This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2940914, IEEE Transactions on Knowledge and Data Engineering

"What Do Your Friends Think?": Efficient Polling Methods for Networks Using Friendship Paradox

Buddhika Nettasinghe, *Student Member, IEEE* and Vikram Krishnamurthy, *Fellow, IEEE*

Abstract—This paper deals with randomized polling of a social network. In the case of forecasting the outcome of an election between two candidates A and B, classical intent polling asks randomly sampled individuals: *who will you vote for?* Expectation polling asks: *who do you think will win?* In this paper, we propose a novel neighborhood expectation polling (NEP) strategy that asks randomly sampled individuals: *what is your estimate of the fraction of votes for A?* Therefore, in NEP, sampled individuals will naturally look at their neighbors (defined by the underlying social network graph) when answering this question. Hence, the mean squared error (MSE) of NEP methods rely on selecting the optimal set of samples from the network. To this end, we propose three NEP algorithms for the following cases: (i) the social network graph is not known but, random walks (sequential exploration) can be performed on the graph (ii) the social network graph is unknown. For both cases, algorithms based on a graph theoretic consequence called *friendship paradox* are proposed. Theoretical results on the dependence of the MSE of the algorithms on the properties of the network are established. Numerical results on real and synthetic data sets are provided to illustrate the performance of the algorithms.

Index Terms—opinion polling, election forecasting, expectation polling, friendship paradox, variance reduction, stochastic ordering, degree distribution, graph sampling, social networks, social sampling

1 Introduction

This paper deals with randomized polling of a social network with a possibly unknown structure. In the case of forecasting the outcome of an election between two candidates A and B, classical intent polling asks uniformly sampled individuals: *who will you vote for?* Expectation polling asks: *who do you think will win?* In this paper, we propose a novel neighborhood expectation polling strategy that asks non-uniformly sampled individuals: *what is your estimate of the fraction of votes for A?* Next, we formally define the problem, explain the solution approach and the related work that motivates it.

Consider a social network represented by an undirected graph $G = (V, E)$ where, each node $v \in V$ has a label $f(v) \in \{0, 1\}$. A pollster can query a total of $|S|$ (called the sampling budget) number of individuals from this social network.

Problem Definition. *Estimate,*

$$\bar{f} = \frac{|\{v \in V : f(v) = 1\}|}{|V|} \tag{1}$$

which is the fraction of nodes with label 1, with a sampling budget $|S| \ll |V|$ for the following cases:

- *Case 1 - graph $G = (V, E)$ is not known but, the graph can be explored sequentially using a random walk*
- *Case 2 - graph $G = (V, E)$ is not known but, the set of nodes V can be uniformly sampled*

We propose a class of polling methods that we call neighborhood expectation polling (NEP) to address the above problem[1]. In NEP, a set $S \subset V$ of individuals from the social network $G = (V, E)$ are selected and asked,

"What is your estimate of the fraction of people with label 1?".

When trying to estimate an unknown quantity about the world, any individual naturally looks at her neighbors. Therefore, each sampled individual $s \in S$ would provide the fraction of their neighbors $\mathcal{N}(s)$, with label 1. In other words, the response of the individual $s \in S$ for the NEP query would be,

$$q(s) = \frac{|\{u \in \mathcal{N}(s) : f(u) = 1\}|}{|\mathcal{N}(s)|}. \tag{2}$$

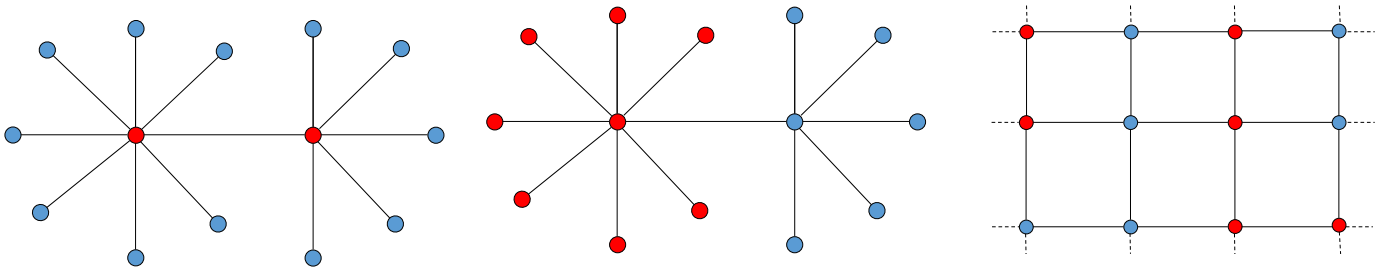
Then, the average of all the responses $\frac{\sum_{s \in S} q(s)}{|S|}$ is used as the NEP estimate of the fraction $\bar{f}$.

1.1 Context

Why call it NEP? NEP takes its name from the fact that, the response $q(s)$ of each sampled individual $s \in S$ is the expected label value among his/her neighbors i.e. $q(s) = \mathbb{E}\{f(U)\}$ where, U is a random neighbor of the sampled individual $s \in S$.

- *Authors are with the School of Electrical and Computer Engineering, Cornell University and Cornell Tech.*
  E-mail: {dwn26, vikramk}@cornell.edu.
- *This material is based upon work supported, in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under grant W911NF-19-1-0365, National Science Foundation under grant 1714180, and the U.S. Airforce Office of Scientific Research.*

[1] Applications of this problem include forecasting the outcome of an upcoming election [1], estimating the fraction of individuals infected with a disease [2], estimating the number of individuals interested in buying a certain product (a market research). More specific real world examples for case 1 and case 2 are discussed in Sec. 3.1 and Sec. 3.2 respectively.



(a) Network G_1: labels are highly correlated with the degrees of nodes

(b) Network G_2: nodes with the same label are clustered (depicting *Homophily*)

(c) Network G_3: a large regular graph with uniformly at random assigned labels

Fig. 1: Consider the case of uniformly sampling nodes and obtaining responses $q(s)$ of sampled nodes $s \in S$ about the fraction of red (i.e. label 1) nodes in the network. In graph G_1 of Fig. 1a, most nodes have their only neighbor to be of color red even though most of the nodes in the network are of color blue. Hence, NEP with uniformly sampled nodes would result in a highly biased estimate in this case. In graph G_2 of Fig. 1b, approximately half the nodes have only a red neighbor and, rest of the nodes have only a blue neighbor. Hence, NEP with uniformly sampled nodes would result in an estimate with a large variance in this case. In graph G_3 of Fig. 1c, average of the NEP responses $q(v)$ of nodes is approximately equal to the fraction $\bar{f}$ of nodes with red labels. Further, $q(v)$ does not vary largely among nodes. Hence, uniformly sampling nodes for NEP in this case would result in an accurate estimate. Similar examples can also be found in [3]. This figure highlights the importance of exploiting network structure and node label distribution when sampling nodes to be used for NEP.

Why (not) use NEP? NEP is substantially different to classical intent polling where, each sampled individual is asked *"What is your label?"*. In intent polling, the response of each sampled individual $s \in S$ is his/her label $f(s)$. In contrast, in NEP, the response $q(s)$ of each sampled individual $s \in S$ is a function of his/her neighborhood (defined by the underlying graph G) as well as the labels of his/her neighbors. Therefore, depending on the graph G, function f and the method of obtaining the samples S, NEP might produce either,

I. an estimate with a larger MSE compared to intent polling (e.g. networks in Fig. 1a and Fig. 1b shows when uniform sampling of individuals for NEP might not work), or,

II. an estimate with a smaller MSE compared to intent polling (e.g. network in Fig. 1c shows when uniform sampling of individuals for NEP might work)

These two possible outcomes highlight the importance of using the available information about the graph G and the function f, when selecting the set S of individuals in NEP. This lead us to the main results of this paper where we combine NEP with *friendship paradox* (reviewed in Sec. 2) based sampling methods to obtain statistically efficient estimates.

