation of $\langle k{-}d | d \rangle$ using $\{p_k\}_{k=1,\ldots}$. For now we assume $\{p_k\}_{k=1,\ldots}$ is given to us. Later we show that for some important families of random networks, the node with maximum observed degree is also the node with the maximum expected excess degree. Let $\zeta_k(t)$ be the probability that a random node in $\bar{\mathcal{B}}(t)$ has degree k. Note that $\zeta_k(0) = p_k$ as, by definition, the initial node in $\mathcal{B}(0)$ is randomly sampled from V. In general we have $\zeta_k(t) = Cp_k(1 - b_k(t))$ (as stated in Sec. 7), where C is a normalization constant. In what follows we omit t for the sake of conciseness.

Let $\mathcal{N}^{(d)}(\mathcal{B}(t)) \subseteq \bar{\mathcal{B}}(t)$ denote the set of nodes with at least d recruited neighbors. Note that $\mathcal{N}^{(1)}(\mathcal{B}(t)) = \mathcal{N}(\mathcal{B}(t))$ and that $\mathcal{N}^{(0)}(\mathcal{B}(t)) = \bar{\mathcal{B}}(t)$. Under the configuration model we can determine $\{\mathcal{N}^{(d)}(\mathcal{B}(t))\}_{d=1,\ldots}$ and $\mathcal{W}(t)$ through the following process that dynamically assigns nodes from $\bar{\mathcal{B}}(t)$ to these sets. Let's assume $N \gg 1$ so we do not need to worry about self-loops. Detach all nodes $v \in V$ from their neighbors such that node v with degree k_v has k_v "active stubs".[4] Iteratively select an active stub in $\mathcal{B}(t)$ to a random active stub in V. Whenever an active stub of a node $u \in \mathcal{N}^{(d)}(t)$, $d \in \{0, 1, \ldots\}$, is selected, we add v to $\mathcal{N}^{(d+1)}(t)$ and mark both stubs of the edge "inactive", that is, we promote u to $\mathcal{N}^{(d+1)}(t)$ but reduce its active degree by one. The following recursion describes the degree distribution $\{\zeta_k^{(d+1)}\}_{k=d+1,\ldots}$ of the nodes in $\mathcal{N}^{(d+1)}(t)$ in terms of the degree distribution $\{\zeta_k^{(d)}\}_{k=d,\ldots}$ of nodes in $\mathcal{N}^{(d)}(t)$

$$\zeta_k^{(d+1)} = \frac{(k-d)\zeta_k^{(d)}}{\langle k \rangle_{\zeta^{(d)}} - d}, \quad k \geq d, \tag{6}$$

and $\zeta_k^{(d)} = 0$ for $k < d$, where $\langle k \rangle_{\zeta^{(d)}} \equiv \sum_{k \geq 0} k\zeta_k^{(d)}$ is the average degree of nodes with at least d recruited (black) neighbors.

We now retrieve $\langle k|d \rangle$ from $\langle k \rangle_{\zeta^{(d)}}$ using the fact that $\mathcal{N}^{(0)} \supseteq \mathcal{N}^{(1)} \supseteq \cdots$. Let N_d be the number of nodes in $\mathcal{N}(\mathcal{B}(t))$ with observed degree d. Note that $N_1 = \mathsf{N}(\mathcal{B}(t))$. For any two sets A, A', such that $A' \subseteq A$, the following holds: $\text{vol}(A - A') = \text{vol}(A) - \text{vol}(A')$, where $\text{vol}(B)$ is the volume of the set B, i.e. the sum of the all the degrees of the nodes in B. Considering $A = \mathcal{N}^{(d)}$ and $A' = \mathcal{N}^{(d+1)}$ it is easy to show that

$$\langle k|d \rangle = \left\langle \frac{N_d}{N_d - N_{d+1}} \right\rangle \langle k \rangle_{\zeta^{(d)}} - \left\langle \frac{N_{d+1}}{N_d - N_{d+1}} \right\rangle \langle k \rangle_{\zeta^{(d+1)}} .$$

We approximate the expectations of $\langle N_d/(N_d - N_{d+1}) \rangle$ and $\langle N_{d+1}/(N_d - N_{d+1}) \rangle$ using the observed values of N_{d+1} and N_d.

Our calculations of $\langle N_d/(N_d - N_{d+1}) \rangle$ and $\langle N_{d+1}/(N_d - N_{d+1}) \rangle$ should be used with caution as they do not consider the extra density of connections inside $\mathcal{B}(t)$ created by the MEED recruitment process. Taking this bias into account is not trivial and is the subject of future work. However, in our MEED simulations we observe that $N_{d+1} \ll N_d$ for large d, and, under such scenario, it is reasonable to make the following simplification $\langle k - d|d \rangle \approx \langle k - d \rangle_{\zeta^{(d)}}$.

Unfortunately, obtaining $\langle k - d \rangle_{\zeta^{(d)}}$ still requires knowing $\zeta_k^{(d)}$, $\forall k, d$, which in turn requires knowing the degree distribution. Note, however, that MEED simplifies to an algorithm that always selects the node $v^\star$ with the maximum observed degree in $\mathcal{N}(\mathcal{B}(t))$ if $\langle k - d(v^\star)|d(v^\star) \rangle \geq \langle k - d|d \rangle$, $d = 1, \ldots, d(v^\star)$. We denote this simplified MEED heuristic **Maximum Observed Degree (MOD)**.

We now see if we should expect to find the property $\langle k - d(v^\star)|d(v^\star) \rangle \geq \langle k - d|d \rangle$, $d = 1, \ldots, d(v^\star)$ in real social networks. In what follows we show that, under certain conditions, two of the most relevant social network models – power law and Erös-Rényi random networks – have this desired property. We later complement these findings with simulation results in the Enron network. In Appendix E we show that

$$\langle k - d \rangle_{\zeta^{(d)}} = \left. \frac{\left(\frac{\partial^{d+1}}{\partial z^{d+1}} H(z) \right)}{\left(\frac{\partial^{d}}{\partial z^{d}} H(z) \right)} \right|_{z=1} , \quad d \geq 1 , \tag{7}$$

where $H(z) = \sum_k z^k \zeta_k$ is the probability generating function (p.g.f.) of ζ_k, and $\partial^d H(z)/\partial z^d$ is the d-th derivative of $H(z)$ with respect to z. Next we use (14) to obtain approximations of $\langle k - d \rangle_{\zeta^{(d)}}$ for Erdös-Rényi and power law networks.

