The Power of Random Neighbors in Social Networks

Silvio Lattanzi
Google, Inc.
New York, NY 10011
silviol@google.com

Yaron Singer
Harvard University
Cambridge, MA 02138
yaron@seas.harvard.edu

ABSTRACT

The friendship paradox is a sociological phenomenon first discovered by Feld [13] which states that individuals are likely to have fewer friends than their friends do, on average. This phenomenon has become common knowledge, has several interesting applications, and has also been observed in various data sets. In his seminal paper Feld provides an intuitive explanation by showing that in *any* graph the average degree of edges in the graph is an upper bound on the average degree of nodes. Despite the appeal of this argument, it does not prove the existence of the friendship paradox. In fact, it is easy to construct networks – *even with power law degree distributions* – where the ratio between the average degree of neighbors and the average degree of nodes is high, but all nodes have the *exact same degree* as their neighbors. Which models, then, explain the friendship paradox?

In this paper we give a strong characterization that provides a formal understanding of the friendship paradox. We show that for *any* power law graph with exponential parameter in $(1, 3)$, when every edge is rewired with constant probability, the friendship paradox holds, i.e. there is an asymptotic gap between the average degree of the sample of polylogarithmic size and the average degree of a random set of its neighbors of equal size. To examine this characterization on real data, we performed several experiments on social network data sets that complement our theoretical analysis. We also discuss the applications of our result to influence maximization.

1. INTRODUCTION

In popular culture, the *friendship paradox* is known as the somewhat discouraging statement that our friends have more friends than we do. This statement is an interpretation of a result discovered by Feld [13] which states that individuals are likely to have fewer friends than the mean number of friends their friends have. Superficially, since the number of friends a person has often translates to social status, the friendship paradox receives considerable attention. Beyond this interpretation, the idea that the degree of a random neighbor may be substantially larger than that of a random node has interesting applications for immunization [8, 23] and influence maximization [36], and it has also been verified experimentally in large online social networks [17, 38].

Despite everything we know about it, from a mathematical perspective the friendship paradox is, at large, a mystery. In his seminal paper, Feld provides an intuitive explanation for the friendship paradox by showing that in *any* graph the average degree of edges in the graph is an upper bound on the average degree of nodes. This however, is far from proving the friendship paradox – that a node is likely to have a smaller degree than the average degree of her friends. In fact, as we will later show, it is easy to construct networks – *even with power law degree distributions* – where the ratio between the average degree of neighbors and the average degree of nodes is high, but all nodes have the *exact same degree* as their neighbors. In other words, not all graphs exhibit the friendship paradox, degree distribution does not explain it, and other structural properties need to be considered. Given the evidence in social network data sets, the question is why the friendship paradox exists.

Which models explain the friendship paradox?

The inception of network science is largely due to seminal mathematical models that explain phenomenon such as small-worlds [21, 40] and power-law degree distributions [4]. In these models we have a mathematical definition for the phenomenon exhibited: short distances in the network are defined as polylogarithmic in its size, heavy tail degree distribution as a power-law. But how should we mathematically define existence of the friendship paradox in a network? A possible definition of the friendship paradox is the existence of an asymptotic gap between a random node and its random neighbor. In a regular network, where all nodes have the same degree, there is no difference between the degree of any node and that of its neighbors, so the friendship paradox does not holds. In a star network on the other hand, where a randomly selected node is likely to be an endpoint with its only neighbor being the center, the ratio between a node and its neighbor is linear in the size of the network, and the friendship paradox holds. So what lies between these extremes that can serve as good model for social networks?

1.1 The structure of social networks

Perhaps the most common assumption about the structure of social networks and complex systems in general, is

Figure 1: An illustration of two power law networks. The network (B) is a set of disjoint cliques generated using the network (A). Despite being power law, every node in (B) has exactly the same degree as its neighbor.

that their degree distribution is heavy tailed. In this paper we focus on the most commonly cited heavy tail distribution, the power law distribution: the likelihood of observing a node with degree i in the network is proportional to $i^{-\beta}$, where β is a constant that depends on the network. [1]

In Figure 1 we illustrate an example showing that the power law property does not suffice to show that a node is likely to have less their friends than her friend. Essentially, one can take any power law network (illustrated as network A) and turn it into a power law network where each node of degree d in the original network is represented as an isolated clique of size $d+1$. A series of such graphs has a power law degree distribution. In Section 3 we show such a construction which formally holds for networks with finite degree distributions. These examples are obviously contrived, but prove an important point: power law degree distributions alone do contain enough information about the network to study the likelihood of a node to have a high degree neighbor. If we wish to *prove* properties on social networks we will need to consider some form of a model. An important property to consider is the presence of long-range edges. Long-range links were first observed by Granovetter [16] and are often modeled as random edges [21, 40].

A general model for social networks. Following the above discussion and inspired by the work of Watts and Strogatz [40], we analyze the following family of graphs: we start from *any* (potentially adversarial) power law network and then we "re-wire" each edge in the network with some fixed probability. The implication is that one can take *any model of a social network* as long as its degree distribution follows a power law, allow every edge to have some fixed probability to be connected at random, and our results will hold. So even the most contrived examples of power law networks can be slightly perturbed and succumb to our analysis.

Additional models. It is important to emphasize that the results we show here also hold for other well-studied models of social networks. In particular, our results also hold for the *configuration model* [5] with power law degree distributions for finite graphs [1](this is a special case of our model when the likelihood of an edge being rewired is 1) and preferential attachment model [4], both widely used in social network analysis (see e.g. [1, 35, 33]), as well as a similar model for generating connected graphs [39]. We omit the proofs for these models here due to lack of space, though the techniques shown here generalize to these models as well.

[1]There is an interesting debate on degree distributions of social network graphs [31]. Different papers support different heavy-tail distributions: power law, log normal, double Pareto, etc. For simplicity we focus on power law distributions.

1.2 Our results

Informally, our main result is that small samples suffice to observe the friendship paradox in social networks. More formally, we show that for any power law graph with parameter $\beta \in (1,3)$, when every edge is rewired with constant probability, there is an asymptotic gap between the average degree of the sample of polylogarithmic size and the average degree of a random set of its neighbors of equal size.

To show the importance of random edges, we first show a construction of a finite power law network where the friendship paradox does not hold – every node has almost the same degree as its neighbor. We then analyze the case where only a single node is sampled from the network. We show that the question of whether an asymptotic gap exists between a set of randomly drawn nodes and their neighbors depends on the parameters of the graph. Our results show a threshold phenomenon: for graphs whose degree distribution is a power law with $\beta \in (1,2]$ an asymptotic gap exists, while for $\beta \in (2,3)$ it does not. Using this analysis we develop our main result for polylogarithmic samples.