Remark 1. The assumption that the graph is not fully known (case 1 and case 2 in problem definition) is applicable to most contexts that deal with large scale real world networks (including online social networks such as Facebook). This is mostly due to the fact that structures of social networks are not made available publicly by online social network network administrators and accurately estimating the network structure would incur costs (computation, memory, querying cost, etc.) that are not feasible in the context of polling. In contrast, our methods do not rely on estimating the network structure and instead, rely on *friendship paradox* based sampling method.

Remark 2. If the graph $G = (V, E)$ is fully known, a greedy (deterministic) optimization method (similar to the one in [4]) can be used to solve the NP hard problem of finding the set $S \subset V$ of $|S|$ individuals whose collective neighborhood is largest, with a $(1 - 1/e)$ approximation guarantee. However, the largest collective neighborhood does not ensure that the set S of individuals would provide an accurate NEP estimate of the fraction $\bar{f}$ defined in (1) e.g. if the sampling budget $|S| = 1$, the node with the largest collective neighborhood in the graph G_2 in Fig. 1b is the red color node with degree seven, whose NEP response (fraction of red neighbors) is $q(s) = 1$, even though $\bar{f} = 4/7$. Hence, our focus is on randomized sampling methods for NEP that do not require the graph to be known.

1.2 Main Results and Organization

The main results of this paper are NEP algorithms for the two cases described in the problem definition and their analysis. The algorithms utilize properties related to the structure of the network to find $|S|$ number of samples. The analysis provides simple and intuitive conditions under which, the proposed algorithms will provide a better estimate compared to intent polling. These results can be summarized as follows.

- For case 1 and case 2, estimation algorithms are obtained by combining NEP with recent statistical results related to a phenomenon called*friendship paradox* [5]. Analytical results characterizing the dependence of bias, variance and MSE of estimates on the properties of the graph G, labels $f(v)$ of individuals $v \in V$ are obtained. These results help to identify conditions on the graph and the labels for which, friendship paradox based NEP produces a better estimate compared to intent polling and naive NEP with uniformly sampled individuals.

- Empirical and simulation results on five real world social network datasets and synthetic datasets are pro-

vided, illustrating the performance of the proposed algorithms compared to classical methods. These empirical and simulation results yield useful insights that complement the analytical results.

Organization: Sec. 2 presents a review of the key results related to friendship paradox. Sec. 3 presents the two NEP algorithms based on the friendship paradox for case 1 and case 2, followed by their theoretical analysis in Sec. 4. Sec. 5 evaluates the proposed algorithms on empirical and synthetic datasets to illustrate and compare their performances. Finally, Sec. 6 provides a discussion about the two algorithms, their theoretical and experimental evaluations and how they relate to each other.

Notation: Table 1 summarizes the parameters and variables used frequently throughout the paper.

1.3 Related work

As described above, in the classical intent polling[2], a set S of nodes is obtained by uniform sampling with replacement and then, the average of their labels

$$I^{|S|} = \frac{\sum_{u \in S} f(u)}{|S|}, \tag{3}$$

is used as the estimate (called intent polling estimate henceforth) of the fraction $\bar{f}$ defined in (1). The main limitation of intent polling is that the sample size needed to achieve a ϵ- additive error is $O(\frac{1}{\epsilon^2})$ [3]. Our work is motivated by two recently proposed methods, namely "expectation polling" [6] and "social sampling" [3], that attempt to overcome this limitation in intent polling.

Firstly, in expectation polling [6], each sampled individual provides an estimate of the label held by the majority of the individuals in the network (i.e. sampled individuals answer the question "Who do you think will win the election?"). Then, each sampled individual will look at her neighbors and provide the value held by the majority of them. This method is more efficient (in terms of sample size) compared to the intent polling method since each sampled individual now provides the putative response of a neighborhood[3,4]. Secondly, in social sampling [3], the response of each sampled individual is a function of the labels, degrees and the sampling probabilities of her neighbors. [3] provides several unbiased estimators for the fraction $\bar{f}$ using this method and, establishes bounds for their variances. The main limitation of social sampling method (compared to NEP) is that it requires the sampled individuals to know a significant amount of information about the underlying network. Therefore, a practical implementation of social sampling might not be feasible in settings with limited

[2]This method is called intent polling because, in the case of predicting the outcome of an election, this is equivalent to asking the voting intention of sampled individuals i.e. asking "Who are you going to vote for in the upcoming election?") [6].

[3]Intent polling and expectation polling have been considered intensively in literature, mostly in the context of forecasting elections and, it is generally accepted that expectation polling is more efficient compared to intent polling [7], [8], [9], [10], [11].

[4] [12], [13] discuss how expectation polling can give rise to misinformation propagation in social learning and, propose Bayesian filtering methods to eliminate the misinformation propagation.

TABLE 1: Summary of Notation

		Network Parameters
$G = (V, E)$	$\triangleq$	Undirected graph with set of nodes V and set of edges E
A	$\triangleq$	Symmetric adjacency matrix of the graph G where $A(u,v) = \begin{cases} 1, & \text{if } (u,v) \in E \\ 0, & \text{otherwise} \end{cases}$
n	$\triangleq$	Number of nodes i.e. $n = \|V\|$
M	$\triangleq$	Number of friends i.e. $M = 2\|E\|$
$\mathcal{N}(v)$	$\triangleq$	The set of neighbors of a node $v \in V$ as defined by the graph G
$d(v)$	$\triangleq$	Degree of node $v \in V$ i.e. $d(v) = \|\mathcal{N}(v)\|$
$f(v)$	$\triangleq$	Binary label of node $v \in V$
$\bar{f}$	$\triangleq$	Fraction of nodes with label 1 i.e. $\bar{f} = \frac{\|\{v \in V : f(v) = 1\}\|}{\|V\|}$
$q(v)$	$\triangleq$	NEP response of node $v \in V$ i.e. $q(v) = \frac{\|\{u \in \mathcal{N}(v) : f(u) = 1\}\|}{\|\mathcal{N}(v)\|}$
D	$\triangleq$	Diagonal matrix with $D(v,v) = d(v)$
$\mathcal{A}$	$\triangleq$	Normalized adjacency matrix $\mathcal{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$
		Random Variables, Distributions and Related Parameters
X	$\triangleq$	Uniformly sampled node from set of nodes V
Y	$\triangleq$	Random friend: uniform sampled end of a uniformly sampled edge from E
Z	$\triangleq$	Random friend of a random node
$P(k)$	$\triangleq$	Degree distribution which gives the probability that a random node X has degree k
$q(k)$	$\triangleq$	Neighbor degree distribution that gives the probability that a random friend Y has degree k
$e(k,k')$	$\triangleq$	Joint degree distribution that gives the probability that a random edge (U,Y) will have nodes with degrees $d(U) = k, d(Y) = k'$
σ_k	$\triangleq$	Standard deviation of the degree $d(X)$ of a random node X i.e. standard deviation of the degree distribution
σ_f	$\triangleq$	Standard deviation of the label $f(X)$ of a random node X
r_{kk}	$\triangleq$	Neighbor degree correlation coefficient defined in (28)
ρ_{kf}	$\triangleq$	Degree-label correlation coefficient defined in (29)
		Polling Estimates and Related Parameters
S	$\triangleq$	Set of the individuals queried by the pollster
$\|S\|$	$\triangleq$	Sampling budget (number of individuals queried by the pollster)
N	$\triangleq$	Length of Random Walk (for Algorithm 1)
$I^{\|S\|}$	$\triangleq$	Intent polling estimate defined in (3)
$T_{UN}^{\|S\|}$	$\triangleq$	Naive NEP estimate with uniformly sampled nodes defined in (7)
$T_{RW}^{\|S\|}$	$\triangleq$	NEP estimate obtained via proposed Algorithm 1
$T_{FN}^{\|S\|}$	$\triangleq$	NEP estimate obtained via proposed Algorithm 2

information about a very large graph. Hence, NEP can be thought of a as a method which asks a question that seeks a finer resolution compared to expectation polling and yet, simpler and intuitive compared to social sampling.