8.1 Maximizing d-excess Degrees in Erdös-Rényi Networks

In an Erdös-Rényi (ER) graph $G(N, q)$ the degree distribution is Binomially distributed, $p_k = \binom{N-1}{k} q^k (1-q)^{N-1-k}$, $k = 1, \ldots, N-1$. We now consider the case where t is small enough such that $\zeta_k(t) \approx p_k$. Then, the probability generating function of ζ_k is $H(z) = (1 - q + qz)^{N-1}$. In such scenario applying (14) yields

$$\langle k - d \rangle_{\zeta^{(d)}} = \frac{(N-1)! q^{d+1}/(N-d-2)!}{(N-1)! q^d/(N-d-1)!} = (N - d - 1)q.$$

As $N \gg 1$ we observe that the average d-excess degree can be approximated by the average degree of the network, $\langle k - d \rangle_{\zeta^{(d)}} = \langle k \rangle$, independent of d. Thus, if $N_{d+1} \ll N_d$, $\forall d$, the scheduler can be degree agnostic and, thus, MOD approximates MEED.

8.2 Maximizing d-excess Degrees in Power Law Networks

In a power law network we cannot ignore the effect of t on $\zeta_k(t)$. Hence, we consider $\zeta_k(t)$ to be power law distributed with an exponential cut-off. The exponential cut-off approximates the behavior observed in Fig. 4. Moreover, in a variety of real world networks the degree distribution can be well approximated by power law distributions with exponential cut-offs [3, 38]. Let

$$\zeta_k = \frac{k^{-\tau} C_t^k}{\text{Li}_\tau(C_t)} , \quad \text{for } k \geq 1,$$

where $C_t < 1$ is a parameter that depends on $\mathcal{B}(t)$ and t and the normalization factor $\text{Li}_h(x) = \sum_{k=1}^{\infty} x^k/k^h$ is the h-th polylogarithm function of x. The probability generating function of ζ_k assumes the form $H(z) = \text{Li}_\tau(zC_t)/\text{Li}_\tau(C_t)$. In Appendix E we use $H(z)$ to show that:

- If $\tau = 1$, $\langle k - d \rangle_{\zeta^{(d)}} \approx \frac{C_t d}{1 - C_t}$. and thus the node with the largest observed degree should be recruited.

- If $\tau = 2$, $\langle k - d \rangle_{\zeta^{(d)}} \approx \Gamma_d$, where $\frac{d}{1+1/d} < \Gamma_d \left(\frac{1-C_t}{C_t} \right) < d + 1$, which implies that for $\langle k - (d + a) \rangle_{\zeta^{(d+a)}} > \langle k - (d) \rangle_{\zeta^{(d)}}$, $a = 2, 3, \ldots$ and $d = 1, \ldots$. Thus, at step $t + 1$ we should recruit the node $v \in \mathcal{N}(\mathcal{B}(t))$ with either the largest or second largest value $d(v)$. We believe that these bounds can be improved to show that the node

[4] This stub analogy is extensively used to describe configuration models [40].

with the largest observed degree should be recruited. An interesting observation is that $\langle k-d \rangle_{\zeta(d)}$ increases with C_t and diverges as $C_t \to 1$.

- With $C_t \to 1$ and $\tau > 0$: The case where $C_t \to 1$ represents the case where $\{p_k\}$ is a pure power law distribution and t has little impact on the degree distribution of $\overline{\mathcal{B}}(t)$. This case is only of theoretical interest as no real world network degree distribution can match an infinite support power law degree distribution. In this scenario $\langle k-d \rangle_{\zeta(d)} < \infty$, $\forall d \leq \lceil \tau - 1 \rceil$, and $\langle k - \lceil \tau \rceil \rangle_{\zeta(\lceil \tau \rceil)} \to \infty$. This means that the node with observed degree degree at least $\lceil \tau \rceil$ should be recruited. We believe that this result can be strengthen to show that $\langle k-d|d \rangle$ monotonically increases with d.

Thus, we conclude that under the above networks MOD is a good approximation to MEED.

8.3 Simulations, Excess Degree & MOD

The above analysis suggests that in some power law networks and in ER networks, $\langle k-d|d \rangle$ increases with d. An important question is whether we observe this phenomenon in practice. Fig. 7 shows estimates of $\langle k-d|d \rangle$ for $d \in \{1, 2, 4\}$ obtained from simulating SI on the Enron network (these estimates are averaged over 1,000 runs). We choose SI to estimate $\langle k-d|d \rangle$ instead of MEED as under MEED very few nodes with observed degrees greater than two remain in $\mathcal{N}(\mathcal{B}(t))$, $t > 0$, thus making the estimates unreliable. In the figure we observe that for t approximately in the range $\{1, \ldots, 2/3N\}$ we have $\langle k-1|1 \rangle < \langle k-2|2 \rangle < \langle k-4|4 \rangle$ and for t approximately in the range $\{2/3N+1, \ldots, N-1\}$ we have $\langle k-1|1 \rangle < \langle k-2|2 \rangle \approx \langle k-4|4 \rangle$. Hence we expect that MOD is a good candidate to approximate MEED on Enron network.

Fig. 2b and Figs. 5a-f show that the MOD heuristic outperforms, sometimes significantly, all previous lookahead one algorithms on all social network datasets. We, however, still notice a significant gap between the Oracle and MOD, which we believe can be reduced using side information to improve the estimation of $\langle k-d \rangle_{\zeta(d)}$.

9. SUMMARY, CONCLUSIONS & RELATED WORK

We have considered the problem of providing an online algorithm that, by recruiting nodes through their neighbors, greedily maximizes the network cover of an online social network. In our setting the network topology was unknown and the only topological information available came from the identity of the neighbors of already recruited nodes.

In this scenario, we have evaluated the efficacy of existing network sampling algorithms (BFS, DFS, RW) and proposed a novel algorithm, Maximum Expected Excess Degree (MEED), inspired by the greedy approximation to the minimum connected dominating set of Guha and Khuller [16], which uses two-hop lookaheads (and, thus, denoted "Oracle" in this work) to recruit at every step the node with the largest excess degree. The MEED heuristic seeks to maximize the expected excess degree of nodes with the help of degree distribution side information (if available). In the absence of degree distribution information, we have shown that on random power law and Erdös-Rényi networks MEED can be approximated by MOD (Maximum Observed Degree), a greedy heuristic that at every step recruits the node with the largest observed degree. We have shown through extensive simulations on real world social network datasets that MOD outperforms all other algorithms, often quite significantly.

We have also provided theoretical analysis of RWs (with and without replacement) and of an algorithm inspired by the Susceptible-Infected epidemic model, which we denoted SI. Our theoretical analysis, to the best of our knowledge, stands as a contribution on its own. We expect that our formulas can aid practitioners in predicting the cover sizes of these algorithms when degree distribution side information is available.

Finally, we have uncovered a puzzling previously unknown fact about DFS: DFS performs remarkably poorly on social networks. In fact, DFS seems to avoid recruiting nodes with large excess degrees. We have argued that this is due to its tendency to keep large degree nodes at the bottom of the its recruitment queue. We note in passing that this property of DFS may find applications in undercover military operations where one seeks to recruit target individuals with the minimum exposure (number of connections) to unrecruited targets.