To examine how these results relate to real data, we performed several experiments on social network data sets. Similar to our analytical results, we observed large gaps between random node samples and their neighbors. We examined other effects as well, as we further describe in Section 6.

1.3 Algorithmic implications

Before we continue, it would be useful to briefly review the algorithmic implications of the friendship paradox in the context of influence maximization, which led to our original interest in this phenomenon. The influence maximization problem [11, 20] is the algorithmic challenge of selecting individuals who can serve as early adopters of a new product or technology in a manner that will trigger a large cascade in the social network. The massive adoption of online social networking technologies in recent years has drawn substantial interest to the study of information diffusion in social networks and influence maximization in particular [3, 6, 7, 15, 24, 27, 29, 32]. Despite the substantial progress on the problem throughout the past decade, naive application of state-of-the-art algorithms is often ineffective. In many applications, one only has access to a small *sample* of the network in which individuals have low influence potential, and algorithms that select their users from such samples have a limited effect. In marketing applications for example, merchants often reward influential users who visit their online store, or who have engaged in other ways (subscribe to a mailing list, follow the brand, install an application etc.). Thinking of users who arrive at a store or follow a brand as being randomly sampled from the network, it follows that observing high-degree users is a rare event simply because the degree distributions of social networks are heavy-tailed. Influence maximization techniques are based on selecting high degree users (not necessarily the *highest* degree), and their application on such samples is therefore ineffective.

Two-stage approaches. As an alternative to spending the entire budget on nodes in the sample, recent work advocates for a two-stage approach called *Adaptive Seeding* [36]: In the first stage, a fraction of the budget is spent on nodes in the sample for the purpose of attracting their neighbors to

Figure 2: An illustration of the friendship paradox in different networks. The two networks above are of the same size in terms of nodes and edges but with different topologies. The graph on the left has edges connected between nodes uniformly at random, and the graph on the right is a random graph with degree distribution close to a power law, i.e. the probability of observing a node of degree i is proportional to $i^{-\beta}$ for some constant β. In the first network the friendship paradox does not hold while in the second one it does. In each network we performed the following experiment: we selected five random nodes to represent the sample S, which are depicted in yellow, their neighbors $\mathcal{N}(S)$ are depicted in orange and their neighbors which is the potential influence of $\mathcal{N}(S)$ are depicted in red. The rest of the nodes are represented in pink.

join the set of potential influencers (e.g. attract neighbors to visit the website, follow the brand, register their email, etc.). In the second stage, after some neighbors have joined, the remainder of the budget is used to select an influential set of individuals from the (hopefully larger) set of accessible nodes. The rationale is that while samples of a graph will likely have low degree nodes, they may have high degree *neighbors*. Intuitively, since high degree nodes have many neighbors (by definition), one would hope such nodes will be connected to the sample. In simple terms, we are guided by the following question:

Are random nodes likely to have high degree neighbors in a social network?

One central contribution in this paper is to explore this question at depth and provide analytical and empirical evidence to show that this is the case. Showing that the friendship paradox exists in social networks implies that adaptive seeding algorithms can indeed enable dramatic improvements for information dissemination. We illustrate how the friendship paradox translates to potential influence of two-stage approaches in Figure 2.

1.4 Related Work

Our results are directly related to Feld's work [13, 14]. Feld gives an explicit characterization of the average degree of a neighbor in terms of the expected degree and variance of the degree distribution in the network and observes that the average degree of neighbors will be strictly greater than the average degree of nodes when the variance is greater than 0. In addition, Feld provides experimental evidence strengthening his thesis. Based on these intuitions, several heuristics for detecting contagions and designing immunization strategies have been suggested [8, 9]. In particular, the knowledge gap we found on this topic can be summarized as follows:

- **Size matters.** To identify asymptotic ratios, we must quantify how many more neighbors a neighbor has.
- **Averages do not imply likelihood.** Perhaps the most critical distinction is that Feld's result is a statement about averages, and **does not imply that a node is likely to have a lower degree than her neighbor**. For the purpose of designing algorithms for example, Feld's result is inapplicable.
- **The friendship paradox is not an amplification of Feld's statement.** It is important to emphasize that the friendship paradox is not a constant probability version of Feld's result. The constant probability version of Feld's result is the ratio between the degree of a randomly selected node v and the degree of a *random neighbor y in the graph*. That is, y is not a neighbor of x. To see that these are two different random variables consider the star of N nodes: the expected degree of friends of a random node is $N - 1/N$, while a constant probability version of Feld gives $N/2$.

Recently, large scale experiments showed that individuals are indeed likely to have less friends than the average number of friends their friends have in the Facebook [38] and Twitter networks [17]. Our work is somewhat complementary to these works; we explain a formal model and provide rigorous analysis of the phenomena observed in their experiments.

From a technical perspective, a related research direction recently explored is social sampling. In this context the goal is usually to sample a subgraph and maintain properties of the original graph [25], compute some statistics in sublinear time [10, 18, 19], or find a subset of the nodes with a certain property [2]. Despite some similarities, our work is different as we focus on finding high degree nodes in a network using an existing sample and we only consider the the node's immediate neighbors.

2. PRELIMINARIES

Power law graphs. We use the same definition of a power law network that is used in [1]. A graph has a power law distribution if the number of nodes of degree i is $\lfloor \frac{e^\alpha}{i^\beta} \rfloor$. Since the sum of the degrees in a network has to be even, the number of nodes of degree 1 should either be $\lfloor e^\alpha \rfloor$ or $\lfloor e^\alpha + 1 \rfloor$ depending on the parity of the sum of the higher degree nodes in the graph. For simplicity, throughout the rest of the paper we assume that the the number of nodes of degree 1 is e^α, and note that all the results naturally extend to the

general definition. For brevity we set $C = e^{\alpha}$ and use N and M to denote the number of nodes and edges in the graph, respectively. We denote the average degree of a set B as $d(B)$.

Social network model. We will analyze the following random graph model, which is inspired by the small-world model of Watts and Strogatz [40]. We start from an arbitrary (potentially adversarial) graph with a power law distribution and then "rewire" each edge in the graph with constant probability $p > 0$. Rewiring is done by first selecting all the edges that have to be rewired and then by applying the technique used in the configuration model [5]: every edge selected to be rewired is split into two stubs attached to the nodes corresponding with the endpoints of that edge, and the set of all stubs are then connected uniformly at random. The importance of this model is that it shows that even when starting from an adversarial structure (**potentially with a strong community structure or even a disconnected network**) a small fraction of randomness suffices to observe an asymptotic gap between the degree of a sampled set of nodes and their neighbors. In case of self-loops, we count a node as a neighbor of itself (in most of the setting the number of self loops is extremely small).