The key idea utilized in our proposed NEP estimators for case 1 and case 2 (stated in problem definition) is the friendship paradox (detailed in Sec. 2), which is a form of network sampling bias observed in undirected graphs. Friendship paradox has recently gained attention in several applications related to networks under the broad theme "how network biases can be used effectively for estimation problems?". For example, [14], [15] show how friendship paradox can be utilized for accurate estimation of a heavy tailed degree distribution, [16], [17] show how friendship paradox can be used for quickly detecting a disease outbreak. Our results for the case 1 and case 2 also fall under this broad theme. Apart from the applications in estimation problems, friendship paradox has been explored also in the contexts of perception biases in social networks [18], [19], [20], information diffusion and opinion formation [21], [22], [23], [24], influence maximization and stochastic seeding [25], [26], [27], node properties other than the degrees [28], [29], [30] and directed social networks [18], [28], [31].

2 What is Friendship Paradox?

"Friendship paradox" is a graph theoretic consequence first presented in [5] by Scott L. Feld in 1991. The friendship paradox states, "*on average, the number of friends of a random friend is always greater than the number of friends of a random individual*". Formally:

Theorem 1. *(Friendship Paradox [5]) Consider an undirected graphs $G = (V, E)$. Let X be a node chosen uniformly from V and, Y be a uniformly chosen node from a uniformly chosen edge $e \in E$. Then,*

$$\mathbb{E}\{d(Y)\} \geq \mathbb{E}\{d(X)\}, \tag{4}$$

where, $d(X)$ and $d(Y)$ denote the degrees of X and Y, respectively.

In Theorem 1, the random variable Y is called a random friend (or a random neighbor) since it is obtained by sampling a pair of friends (i.e. an edge from the graph) uniformly and then choosing one of them by an unbiased coin flip. The intuition behind Theorem 1 is as follows. Individuals with large numbers of friends appear as the friends of a large number of individuals. Hence, such popular individuals can contribute to an increase in the average number of friends of friends. On the other hand, individuals with smaller numbers of friends appear as friends of a smaller number of individuals. Hence, they cannot cause a significant change in the average number of friends of friends. Further, [32] shows that the original version of the friendship paradox (Theorem 1) is a consequence of the monotone likelihood ratio ordering between random variables $d(Y)$ and $d(X)$.

Refinements of Friendship Paradox. Recall that friendship paradox, in its original version given in Theorem 1, is a comparison between the degrees of a random individual X and a random friend Y (obtained by sampling an edge uniformly and then choosing one end of it by an unbiased coin flip). However, a more intuitive comparison would be the comparison of degree $d(X)$ of a random individual X and the degree $d(Z)$ of a random friend Z of a random individual. [32] develops the following important refinement of the friendship paradox which achieves this.

Theorem 2. *[32] Let $G = (V, E)$ be an undirected graph, X be a node chosen uniformly from V and, Z be a uniformly chosen neighbor of a uniformly chosen node from V. Then,*

$$d(Z) \geq_{fosd} d(X) \tag{5}$$

where, $\geq_{fosd}$ denotes the first order stochastic dominance[5].

An immediate consequence of Theorem 2 is,

$$\mathbb{E}\{d(Z)\} \geq \mathbb{E}\{d(X)\}, \tag{6}$$

which says that a random neighbor of a random individual has more friends than a random individual, on average (from the fact that first order stochastic dominance implies larger mean).

With the above background, we present the NEP algorithms that are based on Theorem 1 and Theorem 2.

3 NEP Algorithms Based on Friendship Paradox

In this section, we consider randomized methods for selecting individuals for NEP based on the concept of friendship paradox explained in Sec. 2.

For notational reference, we first describe a naive NEP method that does not exploit the friendship paradox.

Naive NEP Algorithm:

Step 1: Obtain a set S of uniformly sampled nodes from V and the NEP response $q(s)$ (defined in (2)) from each $s \in S$.

Step 2: Compute the naive NEP estimate of $\bar{f}$ in (1) as,

$$T_{UN}^{|S|} = \frac{\sum_{s \in S} q(s)}{|S|} \tag{7}$$

Note from the step 1 of the naive NEP method that the naive NEP estimate $T_{UN}^{|S|}$ of $\bar{f}$ (fraction of nodes with label 1) is based on the NEP responses of uniformly sampled nodes i.e. answers of uniformly sampled individuals to the question "*What is your estimate of the fraction of people with label 1?*" . Hence, the naive NEP algorithm exploits one's knowledge about her neighbors but does not exploit friendship paradox based sampling. Our main contribution below is to develop NEP algorithms that exploit friendship paradox based sampling (Sec. 3.1 and Sec. 3.2) and show that they are more accurate compared to the naive NEP estimate (7) in terms of mean-squared error under various network structures.

[5] A random variable X (with a cumulative distribution function F_X) first order stochastically dominates a discrete random variable Y (with a cumulative distribution function F_Y), denoted $X \geq_{fosd} Y$ if, $F_X(n) \leq F_Y(n)$, for all n.

3.1 Case 1 - Sampling Friends using Random Walks

This subsection considers the case where the graph $G = (V, E)$ is not known initially, but sequential exploration of the graph is possible using multiple random walks (case 1 of problem definition) over the nodes of the graph.

A motivating example for case 1 is a massive online social network where the fraction of user profiles with a certain characteristic needs to be estimated (e.g. profiles with more than ten posts about a product). Web-crawling (using random walks) approaches are widely used to obtain samples from such massive online social networks without requiring the global knowledge of the full network graph [33], [34], [35], [36], [37].

Algorithm 1: NEP with Random Walk Based Sampling

Input: $|S|$ number of samples $\{v_1, v_2, \ldots, v_{|S|}\} \subset V$.
Output: $T_{RW}^{|S|}$ which is the estimate of the fraction $\bar{f}$ of nodes with label 1.