Related Work.

The connections between our work and the literature on MCDS were already presented in Sec. 3. In this section we review the remaining related literature. The work most related to ours is Maiya and Berger-Wolf [29]. Maiya and Berger-Wolf presents a simulation study of the cover sizes (among other metrics) of different algorithms, including BFS, DFS, Oracle (which they denote *Expansion Sampling*), and MOD (which they denote *Sample Edge Count*). Their work considers social (e.g., Enron email) and technological (e.g., Amazon product co-purchase) networks. Surprisingly, their conclusions are remarkably different than ours, arguing that DFS outperforms BFS, RWs, and, most importantly, MOD.

Maiya and Berger-Wolf [29] shows the performance of these algorithms, Figs. 4(e) and 4(f) in their work, on the (non-social) HepTh and Amazon networks (see Sec. 4 for a brief description of these datasets). In order to understand the discrepancy of our conclusions, here we also include simulations of these same two datasets. We show our results in Figs. 8a and 8b. Note the remarkable difference to our social network results in Figs. 2a-b and Figs.5a-f. Indeed, DFS experiences a great improvement in performance while MOD worsens dramatically.

To test whether the sudden improvement of DFS and decline of MOD can be mostly attributed to the more structured nature of these two networks in respect to our social networks, we artificially add randomness to these network by randomly rewiring all endpoints (while making sure nodes form a single connected component). Figs. 8c and 8d provide the results over these randomized networks. The results are unmistaken, adding randomness makes MOD incontestably superior to all other algorithms (in HepTh it even matches the performance of Oracle) and DFS is again noticeably inferior. It remains an open question whether, in respect to the network cover problem, most social networks are more similar to "random networks" or more similar to "structured networks". Our datasets and simulation results suggest the former.

In an earlier preliminary work (Lim et al. [45]) we proposed SI under the name *Randomized Expansion Sampling*

(RXS) used to find the most central nodes in a network. However, in Lim et al. we did not analyze SI cover size. Random walks with lookahead have been the subject of a number of works. Cooper and Frieze [8] studied the cover time of RW. Mihail et al. [31] shows that a RW with one-hop lookahead finds the majority of nodes in sublinear time in an infinite configuration model with heavy tailed power law degree distribution. Our analysis, however, shows that the cover rate is at most linear for any value of t. Adamic et al. [1] proposed a RW with two-hop lookahead and analyzed its cover time on an infinite network. The fast cover of RWs has been used in the context of decentralized search, e.g., when searching for content on unstructured P2P networks (see [20, 24, 28, 44, 46] and references therein).

A closely related problem is the *influence maximization* problem. The influence maximization problem considers that each recruited individual invites its neighbors who can be recruited with some probability. The purpose is to select a set "influential" individuals, in order to cause a cascade of recruitments in the network. Network topology is generally assumed to be known [22]. This problem was first proposed by Domingos and Richardson [9] (please refer to the review in Kleinberg [23] for other references). Directly related to online recruitment, Hartline et al. [17] and Bhattacharya et al.[6] analyze the (paid) recruitment of consumers of a product under the assumption of perfect network knowledge.

APPENDIX

A. RW & INDEPENDENT EDGE SAMPLING APPROXIMATION

Let us assume that the random walk starts from a stationary distribution ($\pi_v = k_v/(2M)$). This is not too restrictive assumption, since in the configuration model after the first step the random walk is asymptotically – with respect to the network size – in the stationary regime. And in the general case, we can add auxiliary uniform jumps, which has been shown [4] to significantly reduce mixing time.

We now show that if the random walk starts in steady state, that is $q_v = \pi_v$, the expression using the complete topological information (1) can be approximated by expression (2). It is enough to consider just one term in (1) corresponding to one node. Without loss of generality, we take $v = 1$ and assume that nodes $u = 2, ..., k_1 + 1$ are neighbours of node 1. Partition the stationary distribution π as $[\pi_1\ \pi_2]$, where π_1 corresponds to node 1 and all its neighbours and π_2 corresponds to all the other nodes. Then, we need to evaluate

$$\pi_{\ \mathcal{N}(v)} P = \pi_2 P_{22}^k \underline{1}.$$

Let us demonstrate that π_2 properly normed is close to $\tilde{\pi}_2$, the quasi-stationary distribution of the substochastic matrix P_{22}:

$$\pi_2 P_{22} = \lambda \pi_2.$$

Even though P_{22} is substochastic, if the graph is large enough, P_{22} will be close enough to a stochastic matrix so that we can apply perturbation theory. Let us consider the following perturbation equation

$$(\pi_2^{(0)} + \varepsilon\pi_2^{(1)} + ...)(S_{22} - \varepsilon D) = (1 - \varepsilon\lambda^{(1)} + ...)(\pi_2^{(0)} + \varepsilon\pi_2^{(1)} + ...), \tag{8}$$

where S_{22} is the stochastic complement (it describes the transitions of the censored Markov chain) and

$$\varepsilon D = P_{21}[I - P_{11}]^{-1} P_{12},$$

with ε as some scaling parameter. Equating terms in (8) with ε^0, we obtain

$$\pi_2^{(0)} S_{22} = \pi_2^{(0)},$$

from which it follows that $\pi_2^{(0)} = c\pi_2$. Then, equating terms in (8) with ε^1, we obtain

$$\pi_2^{(1)} S_{22} - \pi_2^{(0)} D = \pi_2^{(1)} - \lambda^{(1)} \pi_2^{(0)}.$$

Multiplication the above equation by $\underline{1}$ from the right yields

$$\lambda^{(1)} = \frac{1}{\pi_2 \underline{1}} \pi_2 D \underline{1},$$

and consequently,

$$\lambda \approx 1 - \varepsilon\lambda^{(1)} = 1 - \frac{1}{\pi_2 \underline{1}} \pi_2 P_{21} [I - P_{11}]^{-1} P_{12} \underline{1}.$$

Next, from the defining equations for the stationary distribution, we have

$$\pi_2 P_{21} [I - P_{11}]^{-1} = \pi_1,$$

which leads to

$$\lambda \approx 1 - \frac{1}{\pi_2 \underline{1}} \pi_1 P_{12} \underline{1}.$$

First, we note that $\pi_2 \underline{1} \approx 1$. And second, we note that $\pi_1 P_{12} \underline{1}$ is the number of links from the neighbourhood set $\mathcal{N}(1)$ to all the other links divided by the total number of links. Since the number of links to the outside of the neighbourhood set is much large than the number of links inside the set, $\lambda \approx (1 - \alpha_1)$ and expression (1) can be approximated by (2).

B. RW WITHOUT REPLACEMENT

Here we provide a brief description of the steps used to derive the approximation in equations (3).