3. MISBEHAVED POWER LAWS

In this section we show that there exists a family of power law graphs such that the ratio between the degree of a node and the degree of its neighbor is constant for every node in the graph.[2]

PROPOSITION 3.1. *There exist a family of power law graphs with $\beta = 2$ such the ratio between the degree of any node and the degree of any of its neighbor is at most a constant.*

PROOF. First we order the nodes in an increasing degree order. We begin by generating the stubs, initializing each node with unassigned stubs. For all nodes of degree smaller than $\frac{\sqrt{N}}{400}$, starting from the first node we sequentially match each unassigned stub of a node v with an unassigned stub of the minimum ranked node with an available unassigned stub which is not already connected to v.

An interesting property of this ordering is that for every node v of degree smaller than $\frac{\sqrt{N}}{400}$ it is always possible to connect it with the subsequent $d(v)$ nodes. This is true because we consider nodes in an increasing degree order. Note that in this way every node of degree smaller $\frac{\sqrt{N}}{400}$ is connected with nodes that have a degree that is at least half of its degree, and at most double its degree. In fact, every node v connects with at most $d(v)$ subsequent nodes and $\frac{d(v)}{2}$ preceding nodes.

Now we have to assign the edges between all the nodes of degree larger than $\frac{\sqrt{N}}{400}$. But this can be done by assigning edges arbitrarily (for example using the configuration model on the remaining stubs) since the maximum degree node has degree $\sqrt{N}$ and so the ratio between the degree of any two nodes with degree larger than $\frac{\sqrt{N}}{400}$ is at most constant. □

[2]We note that for simplicity we prove a result for $\beta = 2$, although similar counter-examples can be constructed for other values of β. We also note that it is easy to modify our construction to obtain a connected graph.

4. SINGLE SAMPLES

As a preliminary to the main results in Section 5, we investigate the case in which the sample is of a constant size. As we will now show, when we consider small samples the question of whether neighbors yield better results largely depends on the parameters of the graph. While an interesting result in of itself, the lemmas for establishing this will also be instrumental for proving the main result in Section 5.

Throughout the rest of this paper we use $G(\beta, p)$ to denote the family of graphs with power law degree distributions with exponent β in which every edge has some small, constant probability $p > 0$ to be rewired at random.

LEMMA 4.1. *Let u be a node sampled from $G(\beta, p)$ with $\beta \in (1, 2)$ and let v be a neighbor of u drawn u.a.r.. Then, for any constant $\epsilon > 0$, with constant probability the ratio between the degree of u and v is $\Omega\left(N^{\frac{\beta-1}{\beta}-\epsilon}\right)$.*

PROOF. We first compute the average degree of a node in $G(\beta, p)$. Using the fact that the generalized harmonic sum converges to a constant when $\beta > 1$, i.e. $\sum_{i=1}^{\infty} \frac{1}{i^{\beta}} \in O(1)$, we can conclude that the number of nodes in a power law graph with $\beta \in (1, 3)$ is:

$$N = \sum_{i=1}^{\Delta} \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \sum_{i=1}^{\lceil C^{1/\beta} \rceil} \left\lfloor \frac{C}{i^{\beta}} \right\rfloor \in \Theta(C)$$

where Δ is used to denote the largest degree in the graph. The number of edges is:

$$M = \sum_{i=1}^{\lceil C^{1/\beta} \rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = O\left(\sum_{i=1}^{\lceil C^{1/\beta} \rceil} \left\lfloor \frac{C}{i^{\beta-1}} \right\rfloor \right) \in \Theta\left(C^{\frac{2}{\beta}}\right)$$

where we use the fact that that the generalized harmonic sum for $0 < \alpha < 1$ is equal to $\sum_{i=1}^{t} i^{-\alpha} \in \Theta\left(t^{1-\alpha}\right)$. The average degree of a node in $G(\beta, p)$ is therefore in $\Theta\left(C^{\frac{2}{\beta}-1}\right)$. So, by Markov's inequality, we know that when we select a random node with probability $\frac{1}{2}$ its degree is at most twice its expected degree. Thus, with probability $\frac{1}{2}$ we sample a node of degree $O\left(C^{\frac{2}{\beta}-1}\right)$.

To lower bound the degree of a random neighbor, we show that with constant probability an edge is incident to a high degree node. To show this we calculate the number of endpoints of edges incident to nodes of degree at least K, denoted $A_{d>K}$:

$$A_{d>K} = \sum_{i=\lceil K \rceil}^{\lceil C^{1/\beta} \rceil} i \left\lfloor \frac{C}{i^{\beta}} \right\rfloor = \Theta\left(C\left(C^{\frac{2}{\beta}-1} - K^{2-\beta}\right)\right)$$

since that for $\alpha < 1$, $\sum_{i=k}^{n} \frac{1}{i^{\alpha}} = \Theta(n^{1-\alpha} - k^{1-\alpha})$.

Therefore the number of endpoints of edges incident to nodes with degree greater or equal to $K = C^{1/\beta-\epsilon}$ is $\Theta(M)$. If we restrict our attention to endpoints of *rewired* edges incident to nodes with degree greater or equal to $K = C^{1/\beta-\epsilon}$, denoted by $R_{d>K}$, then by linearity of expectation we have that $E[R_{d>K}] = pA_{d>K} \in \Theta(M)$. Unfortunately the probability that the endpoints are rewired is not independent. In fact, two endpoints incident to the same edge are either both

rewired or both not. So to obtain a concentration result we cannot apply the Chernoff bound. Fortunately in this case we can use the method of bounded difference [30] which tells us that the fact that an edge is rewired changes the value of $R_{d>K}$ by an additive factor of 2 at most, and we have that $R_{d>K} \in \Theta(M)$.

So with constant probability a random neighbor of a random node would be connected to the node by a *rewired* edges and this edge with constant probability will point to a node in $R_{d>K} \in \Theta(M)$.

Thus, for $\beta \in (1,2)$ with constant probability the ratio between a node a random neighbor is at least:

$$K\left(\frac{M}{N}\right)^{-1} = \Omega\left(C^{\frac{\beta-1}{\beta}-\epsilon}\right) = \Omega\left(N^{\frac{\beta-1}{\beta}-\epsilon}\right). \quad \square$$

The case when $\beta = 2$: we can apply the same technique as above for $\beta = 2$. In this case as in the above proof the number of nodes is $N \in \Theta(C)$ and the number of edges is:

$$M = \sum_{i=1}^{\lceil C^{1/\beta}\rceil} i\left\lfloor\frac{C}{i^2}\right\rfloor = O\left(\sum_{i=1}^{\lceil C^{1/\beta}\rceil}\left\lfloor\frac{C}{i}\right\rfloor\right) \in \Theta\left(C\log C\right)$$

where this time we use the fact that that the harmonic sum is equal to $\sum_{i=1}^{t} i^{-1} \in \Theta\left(\log t\right)$. When $\beta = 2$ the number of endpoints of edges incident to nodes of degree at least K is:

$$A_{d>K} = \sum_{i=\lceil K\rceil}^{\lceil C^{1/\beta}\rceil} i\left\lfloor\frac{C}{i^\beta}\right\rfloor = \Theta\left(C\left(\log C - \log K\right)\right)$$

where we use the fact that $\sum_{i=k}^{n}\frac{1}{i} = \Theta(\log n - \log k)$. Using the same technique one can show that with constant probability the ratio between the degree of u and v is $\Omega\left(\log^\alpha N\right)$, for any constant $\alpha > 0$.