1) Initialize $|S|$ independent random walks on the social network starting from $v_1, v_2, \ldots, v_{|S|}$.
2) Run each random walk for a N steps. Then collect sample $S = \{s_1, \ldots, s_{|S|}\}$ where, $s_i \in V$ is collected from i^{th} random walk.
3) Query each $s \in S$ to obtain NEP response $q(s)$ (defined in (2)) and, compute the estimate

$$T_{RW}^{|S|} = \frac{\sum_{s \in S} q(s)}{|S|}.$$

We propose Algorithm 1 for estimating the fraction $\bar{f}$ in case 1. The intuition behind Algorithm 1 stems from the fact that the stationary distribution of a random walk on an undirected graph (which is connected and non-bipartite) is the uniform distribution over the set of neighbors [38]. Therefore, Algorithm 1 obtains a set S of $|S|$ neighbors independently from the graph $G = (V, E)$ for sufficiently large N (i.e. one sample from each of the $|S|$ independent random walks) in the step 2. Then, the response $q(s)$ of each sampled individual $s \in S$ for the NEP query is used to compute the estimate $T_{RW}^{|S|}$ in step 3. According to the friendship paradox (Theorem 1), NEP with random neighbors is equivalent to using more node labels (than NEP with random nodes) due to the fact that random neighbors have more neighbors than random nodes on average. Hence, it is intuitive that the variance of this method should be smaller compared to the naive NEP (with uniformly sampled nodes) and intent polling method. In Sec. 4, we verify this claim theoretically and, explore the properties of the underlying network for the estimate $T_{RW}^{|S|}$ to have a smaller MSE compared to the intent polling method.

3.2 Case 2 - Sampling a Random Friend of a Random Individual

In case 1 (Sec. 3.1), we assumed that it is possible to crawl the unknown graph using random walks. Instead, in case 2, we assume that a set of uniform samples $S = \{s_1, \ldots, s_{|S|}\}$ from the set of nodes V can be obtained and, each sampled individual $s_i \in S$ has the ability to answer the question *"What is your (random) friend's estimate of the fraction of individuals with label 1?"*.

A motivating example for case 2 is the situation where random individuals are requested to answer survey questions for an incentive. In such cases, the pollster usually does not have any information about the structural connectivity of the queried individuals and, will only be able to obtain their answer for a question.

For this case, we propose Algorithm 2 to obtain an estimate of the fraction $\bar{f}$ of individuals with label 1.

Algorithm 2: NEP with Random Friend Sampling

Input: $|S|$ number of uniform samples $S = \{s_1, s_2, \ldots, s_{|S|}\} \subset V$.
Output: $T_{FN}^{|S|}$ which is the estimate of the fraction $\bar{f}$ of nodes with label 1.

1) Ask each $s_i \in S$ to provide $q(u_i)$ (defined in (2)) for some randomly chosen neighbor $u_i \in \mathcal{N}(s_i)$.
2) Compute the estimate,

$$T_{FN}^{|S|} = \frac{\sum_{i=1}^{|S|} q(u_i)}{|S|}.$$

In Algorithm 2, each uniformly sampled individual $s_i \in S$ answers the question *"What is your (random) friend's estimate of the fraction of individuals with label 1?"* by providing $q(u_i)$ for a randomly chosen neighbor $u_i \in \mathcal{N}(s_i)$. The reasoning behind this method stems from Theorem 2 which states that, a random friend of a randomly chosen individual has more friends than a randomly chosen individual on average[6]. Therefore, this method should result in a smaller variance compared to naive NEP (7) and intent polling (3).

Remark 3. One can think of Algorithm 2 as a special case of Algorithm 1 with the random walk length set to $N = 1$. By the same argument, the naive NEP algorithm then correspond to a random walk with length $N = 0$ for the purpose of comparing the three NEP algorithms. The length of the random walk is used in Sec. 6 to discuss how friendship paradox based NEP methods achieve a bias-variance trade-off. We refer to [39] which also explores the friendship paradox using random walk length.

4 Statistical Analysis of the Estimates Obtained via Algorithm 1 and Algorithm 2

Algorithm 1 and Algorithm 2 presented in Sec. 3 query random friends (denoted by Y in Theorem 1) and random friends of random nodes (denoted by Z in Theorem 2) respectively, exploiting the friendship paradox. In this context, the aim of this section is to analyze the bias, variance and the

[6]This does not follow from the original version of friendship paradox (Theorem 1) since the random friend is not a uniformly chosen neighbor from the set of all $2|E|$ neighbors. Instead, the response is obtained from a random neighbor of a uniformly sampled node.

mean-squared error (MSE)[7] of the estimates obtained using these proposed algorithms to show that they outperform alternative methods (intent polling and naive NEP without friendship paradox). More specifically,

1) Theorem 3 motivates the use of friendship paradox based NEP algorithms (compared to the naive NEP with uniformly sampled nodes) by considering the case where the label of each node is assigned by an independent and identically distributed coin toss.
2) Theorem 4 relates bias and variance of the estimate $T_{RW}^{|S|}$ obtained using Algorithm 1 to network properties such as degree label correlation and absence of bottlenecks. Then, Corollary 5 gives sufficient conditions on the sampling budget $|S|$ for which the Algorithm 1 has a smaller MSE compared to intent polling.
3) Theorem 6 characterizes the bias and variance of the naive NEP (with uniformly sampled nodes and hence, not exploiting friendship paradox) and Corollary 7 compares the worst case performance of friendship paradox based NEP (Algorithm 1) with naive NEP to highlight how friendship paradox results in a reduced variance.
4) Theorem 8 characterizes the bias and variance of the estimate $T_{FN}^{|S|}$ obtained using Algorithm 2 and relates them to properties of the underlying network.

4.1 Independent and Identically Distributed Labels

Consider graph $G = (V, E)$ where each node $v \in V$ has a binary label $f(v) \in \{0, 1\}$ that is a Bernoulli random variable which is independent of and identically distributed to other labels. The following result shows how friendship paradox based sampling (Algorithm 1 and Algorithm 2) results in reduced variance NEP estimates.

Theorem 3. *Let the set of labels $\{f(v) : v \in V\}$ be independent and identically distributed (iid) Bernoulli random variables. Then,*

$$\mathrm{MSE}\{T_{FN}^{|S|}\} \leq \mathrm{MSE}\{T_{UN}^{|S|}\} \tag{9}$$

$$\mathrm{MSE}\{T_{RW}^{|S|}\} \leq \mathrm{MSE}\{T_{UN}^{|S|}\} \tag{10}$$

where, MSE *denotes mean square error defined in (8),*
$T_{UN}^{|S|}$ is the naive NEP estimate (7),
$T_{RW}^{|S|}$ is the estimate obtained using Algorithm 1,
$T_{FN}^{|S|}$ is the estimate obtained using Algorithm 2.

Proof. By definition,

$$\mathbb{E}\{T_{RW}^{|S|}\} = \mathbb{E}\{q(Y)\} = \mathbb{E}\left\{\frac{\sum_{u\in\mathcal{N}(Y)} f(u)}{d(Y)}\right\}$$

$$\mathbb{E}\{T_{FN}^{|S|}\} = \mathbb{E}\{q(Z)\} = \mathbb{E}\left\{\frac{\sum_{u\in\mathcal{N}(Z)} f(u)}{d(Z)}\right\}$$

$$\mathbb{E}\{T_{UN}^{|S|}\} = \mathbb{E}\{q(X)\} = \mathbb{E}\left\{\frac{\sum_{u\in\mathcal{N}(X)} f(u)}{d(X)}\right\}.$$

[7]The mean-squared error (MSE) of estimate T of a parameter $\bar{f}$ is

$$\mathrm{MSE}\{T\} = \mathbb{E}\{(T - \bar{f})^2\} = \mathrm{Bias}\{T\}^2 + \mathrm{Var}\{T\}. \tag{8}$$

Consider $\mathbb{E}\{T_{RW}^{|S|}\}$.

$$\mathbb{E}\{T_{RW}^{|S|}\} = \mathbb{E}\left\{\frac{\sum_{u\in\mathcal{N}(Y)} f(u)}{d(Y)}\right\} = \mathbb{E}\left\{\mathbb{E}\left\{\frac{\sum_{i=1}^{k} L_i}{k}\middle| d(Y) = k\right\}\right\}$$

where, $L_i,\ i = 1, \ldots, k$ are the iid labels of the neighbors of Y. Since the labels L_i are iid, the inner expectation becomes $\mathbb{E}\{f(X)\}$. Therefore,

$$\mathbb{E}\{T_{RW}^{|S|}\} = \mathbb{E}\{f(X)\} = \bar{f}.$$

Following similar arguments, we also get,

$$\mathbb{E}\{T_{FN}^{|S|}\} = \mathbb{E}\{T_{UN}^{|S|}\} = \mathbb{E}\{f(X)\} = \bar{f}.$$

Therefore, the estimates are unbiased when the labels are iid.