As in the case of the analysis of RW with replacement, we assume that we sample nodes i.i.d. fashion according to the stationary distribution. Recall that $Z(t)$ denotes the number of edges not yet discovered at time t. Then, given the history $\{Z(i), i = 0, 1, ..., t - 1\}$ with $Z(0) = M$, the probability that the node v is uncovered at time t is given by

$$P[v \in \mathcal{W}(t) | \{Z(i), i = 1, ..., t - 1\}] = \prod_{i=0}^{t-1} \left(1 - \frac{1}{2Z(i)} \left(k_v + \sum_{j \in \mathcal{N}_a(v)} k_j\right)\right),$$

Similarly, given the history $\{Z(t), t = 0, 1, ..., t - 1\}$, we can calculate the probability that the link (u, v) is yet uncovered at time step t

$$P[\text{link } (u, v) \text{ is uncovered at } t | \{Z(i), i = 1, ..., t - 1\}] = \prod_{i=0}^{t-1} \left(1 - \frac{k_u + k_v}{2Z(i)}\right).$$

Viewing the link coverage problem as a subset coupon collector problem, we obtain

$$\langle Z(t) | \{Z(i), i = 1, ..., t - 1\} \rangle = \sum_{(u,v) \in E} \prod_{i=0}^{t-1} \left(1 - \frac{k_u + k_v}{2Z(i)}\right).$$

If we use the above equation recursively starting at $Z(0) = M$ and using the approximate value of $\langle Z(t)\rangle$, we obtain the second equation in the approximate formulas (3).

C. BFS ON 2D AND 3D GRIDS

We first observe that BFS discovers the graph by considering progressively large *spheres*. At step k the boundary of the set of discovered nodes is made by all nodes k hops away from the first node. When nodes are embedded in a metric space, the sphere has the property to be the solid with the smallest surface for a given volume. This property justifies intuitively why BFS discovers slowly new nodes.

We first consider an infinite 2D grid. For the 2D grid there are $4t$ nodes at the boundary at step t, then every node on the boundary contributes on the average to add $1 + 1/t$ new nodes when it is visited, even if each of them has degree 4 and at least 2 neighbors not discovered.

For a 3D grid the boundary has size $2 + 4t^2$, then every nodes contributes to discover $(1 + 2(t+1)^2)/(1 + 2t^2)$. Here a node on the boundary has 6 neighbours and at least 3 not discovered at the begin of the step, but still its average contribution converges to 1 as t diverges.

The reason for this slow increase is due to the fact that 1) nodes at the boundary have many connections (roughly half of them, i.e. $\langle k\rangle/2$) to the interior of the sphere, i.e. to already explored or discovered nodes, but also 2) the outgoing connections point to the same nodes. In fact almost every node that is going to be discovered will be discovered through $\langle k\rangle/2$ connections (those pointing towards the sphere). Then, as the sphere becomes larger and the border effects negligible, every node only contributes by discovering one new node.

D. SI ALGORITHM V.S. SI EPIDEMIC

The literature on SI epidemic models is so vast and rich that we dedicate a section of our appendix to contrast our results against previous works. We also provide limited commentary on approximations of our equations. We first note that the previous literature is often interested in a continuous-time version of our SI model. Recall that in our scenario t is the number of recruited nodes, while the related literature considers t as time [34, 41], meaning that at time t, the average number of "infected nodes" may be smaller or larger than t.

However, it is an interesting exercise to try to connect the framework provided by Pastor-Satorras and Vespignani [41] with our approach.[5] Consider a **continuous-time SI epidemic** on G. In an SI epidemic an infected node can be thought play the role of a recruited node and susceptible nodes play the role of non-recruited nodes. Let λ be the per-unit time infection rate, that is, an infected node contaminates (recruits) a susceptible node during a time inteval $\Delta \to 0$ with probability $\lambda\Delta t$, regardless of the state of the infection. Let t' denote the **wall clock time** of the SI epidemic process and $\rho_k(t')$ be the probability that a node with k links is infected. Then [41],

$$\frac{d\rho_k(t')}{dt} = \lambda k(1 - \rho_k(t'))\Theta(\rho(t')), \quad \forall k, \tag{9}$$

where $\rho(t') = (\rho_1(t'), \ldots, \rho_N(t'))$ and

$$\Theta(\rho(t')) = \langle k\rangle^{-1} \sum_k k p_k \rho_k(t') \tag{10}$$

is the probability that any given link points to an infected node. Note that our t is the number of infected nodes, and thus,

$$\sum_k N p_k \rho_k(t') = t. \tag{11}$$

Analytically adding the constraint (9) into the set of equations in (11) is no trivial task, but it can be done numerically. Let $\rho'(t)$ be the solution of (11) with the added constraint (9). It is unclear, however, how the cover size can be derived from (10).

The main difficulty in the above approach is the mapping between wall-clock time and number of recruited nodes. Our formulation in Section 7 avoids this problem entirely and this is why it is much simpler to compute.

E. THE EXPECTED D-EXCESS DEGREE FROM THE P.G.F. OF ζ_K

Our analysis of $\langle k - d|d\rangle$ begins by breaking down $\langle k - d\rangle_{\zeta^{(d)}}$ into the derivatives of the generating function of ζ_k:

$$\begin{aligned}
\langle k - d\rangle_{\zeta^{(d)}} &= \sum_{k=d}^{\infty} (k-d)\zeta_k^{(d)} \\
&= \sum_{h=0}^{\infty} \frac{h(h+1)\zeta_{h+d}^{(d-1)}}{\langle k\rangle_{\zeta^{(d-1)}} - (d-1)} \\
&= \frac{\sum_{h=0}^{\infty} h(h+1)(h+2)\zeta_{h+d}^{(d-2)}}{\prod_{i=d-2}^{d-1}(\langle k\rangle_{\zeta^{(i)}} - i)} \\
&= \frac{\sum_{h=0}^{\infty} h(h+1)\cdots(h+d)\zeta_{h+d}}{\prod_{i=0}^{d-1}(\langle k\rangle_{\zeta^{(i)}} - i)} \\
&= \frac{1}{\prod_{i=0}^{d-1}(\langle k\rangle_{\zeta^{(i)}} - i)} \left.\frac{\partial^{d+1} H(z)}{\partial z^{d+1}}\right|_{z=1},
\end{aligned} \tag{12}$$

where $H(z) = \sum_k z^k \zeta_k$ is the probability generating function of ζ_k. The first several equalities are a consequence of successive applications of (6). The last equality comes from the applying the $d+1$-th derivative to $H(z)$. Multiplying both sides of (12) by $\prod_{i=0}^{d-1}(\langle k\rangle_{\zeta^{(i)}} - i)$ yields

$$\prod_{i=0}^{d}(\langle k\rangle_{\zeta^{(i)}} - i) = \left.\frac{\partial^{d+1} H(z)}{\partial z^{d+1}}\right|_{z=1}, \tag{13}$$

where $\langle k\rangle_{\zeta^{(0)}} = \sum_k k p_k$. Substituting (13) into the denominator of the l.h.s. of (12) yields

$$\langle k - d\rangle_{\zeta^{(d)}} = \left.\frac{\left(\frac{\partial^{d+1} H(z)}{\partial z^{d+1}}\right)}{\left(\frac{\partial^{d} H(z)}{\partial z^{d}}\right)}\right|_{z=1}, \quad d \geq 1. \tag{14}$$

An important property of the polylogarithm function is that for any constant $C > 0$,

$$\frac{\partial \mathrm{Li}_\tau(Cz)}{\partial z} = \frac{\mathrm{Li}_{\tau-1}(Cz)}{z}.$$

IN what follows we consider three special cases: $\tau = 1$, $\tau = 2$, and $C_t \to 1$.