COROLLARY 4.2. *Let u be a randomly selected node from $G(\beta,p)$ with $1 < \beta \leq 2$, and let v a randomly selected neighbor of u. Then with constant probability we have:*

$$\frac{d(v)}{d(u)} \in \omega(1).$$

4.1 Phase transition for $\beta > 2$

Somewhat surprisingly, we now show that when $\beta > 2$ the friendship paradox does not hold for a single node – and not even for a sample of a constant size, and even when all edges are rewired at random. In other words, when $\beta > 2$ a constant number of nodes do not suffice to observe asymptotic gaps between degrees of the sample and their neighbors, even in a random graph model.

LEMMA 4.3. *Let u be a node sampled from $G(\beta,1)$ with $\beta \in (2,3)$, and let v be a random neighbor of u. Then, w.h.p. the ratio between the degree of u and that of v is $\Theta(1)$.*

PROOF. First, note that in this case we also have that the number of nodes is $N \in \Theta(C)$. The number of edges when $2 < \beta < 3$ is:

$$M = \sum_{i=1}^{\lceil C^{1/\beta}\rceil} i\left\lfloor\frac{C}{i^\beta}\right\rfloor = O\left(\sum_{i=1}^{\lceil C^{1/\beta}\rceil}\left\lfloor\frac{C}{i^{\beta-1}}\right\rfloor\right) = \Theta(C)$$

because in this setting $\beta - 1 > 1$. Thus the number of nodes of degree greater or equal to K in the graph is $O\left(\frac{C}{K}\right)$. So for any function $f(N) \in \omega(1)$, the probability of randomly picking a node of degree at least $f(N)$ is $o(1)$. Thus when we sample a node u.a.r, w.h.p. it will have a constant degree.

We will now show that when sampling a single node, w.h.p. also a randomly selected neighbor has degree smaller than $f(N)$, for any function $f(N)$ strictly increasing with N and such that $f(N) \in \omega(1)$ implying that w.h.p. the degree of a random neighbor is also constant. For $2 < \beta < 3$, the number of endpoints of nodes with degree larger than K is:

$$A_{d>K} = \sum_{i=\lceil K\rceil}^{\lceil C^{1/\beta}\rceil} i\left\lfloor\frac{C}{i^\beta}\right\rfloor = \Theta\left(C\left(K^{2-\beta} - C^{\frac{2}{\beta}-1}\right)\right)$$

Here for $\alpha > 1$, $\sum_{i=k}^{n}\frac{1}{i^\alpha} = \Theta(k^{1-\alpha} - n^{1-\alpha})$. The endpoints of edges incident to nodes of degree at least $\lceil f(N)\rceil$ is therefore $A_{d>\lceil f(N)\rceil} \in o(N)$. The number of edges of degree greater than $f(N)$ is therefore sublinear in the number of the nodes and so all but an $o(1)$-fraction of the nodes will have no incident edge of degree at least $f(N)$. $\square$

5. POLY-LOGARITHMIC SAMPLES

In this section we show our main result, namely that for any $\beta \in (1,3)$ polylogarithmic-sized samples suffice to see asymptotic ratios between the degree of a sample and its set of friends. Intuitively, this implies that as long as we have access to a relatively small size of the social network it is possible to reach nodes of relatively high-degree. The results here hold for any set of poly-logarithmic size, but we discuss sizes of $\log N$ for simplicity.

LEMMA 5.1. *Let S be a set of $\Theta(\log N)$ nodes sampled uniformly at random from $G(\beta,p)$, for $\beta \in (2,3)$. Then, w.h.p. the average degree in S is $\Theta(1)$.*

PROOF. From lemma 4.3, we know that the expected degree of a sampled node is $O(1)$. Thus, by linearity of expectation, the expected total degree of the sample is $O(\log N)$. By applying Markov's inequality we get that the probability that the total degree is in $\omega(\log N)$ is $o(1)$. $\square$

PROPOSITION 5.2. *Let S be a set of $\Theta(\log N)$ nodes sampled uniformly at random from $G(\beta,p)$ for $\beta \in (2,3)$, and let T_S be a set obtained by selecting a single neighbor u.a.r. from every node in S. Then, w.h.p.*

$$\frac{d(T_S)}{d(S)} \in \omega(1)$$

PROOF. The main idea behind is to show that w.h.p. T_S contains $\omega(1)$ nodes of degree $O(\log N)$ when we have a sample of size $\log N$. We will first lower-bound the probability of sampling one of these nodes and then show the result holds w.h.p. when we sample $\log N$ elements. From the proof of Lemma 4.3 we know that the number of endpoints of edges incident to nodes with degree at least $\lceil c\log N\rceil$, $A_{d>c\log N} \in \Theta\left(\frac{C}{\lceil c\log N\rceil^{\beta-2}}\right)$. Recall that $R_{d>c\log N}$ is the number of rewired edges incident to nodes of degree bigger than $\log N$. Also in this case by linearity of expectation we have $E[R_{d>c\log N}] = pA_{d>c\log N}$. Applying the method of bounded difference with lipschitz condition 2 we have that $R_{d>c\log N}$ is strongly concentrated around its mean. Thus,

the probability that a random edge of a node in S points to a neighbor with degree at least $c \log N$ is at least:

$$\frac{R_{d>c\log N}-1}{M} = \frac{\Theta\left(\frac{C}{\lceil c\log N\rceil^{\beta-2}}\right)}{\Theta(C)} = \Theta\left(\frac{1}{\lceil c\log N\rceil^{\beta-2}}\right).$$

Now, let Y_i be the random variable which equals 1 if the randomly selected neighbor of the node i in S has degree at least $c \log N$ and 0 otherwise. We have that:

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in S} Y_i\right] &= \log N \cdot \mathbb{E}[Y_i] = \log N\left(\frac{E_{d>c\log N}}{M}\right)\\ &= \Theta\left(\log N^{3-\beta}\right).\end{aligned}$$

Unfortunately the random variables Y_i are neither independent nor nicely correlated so we cannot use a Chernoff bound to get a high probability result directly. To overcome this difficulty, we show that the probability the sum of the Y_is is bigger than a value D dominates the probability that the sum of a random variable X_i that counts the number of heads of a coin that gives heads with probability $\frac{(E_{d>c\log N})-\log N}{M}$ and tails otherwise is bigger than a value D. Note that we are sampling only $\log N$ edges, thus for every i the number of edges of degree bigger than $c \log N$ that we still have not used is bigger or equal than $E_{d>c\log N} - \log N$ for all i. Thus the probability that Y_i is equal 1 is higher than the probability of the same event for X_i for all i, hence the sum of the random variables Y_i dominates the sum of the random variables X_i. But $\mathbb{E}[\sum_{i\in S} X_i] = \Theta\left(\log N^{3-\beta}\right)$. Furthermore the X_i are independent so by Chernoff we get that $\sum_{i\in S} X_i \in \Omega\left(\log N^{3-\beta}\right)$ with probability $1-o(1)$. Thus using stochastic domination, we have that w.h.p. at least $\Omega\left(\log N^{3-\beta}\right)$ sampled neighbors have degree bigger than $c \log N$.