Next, consider the variances of the estimate $T_{RW}^{|S|}$. Since all $|S|$ samples are independent,

$$\mathrm{Var}\{T_{RW}^{|S|}\} = \frac{1}{|S|}\,\mathrm{Var}\left\{\frac{\sum_{u\in\mathcal{N}(Y)} f(u)}{d(Y)}\right\}$$

By applying the law of total variance, we get,

$$\mathrm{Var}\{T_{RW}^{|S|}\} = \frac{1}{|S|}\left[\mathrm{Var}\left\{\mathbb{E}\left\{\frac{\sum_{u\in\mathcal{N}(Y)} f(u)}{d(Y)}\middle| d(Y)\right\}\right\} + \mathbb{E}\left\{\mathrm{Var}\left\{\frac{\sum_{u\in\mathcal{N}(Y)} f(u)}{d(Y)}\middle| d(Y)\right\}\right\}\right]$$

$$= \frac{\sigma_f^2}{|S|}\mathbb{E}\left\{\frac{1}{d(Y)}\right\} \text{ (since the labels are iid)}$$

where, σ_f^2 denotes the variance of iid labels i.e. $\sigma_f^2 = \mathrm{Var}\{f(X)\}$. Following similar steps, we obtain,

$$\mathrm{Var}\{T_{FN}^{|S|}\} = \frac{\sigma_f^2}{|S|}\mathbb{E}\left\{\frac{1}{d(Z)}\right\},\ \mathrm{Var}\{T_{UN}^{|S|}\} = \frac{\sigma_f^2}{|S|}\mathbb{E}\left\{\frac{1}{d(X)}\right\}.$$

Then, the result follows by noting that

$$\frac{1}{d(X)} \geq_{fosd} \frac{1}{d(Y)}, \quad \frac{1}{d(X)} \geq_{fosd} \frac{1}{d(Z)} \tag{11}$$

where, $\geq_{fosd}$ denotes the first order stochastic dominance defined in Footnote 5 in Sec. 2. Eq. (11) follows immediately from Theorem 1 and Theorem 2 (note that $d(\cdot)$ is strictly positive for connected graphs). □

Theorem 3 shows that friendship paradox based NEP methods (Algorithm 1 and Algorithm 2) have smaller MSE compared to naive NEP (7) when the node labels are iid Bernoulli random variables. A natural question is "How do friendship paradox based NEP methods perform when the node labels are from an arbitrary joint distribution?". We consider this next.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2940914, IEEE Transactions on Knowledge and Data Engineering

4.2 Arbitrarily Assigned Node Labels

In the remainder of this section, we assume that node labels $\{f(v) : v \in V\}$ are already assigned from an arbitrary joint distribution or deterministically specified.

We first characterize the bias $\text{Bias}\{T_{RW}^{|S|}\}$ and the variance $\text{Var}\{T_{RW}^{|S|}\}$ of the estimate $T_{RW}^{|S|}$ obtained via Algorithm 1 as the random walk length N goes to infinity. Define the $|V| \times |V|$ dimensional diagonal matrix D and the normalized adjacency matrix $\mathcal{A}$ as,

$$D(v,v) = d(v), \quad \mathcal{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}. \tag{12}$$

Let $||Q||$ denote the spectral norm of a matrix Q (recall that the spectral norm is the maximum singular value).

Theorem 4. *Let $G = (V, E)$ be a connected, non-bipartite graph. Then, as the random walk length N tends to infinity, the bias $\text{Bias}\{T_{RW}^{|S|}\}$ and the variance $\text{Var}\{T_{RW}^{|S|}\}$ of the estimate $T_{RW}^{|S|}$, obtained via Algorithm 1 are given by,*

$$\begin{aligned} \text{Bias}(T_{RW}^{|S|}) &= \mathbb{E}\{f(Y)\} - \mathbb{E}\{f(X)\} \\ &= \frac{\text{Cov}\{f(X), d(X)\}}{\mathbb{E}\{d(X)\}} \end{aligned} \tag{13}$$

$$\begin{aligned} \text{Var}\{T_{RW}^{|S|}\} &= \frac{1}{|S|M} f^T D^{\frac{1}{2}} \Big(\mathcal{A}^2 - \frac{1}{M} D^{\frac{1}{2}} \mathbb{1}\mathbb{1}^T D^{\frac{1}{2}} \Big) D^{\frac{1}{2}} f \\ &\leq \frac{1}{|S|} \lambda_2^2 \mathbb{E}\{f(Y)\} \end{aligned} \tag{14}$$

where, X is a random node, Y is a random friend, M is the total number of friends, λ_2 is the second largest singular value of the normalized adjacency matrix $\mathcal{A}$ (defined in (12)) and f is a column vector with label $f(v) \in \{0, 1\}$ of node v at v^{th} element.

Proof. If $G = (V, E)$ is a connected, non-bipartite graph, then the stationary distribution of a random walk on G samples each $v \in V$ with a probability proportional to the degree $d(v)$ of v (page 298, [40]). Equivalently, sampling from the stationary distribution of a random walk on a finite connected, non-bipartite graph is equivalent to sampling friendships $(U, Y) \in E$ uniformly. Therefore,

$$\begin{aligned} \text{Bias}(T_{RW}^{|S|}) &= \mathbb{E}\{T_{RW}^{|S|}\} - \bar{f} = \mathbb{E}\{q(U)\} - \bar{f} \\ &= \mathbb{E}\{f(Y)\} - \mathbb{E}\{f(X)\} \\ &= \sum_{v \in V} f(v) \frac{d(v)}{\sum_{v \in V} d(v)} - \frac{\sum_{v \in V} f(v)}{|V|} \\ &= \frac{\mathbb{E}\{f(X)d(X)\} - \mathbb{E}\{f(X)\}\mathbb{E}\{d(X)\}}{\mathbb{E}\{d(X)\}} \\ &= \frac{\text{Cov}\{f(X), d(X)\}}{\mathbb{E}\{d(X)\}} \end{aligned}$$

To obtain the variance of $q(Y)$, let e_v denote the $n \times 1$ dimensional unit vector with 1 at the v^{th} element and zeros elsewhere. Then, $q(v) = e_v^T D^{-1} A f$. Hence,

$$\begin{aligned} \mathbb{E}\{q(Y)\} &= \sum_{v \in V} \frac{d(v)}{M} e_v^T D^{-1} A f = \frac{1}{M} \mathbb{1}^T D D^{-1} A f \\ &= \frac{1}{M} \mathbb{1}^T A f = \frac{1}{M} \mathbb{1}^T D f \end{aligned} \tag{15}$$

$$\begin{aligned} \mathbb{E}\{q^2(Y)\} &= \sum_{v \in V} \frac{d(v)}{M} f^T A D^{-1} e_v e_v^T D^{-1} A f \\ &= \frac{1}{M} f^T A D^{-1} A f. \end{aligned} \tag{16}$$