[5] The authors would like to acknowledge N. Perra and A. Baronchelli for helpful discussions on this topic and N. Perra for suggesting the connection.

Case $\tau = 1$: Let's first consider the case $\tau = 1$, which implies that ζ_k is very heavy tailed. We use the fact that $\mathrm{Li}_1(a) = -\log(1-a)$. With $\tau = 1$ we have

$$\frac{\partial H(z)}{\partial z} = -\frac{1}{\log(1-C_t)} \frac{\partial \mathrm{Li}_1(C_t z)}{\partial z} = -\frac{1}{\log(1-C_t)(1-C_t z)}$$

and thus,

$$\frac{\partial^d H(z)}{\partial z^d} = -\frac{1}{\log(1-C_t)} \frac{(d-1)! C_t^d}{(1-C_t)^d},$$

which yields

$$\langle k-d \rangle_{\zeta^{(d)}} \approx \frac{C_t d}{1-C_t},$$

showing that the d-excess degree grows linearly with d when $\tau = 1$. Then, we conclude that recruiting the node with the largest observed degree from $\mathcal{N}(\mathcal{B}(t))$ maximizes the expected cover increase.

Case $\tau = 2$: Another case of interest is $\tau = 2$. In this case note that

$$\langle k-d \rangle_{\zeta^{(d)}} \approx \left. \frac{\frac{\partial^d}{\partial z^d}\left(-\log(1-C_t z)/z\right)}{\frac{\partial^{d-1}}{\partial z^{d-1}}\left(-\log(1-C_t z)/z\right)} \right|_{z=1}.$$

Using the Taylor series expansion of

$$-\log(1-C_t z)/z = \sum_{h\geq 1} \frac{C_t^h z^{h-1}}{h}$$

yields

$$\langle k-d \rangle_{\zeta^{(d)}} \approx \Gamma_d\,,$$

where

$$\begin{aligned}\Gamma_d &= \frac{\sum_{h\geq d+2} \frac{C_t^h (h-1)!}{h(h-d-2)!}}{\sum_{h\geq d+1} \frac{C_t^h (h-1)!}{h(h-d-1)!}} \\ &= \frac{\sum_{m\geq 0} \frac{C_t^{m+1}(m+d+1)!}{(m+d+2)m!}}{\sum_{m\geq 0} \frac{C_t^{m}(m+d)!}{(m+d+1)m!}}.\end{aligned} \qquad (15)$$

We now obtain upper and lower bounds of eq. (15) for any $d = 1, 2, \ldots$. To derive a lower bound note that

$$\frac{m+d+2}{(1+1/d)(m+d+1)} < 1$$

and that $(m+d)/(m+d+1) < 1$, for $m = 0, 1, \ldots$. Using the above we decrease the numerator and increase the denominator of (15) obtaining the lower bound:

$$\Gamma_d > \frac{\sum_{m\geq 0} \frac{C_t^{m+1}(m+d+1)!}{(1+1/d)(m+d+1)m!}}{\sum_{m\geq 0} \frac{C_t^{m}(m+d)!}{(m+d)m!}} = \frac{C_t d}{(1+1/d)(1-C_t)}.$$

Similarly, an upper bound of (15) can be obtained from applying the inequalities

$$\frac{m+d+1}{(1+1/d)(m+d)} \leq 1$$

and $(m+d+1)/(m+d+2) < 1$, for $m = 0, 1, \ldots$ into (15)

$$\Gamma_d < \frac{\sum_{m\geq 0} \frac{C_t^{m+1}(m+d+1)!}{(m+d+1)m!}}{\sum_{m\geq 0} \frac{C_t^{m}(m+d)!}{(1+1/d)(m+d)m!}} = \frac{(d+1)C_t}{(1-C_t)}.$$

Thus,

$$\frac{d}{1+1/d} < \Gamma_d \left(\frac{1-C_t}{C_t}\right) < d+1$$

which implies that for $\langle k-(d+a)|d+a \rangle > \langle k-d \rangle_{\zeta^{(d)}}$, for all $a \geq 2$ and $d = 1, \ldots$. This means that at step $t+1$ we should recruit the node $v \in \mathcal{N}(\mathcal{B}(t))$ with either the largest value $d(v)$ of all nodes in $\mathcal{N}(\mathcal{B}(t))$ or the second largest value. We hypothesize that improving the above bounds will reveal that node v should have the largest value of $d(v)$. Thus, if the above hypothesis holds, recruiting the node with the largest observed degree from $\mathcal{N}(\mathcal{B}(t))$ maximizes the expected cover increase. An interesting observation is that $\langle k-d \rangle_{\zeta^{(d)}}$ increases with C_t and diverges as $C_t \to 1$. We now explore the case $C_t \to 1$.

Case $C_t = 1$: The case when $C_t = 1$ represents the case when $\{p_k\}$ is a pure power law distribution – that is, $\zeta_k = k^{-\tau}/\zeta(\tau)$, where $\zeta(\tau)$ is the Riemann zeta function with parameter τ (note that $\mathrm{Li}_\tau(1) = \zeta(\tau)$) – and t has little impact on the degree distribution of $\overline{\mathcal{B}}(t)$. This case requires assuming $N \to \infty$ and $t = o(N)$ and thus it is just of theoretical interest as no real world network is an infinite power law network. Because $\zeta(a) \to \infty$ for $a \leq 1$ and $\zeta(a) < \infty$ for $a > 1$, we observe that

$$\frac{\partial^{\lceil \tau-1 \rceil} H(z)}{\partial z^{\lceil \tau-1 \rceil}} < \infty$$

and

$$\frac{\partial^{\lceil \tau \rceil} H(z)}{\partial z^{\lceil \tau \rceil}} \to \infty\,.$$

Thus, $\langle k-d|d \rangle$ diverges for $d = \lceil \tau \rceil$ and converges for $d < \lceil \tau \rceil$. This implies that the expected d-excess degree of a node with $d = \lceil \tau \rceil$ recruited neighbors is infinite. Due to the appearance of subtractions between two infinite quantities we were unable to verify whether $\langle k-d|d \rangle \to \infty$ for all $d > \lceil \tau \rceil$ or, more importantly, whether or not $\langle k-d-1|d+1 \rangle / \langle k-d|d \rangle > 1$ holds for all $d > \lceil \tau \rceil$. We, however, hypothesize that $\langle k-d|d \rangle \to \infty$ for all $d \geq \lceil \tau \rceil$. And again we see that recruiting the node with the largest observed degree from $\mathcal{N}(\mathcal{B}(t))$ maximizes the expected cover increase.