Now in order to conclude the proof we have to show that those $\Omega\left(\log N^{3-\beta}\right)$ selected neighbors of high degree are different nodes. To this end we will show that any node in the graph is selected at most a constant number of times with high probability, which will imply the result.

The node of maximum degree is the most likely to be selected and it has degree $C^{\frac{1}{\beta}} = \Theta\left(C^{\frac{1}{\beta}}\right)$ so the probability of sampling it in one sample is $\Theta(C^{-1+\frac{1}{\beta}}) \in O(C^{-\frac{1}{2}})$. Thus the probability of sampling the highest degree node three times in $\log N$ samples is smaller than $\binom{\log N}{3}\Theta(C^{-\frac{3}{2}}) = o(C^{-1})$. Thus, using the union bound, no node is sampled more than twice with high probability. So the set of neighbors with high probability contains $\Omega\left(\log N^{3-\beta}\right)$ distinct nodes that have degree bigger or equal to $c \log N$, thus the average degree is w.h.p. at least $\Omega\left(\log N^{3-\beta}\right)$. □

The case when $\beta \in (1, 2]$: The proof requires a different approach when $1 < \beta \leq 2$. Roughly, the core idea is to amplify the results of the single sample case to show that w.h.p. we get at least one high degree node.

PROPOSITION 5.3. *Let S be a set of $\log N$ nodes sampled uniformly at random from $G(\beta)$ where $1 < \beta \leq 2$, and let $N(S)$ be a set obtained by selecting a single neighbor u.a.r. from every node in S. Then, w.h.p. the ratio between the sum of degrees of S and that of $N(S)$ is $\omega(1)$.*

Network	*# of Nodes*	*# of Edges*	β	C
Orkut	3,072,441	117,185,083	0.7470	223,989
LiveJournal	3,997,962	34,681,189	1.0322	520,041
Wikipedia	2,394,385	5,021,410	1.9548	80,033
YouTube	1,134,890	2,987,624	1.4212	160,927
DBLP	317,080	1,049,866	1.2048	64,983
SlashDot	82,168	948,464	1.2146	13,805
Enron	36,692	367,662	1.1636	11,322

Table 1: Networks' statistics.

In order to prove the proposition we first bound the average degree of the set of sample nodes:

LEMMA 5.4. *Let S be a set of $\Theta(\log N)$ nodes sampled uniformly at random from $G(\beta)$, for $1 < \beta < 2$. Then, w.h.p. the average degree in S is $\Theta\left(C^{\frac{2}{\beta}-1}\right)$.*

The proof uses the same Markov inequality argument as in Lemma 5.1 and is omitted. To conclude the main proof we show that with high probability we pick at least one very high degree node.

LEMMA 5.5. *Let S be a set of $\Theta(\log N)$ nodes sampled uniformly at random from $G(\beta, p)$ for $\beta \in (1, 2)$, and let T_S be a set obtained by selecting a single neighbor u.a.r. from every node in S. Then, w.h.p. $d(T_S) \in \Omega\left(N^{\frac{\beta-1}{\beta}-\epsilon}\right)$, for any constant $\epsilon > 0$.*

PROOF. The results follow directly from the previous lemma and by the fact that we can amplify the probability of getting a high degree node using $\Theta(\log N)$ samples. □

Using similar technique we can prove for $\beta = 2$ that the ratio is in $\Omega\left(\log^{\alpha}(N)\right)$, for any constant $\alpha > 0$.

THEOREM 5.6. *For any $\beta \in (1, 3)$, let S be a set of $\Theta(\log N)$ nodes sampled uniformly at random from $G(\beta, p)$, and let T_S be a set obtained by selecting a single neighbor u.a.r. from every node in S. Then, w.h.p. there is ratio between $d(T_S)$ and $d(S)$ is in $\omega(1)$.*

6. EXPERIMENTS

We conducted several experiments to evaluate and further study the friendship paradox in different online communities and social networks. Since analytical results only hold on stylized models, our primary motivation was to witness the existence of the phenomenon in the data. We used 8 different publicly available data sets: the Orkut social network [41], the Live Journal blogging community [41], the Wikipedia author network [26], the YouTube social network [41], the DBLP author network [41], the Slashdot user community network [28], and the Enron email communication network [22]. These networks vary in size and degree distribution and provide insight on the effect network parameters have on the friendship paradox.

We began by characterizing the networks according to their degree distribution. We fitted each network to a power law graph with different parameters of β and $C = e^{\alpha}$, using methods that optimize over final sum of squares of residuals. We summarize the main statistics in Table 1. Note that for almost all the networks $\beta \in (1, 3)$. We first considered the effects of the sample size and network topology.

The effects of sample size. To experimentally observe the way in which the sample size affects the friendship paradox, we compared the degree distributions of sampled sets and those of their neighbors as a function of the sample size. For each network, for every $j \in \{10, 20, 30, \ldots, 500\}$ we sampled j nodes u.a.r. and computed the degree distribution of the sampled nodes and the average degree distribution of all their neighbors. Note that the average degree distribution of all their neighbors is the average of selecting a neighbor at random from each node in the sample. Each such iteration is a snapshot of the friendship paradox for a set of size j. For each sample size j, we repeated this experiment 5 times. We then computed the ratio between the average degree of neighbors and the average degree in the sample. We plot the results in Figure 3.

As the figure shows, the friendship paradox varies across the different networks. The lowest gap we observed experimentally was in the DBLP network where the ratio averaged over all sample size iterations was 2.5 and the largest was 150 in the YouTube network. In all networks we see a large gap in the first iteration when the sample size is 10 and the neighbors' average degree distribution is large.

In all networks there is a trend of decrease in the friendship paradox as the sample size increases. This trend can be intuitively explained using the analysis from the previous sections: Since we showed that the friendship paradox is large at already the small sample size of $\log N$, this implies that when the sample S is of size $\log N$, the high degree nodes of the graph are in the neighborhood S. As the sample size increases we are more likely to sample high degree nodes in S while the degree distribution of S will becomes smaller as the high degree nodes are exhausted early.