Therefore,

$$\begin{aligned} &\text{Var}\{q(Y)\} = \mathbb{E}\{q^2(Y)\} - \mathbb{E}\{q(Y)\}^2 \\ &= \frac{1}{M} f^T A D^{-1} A f - \frac{1}{M^2} f^T D \mathbb{1}\mathbb{1}^T D f \\ &= \frac{1}{M} f^T D^{\frac{1}{2}} \Big(\Big(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \Big)^2 - \Big(\frac{D^{\frac{1}{2}} \mathbb{1}}{\sqrt{M}} \Big) \Big(\frac{\mathbb{1}^T D^{\frac{1}{2}}}{\sqrt{M}} \Big) \Big) D^{\frac{1}{2}} f \\ &= \frac{1}{M} f^T D^{\frac{1}{2}} \Big(\mathcal{A}^2 - \Big(\frac{D^{\frac{1}{2}} \mathbb{1}}{\sqrt{M}} \Big) \Big(\frac{\mathbb{1}^T D^{\frac{1}{2}}}{\sqrt{M}} \Big) \Big) D^{\frac{1}{2}} f, \end{aligned}$$

where $\mathcal{A}$ denotes the normalized adjacency matrix defined in (12). Note that $\frac{D^{\frac{1}{2}} \mathbb{1}}{\sqrt{M}}$ is the eigenvector corresponding to the largest eigenvalue 1 of $\mathcal{A}^2$. Therefore, we get

$$\begin{aligned} &\Big| \frac{1}{M} f^T D^{\frac{1}{2}} \Big(\mathcal{A}^2 - \Big(\frac{D^{\frac{1}{2}} \mathbb{1}}{\sqrt{M}} \Big) \Big(\frac{\mathbb{1}^T D^{\frac{1}{2}}}{\sqrt{M}} \Big) \Big) D^{\frac{1}{2}} f \Big| \\ &\quad \leq \Big|\Big| \frac{D^{\frac{1}{2}} f}{\sqrt{M}} \Big|\Big| \times \Big|\Big| \Big(\mathcal{A}^2 - \Big(\frac{D^{\frac{1}{2}} \mathbb{1}}{\sqrt{M}} \Big) \Big(\frac{\mathbb{1}^T D^{\frac{1}{2}}}{\sqrt{M}} \Big) \Big) \frac{D^{\frac{1}{2}} f}{\sqrt{M}} \Big|\Big| \\ &\quad \text{(by Cauchy-Schwarz inequality)} \\ &\quad \leq \Big|\Big| \Big(\mathcal{A}^2 - \Big(\frac{D^{\frac{1}{2}} \mathbb{1}}{\sqrt{M}} \Big) \Big(\frac{\mathbb{1}^T D^{\frac{1}{2}}}{\sqrt{M}} \Big) \Big) \Big|\Big| \times \Big|\Big| \frac{D^{\frac{1}{2}} f}{\sqrt{M}} \Big|\Big|^2 \\ &\quad \text{(where, } ||Q|| \text{ denotes operator norm of a matrix } Q) \\ &\quad = \lambda_2^2 \mathbb{E}\{f(Y)\} \end{aligned}$$

and (14) follows. □

Theorem 4 gives insight into the network properties that affect the performance of the Algorithm 1. Eq. (13) states that, the bias of the estimate $T_{RW}^{|S|}$ is proportional to the covariance between the degree $d(X)$ and the label $f(X)$ of a random node X. Theorem 4 also shows that the variance of the estimate $T_{RW}^{|S|}$ is bounded above by a function of the second largest singular value λ_2 of the normalized adjacency matrix $\mathcal{A}$ and the expected label value of a random friend Y. Hence, a smaller λ_2 which indicates that the network has a good expansion[8] (i.e. absence of bottlenecks) [41] will result in a smaller variance in the estimate $T_{RW}^{|S|}$.

The following corollary gives a sufficient condition for the estimate $T_{RW}^{|S|}$ to be more statistically efficient (i.e. smaller MSE) compared to the classical intent polling method. Recall that the sampling budget $|S|$ denotes the number of nodes queried by the pollster.

Corollary 5. *If the sampling budget $|S|$ satisfies*

$$|S| \leq \frac{\big(\text{Var}\{f(X)\} - \lambda_2^2 \mathbb{E}\{f(Y)\} \big) \mathbb{E}\{d(X)\}^2}{\text{Cov}\{f(X)d(X)\}^2}, \tag{17}$$

then the estimate $T_{RW}^{|S|}$ obtained from Algorithm 1 has a smaller MSE compared to the intent polling estimate $I^{|S|}$ in (3), i.e. $\text{MSE}\{T_{RW}^{|S|}\} \leq \text{MSE}\{I^{|S|}\}$.

[8] A network is considered to have "good expansion" if every subset S of nodes ($S \leq 50\%$ of the nodes) has a neighborhood that is larger than some "expansion factor" multiplied by the number of nodes in S. Hence, a good expansion factor indicates that that there are no bottlenecks i.e. there is no small set of edges whose removal will fragment the network into two large connected components [41].

Proof. From (13) and (14) we get,

$$\begin{aligned}\mathrm{MSE}\{T_{RW}^{|S|}\} &= \mathrm{Bias}\{T_{RW}^{|S|}\}^2 + \mathrm{Var}\{T_{RW}^{|S|}\} \\ &\leq \left(\frac{\mathrm{Cov}\{f(X), d(X)\}}{\mathbb{E}\{d(X)\}}\right)^2 + \frac{\lambda_2^2 \mathbb{E}\{f(Y)\}}{|S|}. \end{aligned} \tag{18}$$

Also,

$$\mathrm{MSE}\{I^{|S|}\} = \mathrm{Var}\{I^{|S|}\} = \frac{\mathrm{Var}\{f(X)\}}{|S|}. \tag{19}$$

Hence, the result follows from (18) and (19). □

Corollary 5 indicates that a smaller degree-label correlation and the absence of bottlenecks result in the estimate $T_{RW}^{|S|}$ outperforming intent polling (3) for a larger range of sampling budgets $|S|$. This is because smaller label-degree correlation and the absence of bottlenecks make the bias and variance of $T_{RW}^{|S|}$ smaller according to Theorem 4 and therefore, makes the MSE of $T_{RW}^{|S|}$ smaller.

Next, we characterize bias and variance of the naive NEP estimate $T_{UN}^{|S|}$ (defined in (7)), thereby allowing us to compare it with friendship paradox based NEP methods (Algorithm 1 and Algorithm 2).