References

[1] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 64(4), 2001.

[2] I. Adler and S.M. Ross. The coupon subset collection problem. *Journal of Applied Probability*, 38(3):737–746, 2001.

[3] L. Amaral, A. Scala, M. Barthelemy, and H. Stanley. Classes of small-world networks. *Proc. Natl. Acad. Sci.*, (97):11149–11152, 2000.

[4] K. Avrachenkov, B. Ribeiro, and D. Towsley. Improving random walk estimation accuracy with uniform restarts. In *Proc. of the 7th Workshop on Algorithms and Models for the Web-Graph (WAW), also Springer LNCS*, volume 6516, pages 98–109, 2010.

[5] Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge University Press, 2008.

[6] S. Bhattacharya, D. Korzhyk, and V. Conitzer. Computing a profit-maximizing sequence of offers to agents in a social network. In *Proc. of the 8th Intl. Workshop on Internet and Network Economics (WINE)*, 2012.

[7] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489:295–298, Sept. 2012.

[8] Colin Cooper and Alan Frieze. Random walks with look-ahead in scale-free random graphs. *SIAM J. Discret. Math.*, 24(3):1162–1176, September 2010.

[9] Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proc. of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pages 57–66, New York, NY, USA, 2001. ACM.

[10] Nikolaos Fountoulakis, Konstantinos Panagiotou, and Thomas Sauerwald. Ultra-fast rumor spreading in social networks. In *Proc. of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pages 1642–1660. SIAM, 2012.

[11] O. Frank and T. Snijders. Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10:53–67, 1994.

[12] Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness.* W. H. Freeman & Co., New York, NY, USA, 1990.

[13] Krista Gile and Mark S. Handcock. Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology*, 40(285–327), 2010.

[14] M. Gjoka, M. Kurant, C.T. Butts, and A. Markopoulou. Practical recommendations on crawling online social networks. *Selected Areas in Communications, IEEE Journal on*, 29(9):1872 –1892, october 2011.

[15] Sharad Goel and Matthew J. Salganik. Assessing respondent-driven sampling. *Proc. of the National Academy of Sciences*, 107(6743–6747), 2010.

[16] S. Guha and S. Khuller. Approximation algorithms for connected dominating sets. *Algorithmica*, 20(4):374–387, April 1998.

[17] Jason Hartline, Vahab Mirrokni, and Mukund Sundararajan. Optimal marketing strategies over social networks. In *Proc. of the 17th international conference on World Wide Web*, WWW '08, pages 189–198, New York, NY, USA, 2008. ACM.

[18] Douglas D. Heckathorn. Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, 44(174–199), 1997.

[19] CNN.com. *How Obama's data crunchers helped him win.* `http://liveweb.archive.org/http://www.cnn.com/2012/11/07/tech/web/obama-campaign-tech-team/index.html`.

[20] S. Ioannidis and P. Marbach. Absence of evidence as evidence of absence: A simple mechanism for scalable p2p search. In *INFOCOM 2009, IEEE*, pages 576 –584. april 2009.

[21] Viggo Kann. *On the Approximability of NP-complete Optimization Problems.* PhD thesis, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, 1992.

[22] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '03, pages 137–146, New York, NY, USA, 2003. ACM.

[23] J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In N. Nisan, T. Roughgarden, E. Tardos, and V. Vazirani, editors, *Algorithmic Game Theory.* Cambridge University Press.

[24] J. Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings oh the International Congress of Mathematicians: Madrid, August 22-30, 2006: invited lectures*, pages 1019–1044, 2006.

[25] B. Klimmt and Y. Yang. Introducing the enron corpus. In *Proc. of the 1st Conference on Email and Anti-Spam (CEAS)*, 2004.

[26] M. Kurant, A. Markopoulou, and P. Thiran. Towards Unbiased BFS Sampling. *Selected Areas in Communications, IEEE Journal on*, 29(9):1799 –1809, october 2011.

[27] Maciej Kurant, Minas Gjoka, Carter T. Butts, and Athina Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *Proc. of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, SIGMETRICS '11, pages 281–292, New York, NY, USA, 2011. ACM.

[28] Q. Lv, P. Cao, E. Cohen, K. Li, and S. Shenker. Search and replication in unstructured peer-to-peer networks. In *Proc. of the 16th international conference on Supercomputing*, pages 84–95. ACM, 2002.

[29] Arun S. Maiya and Tanya Y. Berger-Wolf. Benefits of bias: towards better characterization of network sampling. In *Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 105–113, New York, NY, USA, 2011. ACM.

[30] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(415–444), 2001.

[31] M. Mihail, A. Saberi, and P. Tetali. Random walks with lookahead on power law random graphs. *Internet Mathematics*, 3(2):147–152, 2006.

[32] V. Mikhailov. A poisson limit theorem in the scheme of group disposal of particles. *Theory of Probability and its Applications*, 22:152–156, 1977.

[33] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, IMC '07, pages 29–42, New York, NY, USA, 2007. ACM.

[34] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B*, 26:521–529, 2002.

[35] Marc Najork and Janet L. Wiener. Breadth-first crawling yields high-quality pages. In *Proc. of the 10th international conference on World Wide Web*, WWW '01, pages 114–118, New York, NY, USA, 2001. ACM.

[36] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14(1):265–294, 1978.

[37] Jure Leskovec. SNAP network data repository. `http://snap.stanford.edu/data/`.

[38] M. E. J. Newman. The structure of scientific collaboration networks. *Proc. Natl. Acad. Sci.*, 98:404–409, 2001.

[39] M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45:167–256, 2003.

[40] M. E. J. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[41] R Pastor-Satorras and A Vespignani. Epidemic dynamics in finite size scale-free networks. *Physical Review E*, 65(3), 2002.

[42] Bruno Ribeiro and Don Towsley. Estimating and sampling graphs with multidimensional random walks. In *Proc. of the 10th ACM SIGCOMM conference on Internet measurement*, IMC '10, pages 390–403, New York, NY, USA, 2010. ACM.