Recall that in Section 4 we showed that with parameter $\beta \in (1, 2]$ the friendship paradox occurs even when the sample sizes are small. Interestingly, this phenomenon can also be observed on the data sets we examined, all of which were found to have parameters in this range, with the exception of the Orkut network for which $\beta < 1$.

One hypothesis is that the friendship paradox occurs due to the fact that in social networks neighbors indeed are more likely to have a high degree. An alternative hypothesis however could be that when considering neighbors, the size of the sample is larger. That is, a set of S nodes sampled at random generates a neighborhood of size $n = |\cup_{\text{u is neighbor of S}} \{u\}|$ and it may be that a set of size n sampled uniformly at random from the graph could have a similar effect. To test this hypothesis, we conducted the following experiment. For each $j = 1, \ldots, 100$ we sampled j nodes u.a.r. from the network which gave us a set of neighbors N_j. We then sampled $|N_j|$ nodes u.a.r. from the network, and compared the ratios between their average degree distribution. We depict the results in Figure 4 which supports our hypothesis.

To further investigate the effect of the sample size on the friendship paradox, we computed the degree distribution of *all* the nodes in the networks and for each node we averaged its' neighbors' degree distribution. In Figure 6 we plot the CDF of these distributions for the DBLP, Epinions, Slashdot, and the Enron networks. In the figure, the nodes' degree distribution is depicted in red and the average neighbors' degree distribution in blue. The expansion is the right shift between the distributions. The CDF shows that the majority of nodes' degree distribution is well below the average

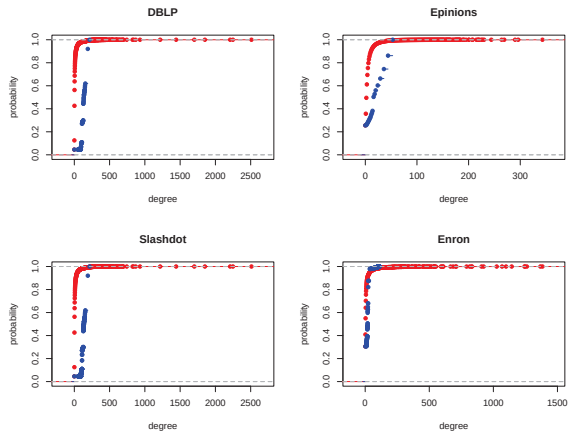

Figure 6: The CDFs of the degrees of nodes and the average degree of the neighbors in the network.

of their neighbors; however for nodes in top percentiles, this is no longer true. As the sample size increases these nodes are more likely to appear in the sample and diminish the expansion effect.

The effects of network topology. It is easy to see analytically that the friendship paradox cannot occur in regular graphs, where every node has the same degree. To view this experimentally, for each network $G = (V, E)$ in our data sets we also generated a random graph by running a process which assigns $|E|$ edges uniformly at random between $|V|$ nodes. We used the same process as above to observe the gap under different sample sizes. As expected, for these graphs, regardless of sample size, the average degree of the sample and that of their neighbors were nearly identical.

High degree nodes in neighbors vs. sample. To strengthen our results, we compared the average degree of the top 1% and 10% of neighbors and degrees from the sample. We plot the results in Figure 5 which show a clear dominance of selecting from the set of neighbors.

Beyond the first circle. It seems natural to ask whether the results extend beyond the immediate circle of friends. The first question is whether nodes which are two hops away from a randomly sampled node have a higher degree on average as well. The second question is whether the average degree grows as we further explore the graph. For a given set S, we use $\mathcal{N}_i(S)$ to denote the neighbors who are exactly i hops away from S in the graph. To observe the change in the degree distribution as a function of distance from a node, we conducted the following experiment. For each network, we sampled 5 nodes u.a.r., and conducted a BFS crawl of 4 levels from all nodes in the set. In each level we measured the degree distribution of all nodes in the level, and computed its average. In Figure 7 we plot the average degree as a function of the level, as well as the percentage of increase in average degree between levels.

The figure above gives experimental evidence that suggests that the friendship paradox would be most dramatic between the sampled set and its immediate neighborhood.

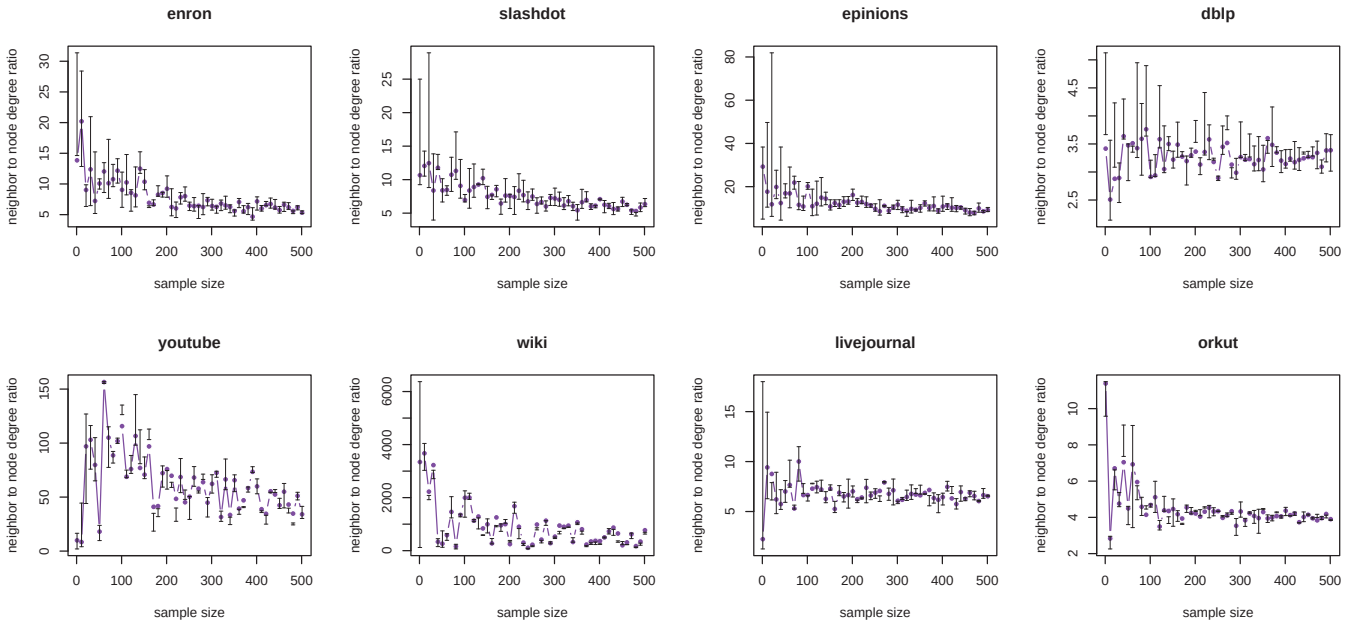

Figure 3: The ratio between average degrees of a sampled set and its neighbors.