Theorem 6. *The bias* $\mathrm{Bias}\{T_{UN}^{|S|}\}$ *and the variance* $\mathrm{Var}\{T_{UN}^{|S|}\}$ *of the naive NEP estimate* $T_{UN}^{|S|}$ *(defined in (7)) are given by,*

$$\mathrm{Bias}(T_{UN}^{|S|}) = \mathbb{E}\{f(Z)\} - \mathbb{E}\{f(X)\} \tag{20}$$

$$\begin{aligned}\mathrm{Var}\{T_{UN}^{|S|}\} &= \frac{1}{|S|n} f^T D^{\frac{1}{2}} \mathcal{A} D^{-\frac{1}{2}} \left(I - \frac{\mathbb{1}\mathbb{1}^T}{n}\right) D^{-\frac{1}{2}} \mathcal{A} D^{\frac{1}{2}} f \\ &\leq \frac{1}{|S|} \frac{\mathbb{E}\{f(Y)\}\mathbb{E}\{d(X)\}}{d_{min}} \end{aligned} \tag{21}$$

where, n is the total number of nodes, X is a random node, Y is a random friend, Z is a random friend of a random node, $\mathcal{A}$ is the normalized adjacency matrix defined in (12) and f is a column vector with label $f(v) \in \{0, 1\}$ of node v at v^{th} element.

Proof. Note that,

$$\begin{aligned}\mathbb{E}\{q(X)\} &= \mathbb{E}\left\{\frac{\sum_{u \in \mathcal{N}(X)} f(u)}{d(X)}\right\} \\ &= \mathbb{E}\{\mathbb{E}\{f(Z)|X\}\} = \mathbb{E}\{f(Z)\},\end{aligned}$$

from which, (20) follows.

Next, recall that $q(v) = e_v^T D^{-1} A f$. Hence,

$$\begin{aligned}\mathbb{E}\{q(X)\} &= \sum_{v \in V} \frac{1}{n} e_v^T D^{-1} A f = \frac{1}{n} \mathbb{1}^T D^{-1} A f \text{ and,} \\ \mathbb{E}\{q^2(X)\} &= \sum_{v \in V} \frac{1}{n} f^T A D^{-1} e_v e_v^T D^{-1} A f \\ &= \frac{1}{n} f^T A D^{-2} A f\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Var}\{q(X)\} &= \mathbb{E}\{q^2(X)\} - \mathbb{E}\{q(X)\}^2 \\ &= \frac{1}{n} f^T A D^{-2} A f - \frac{1}{n^2} f^T A D^{-1} \mathbb{1}\mathbb{1}^T D^{-1} A f \\ &= \frac{1}{n} f^T D^{\frac{1}{2}} \Big(\Big(D^{-\frac{1}{2}} A D^{-\frac{1}{2}}\Big) D^{-1} \Big(D^{-\frac{1}{2}} A D^{-\frac{1}{2}}\Big) \\ &\qquad - \frac{1}{n} D^{-\frac{1}{2}} A D^{-1} \mathbb{1}\mathbb{1}^T D^{-1} A D^{-\frac{1}{2}} \Big) D^{\frac{1}{2}} f \\ &= \frac{1}{n} (f^T D^{\frac{1}{2}}) (\mathcal{A} D^{-\frac{1}{2}}) \left(I - \frac{\mathbb{1}\mathbb{1}^T}{n}\right) (D^{-\frac{1}{2}} \mathcal{A}) (D^{\frac{1}{2}} f)\end{aligned}$$

By the sub-multiplicative property of matrix norms,

$$\begin{aligned}&\left|\left| (\mathcal{A} D^{-\frac{1}{2}}) \left(I - \frac{\mathbb{1}\mathbb{1}^T}{n}\right) (D^{-\frac{1}{2}} \mathcal{A}) \right|\right| \\ &\qquad \leq ||\mathcal{A}||^2 ||D^{-\frac{1}{2}}||^2 \left|\left| I - \frac{\mathbb{1}\mathbb{1}^T}{n} \right|\right|^2 = \frac{1}{d_{min}} \end{aligned} \tag{22}$$

(where, $||Q||$ denotes operator norm of a matrix Q).

Therefore, by applying Cauchy-Schwarz inequality and then using (22), we get

$$\mathrm{Var}\{q(X)\} \leq \frac{||D^{\frac{1}{2}} f||^2}{n} \frac{1}{d_{min}} = \mathbb{E}\{f(Y)\} \frac{\mathbb{E}\{d(X)\}}{d_{min}},$$

and (21) follows. □

The following corollary is a consequence of Theorem 4 and Theorem 6. It compares the worst case performances of friendship paradox based NEP estimate $T_{RW}^{|S|}$ (obtained via Algorithm 1) and naive NEP estimate $T_{UN}^{|S|}$ (defined in (7)). The result shows how friendship paradox based sampling reduces variance of NEP methods.

Corollary 7. *The upper bound (14) for the variance of the estimate $T_{RW}^{|S|}$ (from Algorithm 1) and the upper bound (21) for the variance of the estimate $T_{UN}^{|S|}$ (naive NEP) satisfy,*

$$\frac{1}{|S|} \lambda_2^2 \mathbb{E}\{f(Y)\} \leq \frac{1}{|S|} \frac{\mathbb{E}\{f(Y)\}\mathbb{E}\{d(X)\}}{d_{min}}. \tag{23}$$

Proof. The proof follows by the fact that $0 \leq \lambda_2^2 < 1 \leq \frac{\mathbb{E}\{d(X)\}}{d_{min}}$. □

Finally, we characterize bias and variance of the estimate $T_{FN}^{|S|}$ obtained via Algorithm 2 which exploits the second version of the friendship paradox (Theorem 2).

Theorem 8. *The bias* $\mathrm{Bias}\{T_{FN}^{|S|}\}$ *and the variance* $\mathrm{Var}\{T_{FN}^{|S|}\}$ *of the estimate $T_{FN}^{|S|}$, obtained via Algorithm 2 satisfy,*

$$\begin{aligned}\mathrm{Bias}\{T_{FN}^{|S|}\}^2 &= \frac{1}{n} \mathbb{1}^T D^{-\frac{1}{2}} (\mathcal{A}^2 - I) D^{\frac{1}{2}} f \\ &\leq (\lambda_n^2 - 1)^2 \mathbb{E}\{f(Y)\} \frac{\mathbb{E}\{d(X)\}}{\bar{d}_{hm}} \end{aligned} \tag{24}$$

$$\mathrm{Var}\{T_{FN}^{|S|}\} = \frac{1}{|S|n} f^T A D_{hm}^{-\frac{1}{2}} \left(D^{-1} - \frac{D_{hm}^{-\frac{1}{2}} \mathbb{1}\mathbb{1}^T D_{hm}^{-\frac{1}{2}}}{n}\right) D_{hm}^{-\frac{1}{2}} A f \tag{25}$$

where, λ_n is the smallest singular value of the normalized adjacency matrix $\mathcal{A}$, $\bar{d}_{hm} = \mathbb{E}\{\frac{1}{d(X)}\}^{-1}$ is the harmonic mean degree of the graph and D_{hm} is a diagonal matrix with harmonic mean of the neighbor degrees of node $v \in V$ at the v^{th} element.

Proof. Note that $\mathbb{P}\{Z = v\} = \frac{1}{n} e_v^T A D^{-1} \mathbb{1}$ and recall that $q(v) = e_v^T D^{-1} A f$. Hence,

$$\begin{aligned}\mathbb{E}\{q(Z)\} &= \sum_{v \in V} \mathbb{P}\{Z = v\} e_v^T D^{-1} A f \\ &= \sum_{v \in V} \frac{1}{n} (\mathbb{1}^T D^{-1} A e_v)(e_v^T D^{-1} A f) \\ &= \frac{1}{n} \mathbb{1}^T D^{-1} A D^{-1} A f \end{aligned} \tag{26}$$

Following similar steps to the above, we get,

$$\begin{aligned}\mathbb{E}\{q^2(Z)\} &= \sum_{v \in V} \mathbb{P}\{Z = v\} f^T A D^{-1} e_v e_v^T D^{-1} A f \\ &= \frac{1}{n} f^T A D^{-1} \Big(\sum_{v \in V} e_v e_v^T A D^{-1} \mathbb{1} e_v^T \Big) D^{-1} A f \\ &= \frac{1}{n} f^T A D_{hm}^{-1} D^{-1} A f \quad \text{where,} \end{aligned} \tag{27}$$

D_{hm} is a diagonal matrix with harmonic mean of the neighbors of node $v \in V$ at v^{th} diagonal element i.e. $D_{hm}(v,v) = d(v) \Big(\sum_{u \in \mathcal{N}(v)} \frac{1}{d(u)} \Big)^{-1}$. Then, (25) follows from (26) and (27).