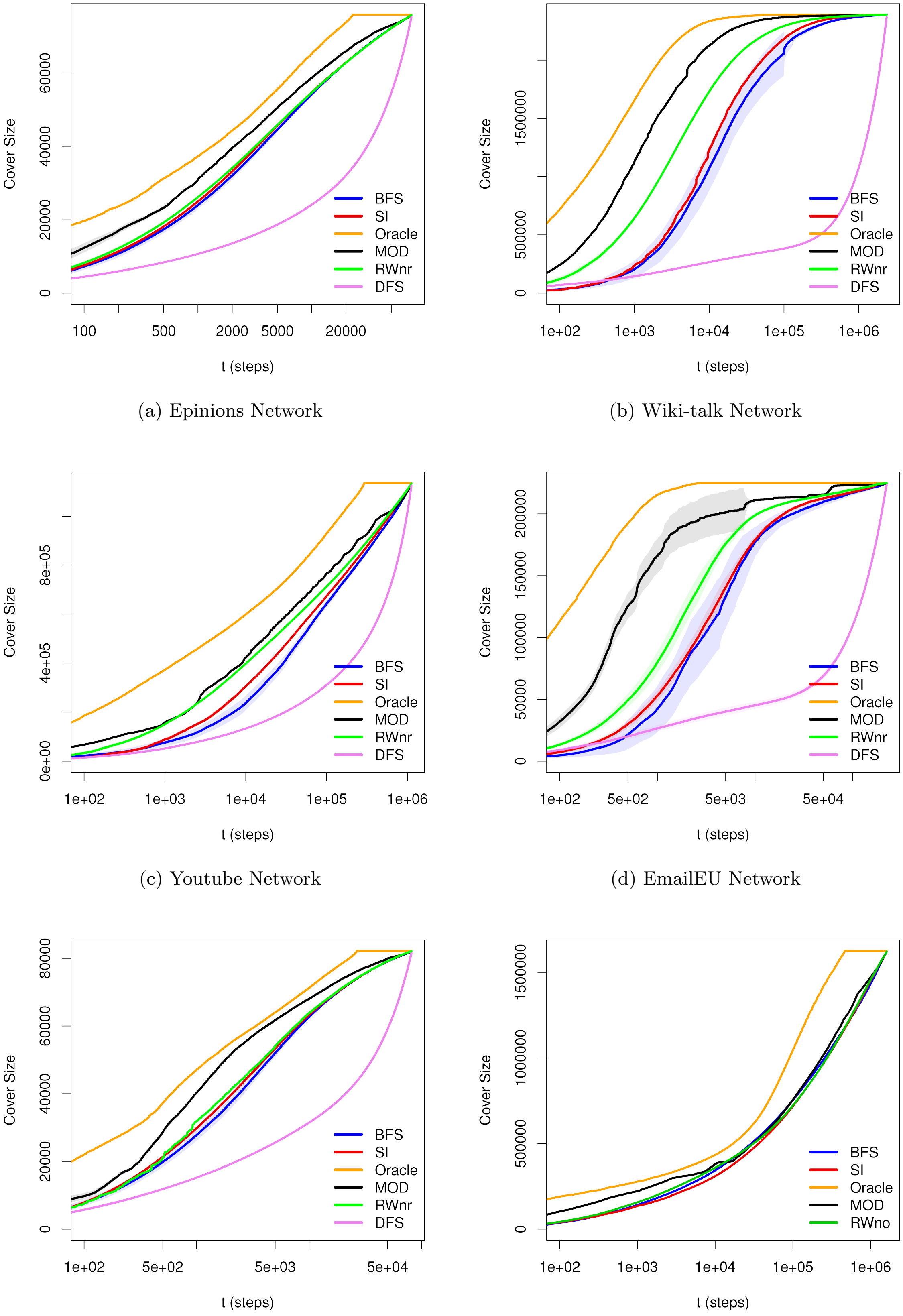
[43] Matei Ripeanu, Adriana Iamnitchi, and Ian Foster. Mapping the Gnutella Network. *IEEE Internet Computing*, 6(1):50–57, January 2002.

[44] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger. On unbiased sampling for unstructured peer-to-peer networks. *IEEE/ACM Trans. Netw.*, 17(2):377–390, 2009.

[45] Yeon sup Lim, D. S. Menasche, B. Ribeiro, D. Towsley, and P. Basu. Online estimating the k central nodes of a network. In *Proc. of the 2011 IEEE Network Science Workshop*, NSW '11, pages 118–122, Washington, DC, USA, 2011. IEEE Computer Society.

[46] D. Tsoumakos and N. Roussopoulos. Analysis and comparison of p2p search methods. In *Proc. of the 1st international conference on Scalable information systems*, page 25. ACM, 2006.

[47] G. Watson. Three triple integrals. *Quart. J. Math.*, 2(10):266–276, 1939.



(a) Epinions Network

(b) Wiki-talk Network

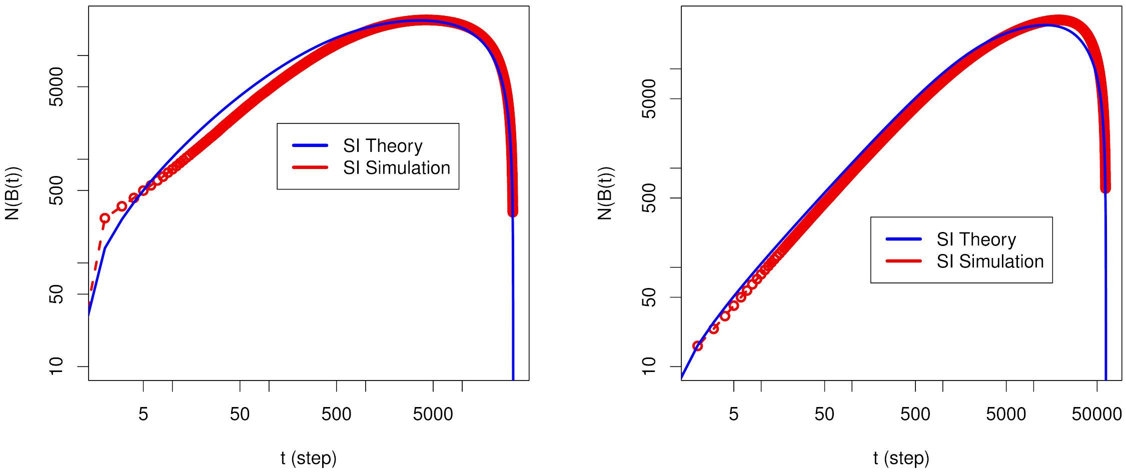
(c) Youtube Network

(d) EmailEU Network

(e) Slashdot Network

(f) Flickr Network

Figure 5: Empirical average cover size $\langle \overline{W}(t) \rangle$ of various social networks. Comparison between Oracle, RWnr, SI, BFS, DFS, and MOD algorithms. x-axis in log-scale.