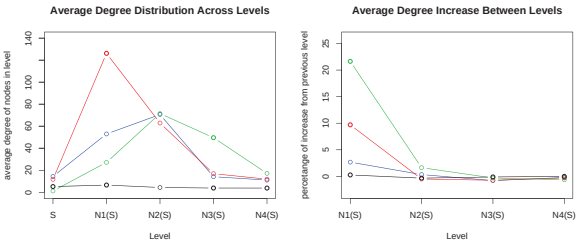

Figure 7: The degree distribution of a sampled set and its neighborhoods. Enron, Slashdot, Epinions, and DBLP correspond to the blue,red, green,black lines, respectively.

In the Slashdot and Epinions networks depicted in red and black, respectively, the degree distribution is maximized in $\mathcal{N}_1(S)$, and in the Enron and DBLP networks the maximum is achieved in $\mathcal{N}_2(S)$. In all networks the largest increase in percentage is between S and $\mathcal{N}_1(S)$. This phenomenon also seems to be a derivative of the fact that in power law graphs high degree nodes are likely to be found in the immediate neighborhoods of many nodes.

Applications. Besides being a fundamental sociological phenomena the friendship paradox has also practical applications. For example, it has already been used successfully to design immunization strategy [8, 9]. In this section we study an application of the friendship paradox to influence maximization. In particular we consider the case where we do not have knowledge about the network. From our experimental results in the previous setting it is clear that using random neighbors would be the best choice in the well-studied *voter model* [11, 12, 34, 37]. To strengthen our results we consider the classic *independent cascade* and *linear threshold* [20] and we study experimentally the effect of seeding a cascade in a random set or in random set of neighbors. Figure 8 shows that also in this more challenging setting using a set of random neighbors clearly outperforms a set of random nodes.

7. CONCLUSIONS

The friendship paradox is a fundamental phenomena in social networks. In this paper we strengthen the understanding of this phenomena by formalizing this concept and characterizing a large family of graphs where this phenomenon exists. One of the implications of our analysis is that for a large class of networks, the effectiveness of influence maximization strategies can be dramatically improved by using two-step approaches. Our theoretical results are supported by experiments that show this phenomena exists in social network data sets, and applies to influence maximization.

8. REFERENCES

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for power law graphs. In *STOC*, 2000.

[2] L. Backstrom and J. M. Kleinberg. Network bucket testing. In *WWW*, 2011.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, 2011.

[4] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science 286*.

[5] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European J. Combin.*, 1980.

[6] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier. Influence maximization in social networks: Towards

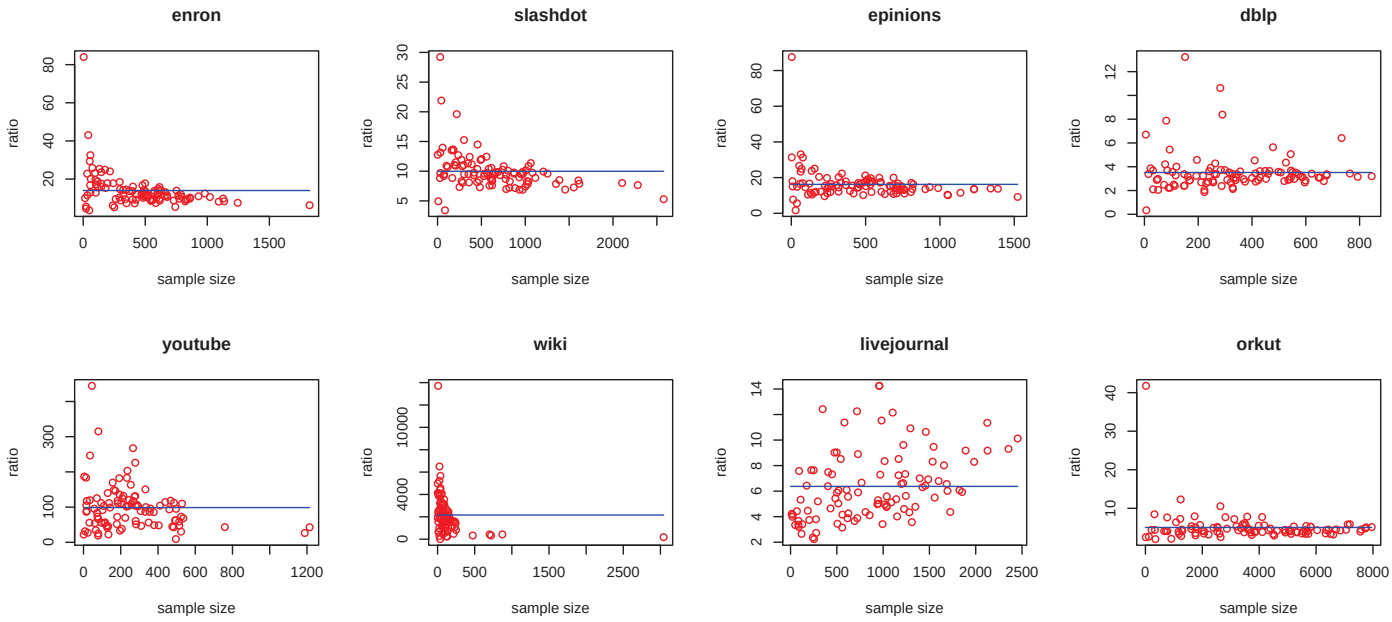

Figure 4: The ratio between average degree of a set of neighbors and a set of randomly sampled nodes from the graph.

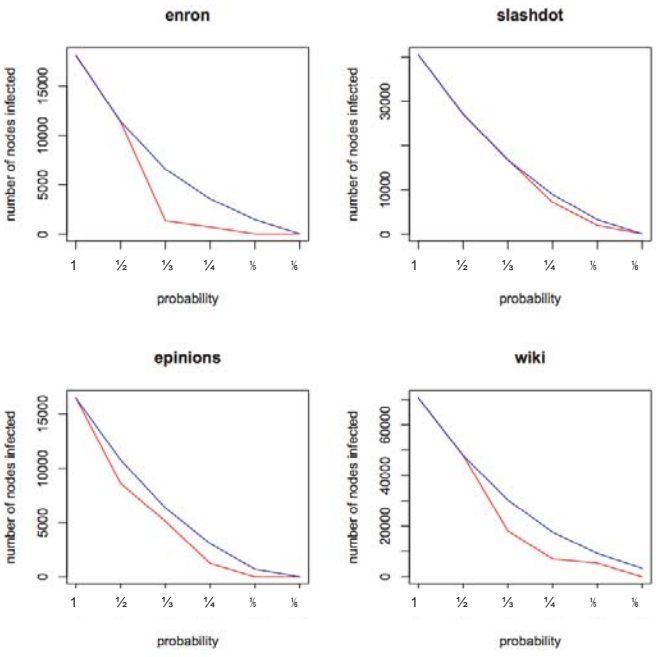

Figure 8: Number of the nodes influenced when a random set of node of size 10 is seeded(red line) as source of a cascade and when 10 random neighbors are seeded(blue line) as a function of the infection probability. Clearly the friendship paradox plays an important role also in this setting.

an optimal algorithmic solution. *arXiv preprint arXiv:1212.0884*, 2012.

[7] N. Chen. On the approximability of influence in social networks. In *SODA*, 2008.

[8] N. A. Christakis and J. H. Fowler. Social Network Sensors for Early Detection of Contagious Outbreaks. . In *PLoS ONE*, 2010.

[9] R. Cohen, S. Havlin, and D. Ben-Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 2003.

[10] A. Dasgupta, R. Kumar, and D. Sivakumar. Network bucket testing. In *KDD*, 2012.

[11] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD*, pages 57–66, 2001.

[12] E. Even-Dar and A. Shapira. A note on maximizing the spread of influence in social networks. *Inf. Process. Lett.*, 111(4):184–187, 2011.

[13] S. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 1991.

[14] S. L. Feld and B. Grofman. Variation in class size, the class size paradox, and some consequences for students. *Research in Higher Education*, 6, 1977.

[15] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD*, 2010.

[16] M. Granovetter. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1, 1983.

[17] N. O. Hodas, F. Kooti, and K. Lerman. Friendship paradox redux: Your friends are more interesting than you. *CoRR*, abs/1304.3480, 2013.

[18] L. Katzir, E. Liberty, and O. Somekh. Estimating sizes of social networks via biased sampling. . In *WWW*, 2011.

[19] L. Katzir, E. Liberty, and O. Somekh. Framework and algorithms for network bucket testing. . In *WWW*, 2012.

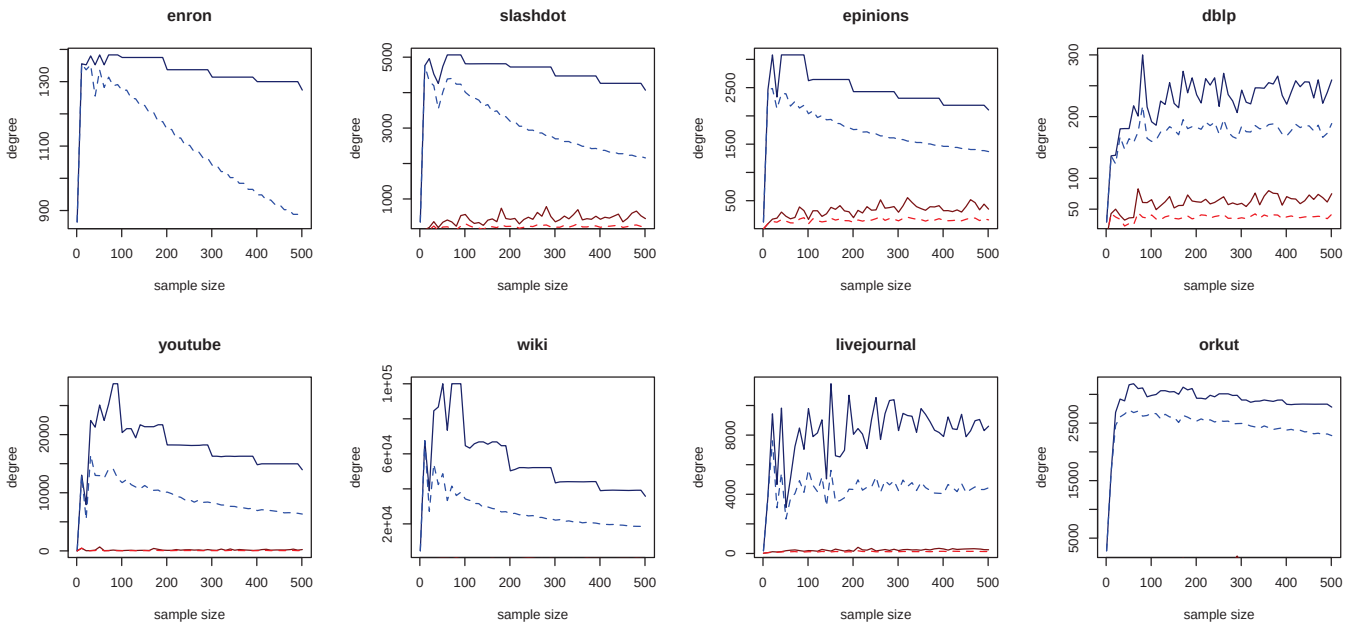

Figure 5: High degree nodes via neighbors vs. via sample. Blue lines depict top 1% and top10% (dotted) of neighbors; Red lines depict op 1% and top10% (dotted) of nodes from the sample.

[20] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*.

[21] J. Kleinberg. Navigation in a small world. *Nature*, 406:257–275, 2000.

[22] B. Klimt and Y. Yang. In *First Conference on Email and Anti-Spam (CEAS)*, Mountain View, CA.

[23] M. Lelarge. Efficient Control of Epidemics over Random Networks. In *SIGMETRICS*, 2009.

[24] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *ACM Conference on Electronic Commerce*, 2006.

[25] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD*, 2006.

[26] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting Positive and Negative Links in Online Social Networks. In *WWW*, 2010.

[27] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[28] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6, 2009.

[29] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen. Sparsification of influence networks. In *KDD*, 2011.

[30] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, 1989.

[31] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics 1 (2)*, 2004.

[32] E. Mossel and S. Roch. On the submodularity of influence in social networks. In *STOC*, 2007.

[33] M. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[34] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, pages 61–70, 2002.

[35] F. Santos and J. P. T. Lenaerts. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. *PNAS*, 2006.

[36] L. Seeman and Y. Singer. Adaptive seeding in social networks. In *FOCS*, 2013.

[37] Y. Singer. How to win friends and influence people, truthfully: influence maximization mechanisms for social networks. In *WSDM*, pages 733–742, 2012.

[38] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *CoRR*, abs/1111.4503, 2011.

[39] F. Viger and M. Latapy. Efficient and simple generation of random simple connected graphs with prescribed degree sequence. In *Proceedings of the 11th Annual International Conference on Computing and Combinatorics*, COCOON'05, pages 440–449, Berlin, Heidelberg, 2005. Springer-Verlag.

[40] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 409-410.

[41] J. Yang and J. Leskovec. Defining and Evaluating Network Communities based on Ground-truth. In *ICDM*, 2012.