(a) Enron Network

(b) Gnutella Network

Figure 6: Mean field approximation of $\langle \mathsf{N}(B(t)) \rangle$ eq. (4) against the true average of 1,000 simulations on Enron and Gnutella networks.

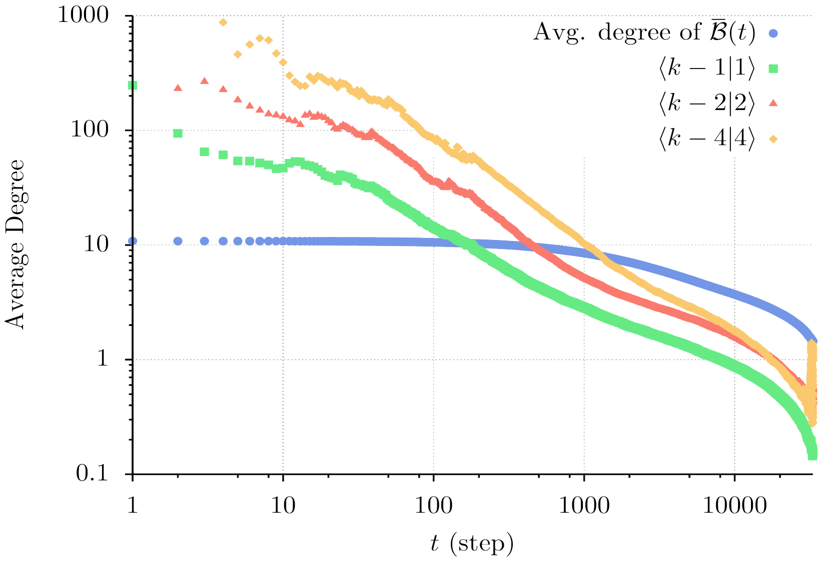
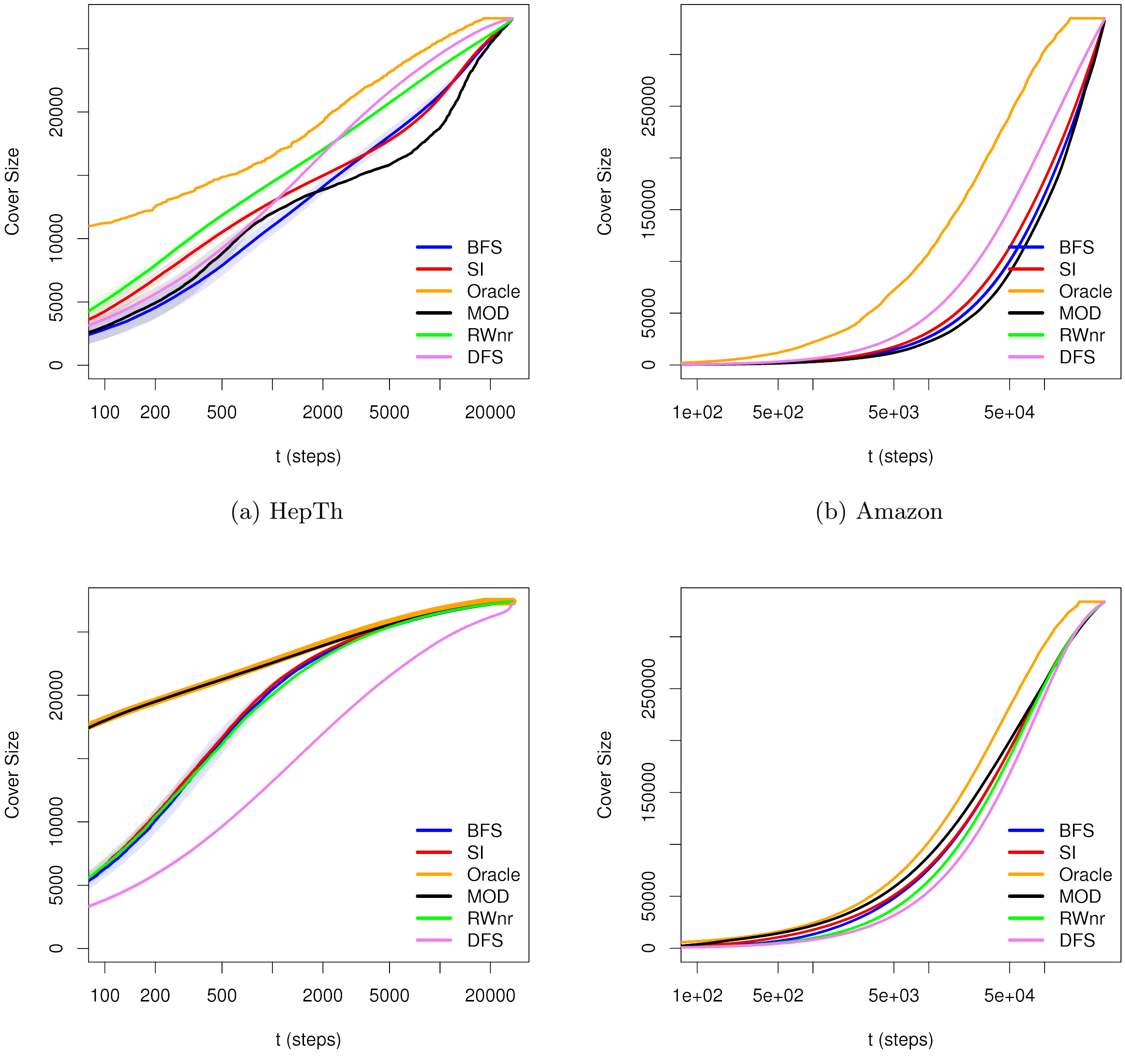


Figure 7: **(Enron Network)** Average d-excess degree of $\overline{\mathcal{B}}(t)$ in SI as a function of step $t = 1, \ldots, N-1$. Observe that $\langle k-4|4 \rangle$ is consistently larger for all values of t than $\langle k-1|1 \rangle$ and $\langle k-2|2 \rangle$. Also note that this happens notwithstanding how little the average degree of $\overline{\mathcal{B}}(t)$ varies over t (blue dots).



(a) HepTh

(b) Amazon

(c) Randomized HepTh

(d) Randomized Amazon

Figure 8: The two datasets in which DFS outperforms MOD. Comparison between the empirical average cover size $\langle \overline{W}(t) \rangle$ of Oracle, RWnr, SI, BFS, DFS, and MOD algorithms. Figs. 8a and 8b show the results on the original networks and Figs. 8c and 8d show the results in their randomized counterparts. Note that when randomized, we see similar results seen in the social networks. Thus, the good performance of DFS and poor performance of MOD are due to the peculiar network topology of these graphs. x-axis in log-scale.