Next we prove (24).

$$\begin{aligned}\text{Bias}\{T_{FN}^{|S|}\} &= \mathbb{E}\{q(Z)\} - \mathbb{E}\{f(X)\} \\ &= \frac{1}{n} \mathbb{1}^T D^{-1} A D^{-1} A f - \frac{\mathbb{1}^T f}{n} \\ &= \frac{\mathbb{1}^T D^{-\frac{1}{2}}}{n} \Big(\mathcal{A}^2 - I \Big) D^{\frac{1}{2}} f \end{aligned}$$

Hence,

$$\begin{aligned}|\text{Bias}\{T_{FN}^{|S|}\}| &\leq \Big|\Big| \frac{\mathbb{1}^T D^{-\frac{1}{2}}}{n} \Big(\mathcal{A}^2 - I \Big) D^{\frac{1}{2}} f \Big|\Big| \\ &= \frac{1}{n} (\lambda_n^2 - 1) ||D^{\frac{1}{2}} f|| \times ||\mathbb{1}^T D^{-\frac{1}{2}}|| \end{aligned}$$

which implies,

$$\begin{aligned}\text{Bias}\{T_{FN}^{|S|}\}^2 &\leq \frac{M}{n} (\lambda_n^2 - 1)^2 \frac{||D^{\frac{1}{2}} f||^2}{M} \times \frac{\sum_{v \in V} \frac{1}{d(v)}}{n} \\ &= (\lambda_n^2 - 1)^2 \mathbb{E}\{f(Y)\} \times \frac{\mathbb{E}\{d(X)\}}{\bar{d}_{hm}} \end{aligned}$$

and (24) follows. □

Eq. (24) shows that the bias of the estimate $T_{FN}^{|S|}$ depends on the smallest singular value of the normalized adjacency matrix $\mathcal{A}$. This suggests that, the bias of the estimate $T_{FN}^{|S|}$ based on second version of friendship paradox depends on spectral properties of the network as opposed to the estimate $T_{RW}^{|S|}$ (obtained via Algorithm 1) based on the first version of the friendship paradox (Theorem 1).

Summary of Statistical Analysis: The above results (Theorem 3 to Theorem 8) motivate the use of NEP with friendship paradox based sampling (Algorithm 1 and Algorithm 2) compared to the intent polling and NEP without friendship paradox (i.e. naive NEP). Theorem 3 showed that the two friendship paradox based NEP algorithms have smaller MSE compared to the naive NEP method when labels are independently and identically distributed. Then, Theorem 4 characterized the bias and variance of the estimate $T_{RW}^{|S|}$ obtained via Algorithm 1 and Corollary 5 illustrated that it has a smaller MSE compared to intent polling for small sampling budget $|S|$ values. Further, Theorem 4 also showed that the bias and variance of the estimate $T_{RW}^{|S|}$ are affected by the degree-label correlation and the expansion of the network respectively. Next, Theorem 6 characterized the bias and variance of the naive NEP estimate $T_{UN}^{|S|}$ and Corollary 7 illustrated how NEP with friendship paradox outperforms naive NEP (without friendship paradox). Finally, Theorem 8 characterized the bias and variance of estimate $T_{FN}^{|S|}$ produced by the Algorithm 2 based on the second version of the friendship paradox (Theorem 2). It shows that the bias of estimate $T_{FN}^{|S|}$ depends on the spectral properties of the network as opposed the estimate $T_{RW}^{|S|}$ based on the first version of the friendship paradox.

5 Empirical and Simulation Results

The aim of this section is to evaluate Algorithm 1 and Algorithm 2 on five large scale real world social networks as well as synthetic network datasets in order to obtain insights that complement the analytical results presented in Sec. 4. More specifically,

1) Sec. 5.2 evaluates Algorithm 1, Algorithm 2, naive NEP and intent polling on four real world social networks with different degree-label correlation coefficients.
2) Sec. 5.2 evaluates Algorithm 1, Algorithm 2, naive NEP and intent polling on networks that are obtained from two well known models: configuration model [42] and Erdős-Rényi ($G(n,p)$) model [43].

The key conclusions that can be drawn from these experiments and simulations, and how they relate to the analytical results, are then discussed in detail in Sec. 6.

Before proceeding to present the results, we define three key variables that are widely used in social network analysis.

1) **Degree distribution** $P(k)$ is the probability that a randomly chosen node has k neighbors.

2) **Neighbor degree correlation (assortativity) coefficient** is defined as,

$$r_{kk} = \frac{1}{\sigma_q^2} \sum_{k,k'} kk' \Big(e(k,k') - q(k) q(k') \Big) \tag{28}$$

where, $e(k,k')$ is the probability of nodes at the ends of a randomly chosen edge have degrees k and k' (joint degree distribution of neighbors), $q(k)$ is the probability that a random neighbor has k neighbors (marginal distribution of $e(k,k')$) and σ_q is the standard deviation with respect to q.

3) **Degree-label correlation coefficient** is defined as,

$$\begin{aligned}\rho_{kf} = \frac{1}{\sigma_k \sigma_f} \sum_k k \Big(&\mathbb{P}(f(X) = 1, d(X) = k) \\ &- \mathbb{P}(f(X) = 1) P(k) \Big) \end{aligned} \tag{29}$$

where, σ_k and σ_f are the standard deviations of the degree of a random node and the label of a random node respectively.

A detailed discussion of these variables and their effects can be found in [20].

5.1 Real World Networks

Dataset Description: The datasets used in this subsection are openly available from the Stanford Network Analysis Project (SNAP) [44]. Below, we describe each dataset briefly.

1) *Facebook Social Circles* [45]: This dataset consists of "circles" (or "friends lists") from Facebook that were collected using the Facebook App. Total number of nodes and edges in the network constructed from this dataset are 4039 and 88234 respectively. The neighbor degree correlation coefficient r_{kk} (defined in (28)) of the network is 0.06
2) *Co-authorship Network* [46]: This dataset contains the scientific collaborations between authors of papers submitted to General Relativity and Quantum Cosmology category in the Arxiv website. More specifically, an author i co-authoring a paper with author j will be represented by an undirected edge between the two nodes i and j in the network. Total number of nodes and edges in the network constructed from this dataset are 5242 and 14496 respectively. The neighbor degree correlation coefficient r_{kk} (defined in (28)) of the network is 0.66.
3) *Athlete Network* [47]: This dataset contains Facebook page networks of athletes. The nodes in the network represent the Facebook pages of athletes and the edges represent mutual likes among them. Total number of nodes and edges in the network constructed from this dataset are $13,866$ and $86,858$ respectively. The neighbor degree correlation coefficient r_{kk} (defined in (28)) of the network is -0.03.
4) *Politician Network* [47]: This dataset contains Facebook page networks of politicians. The nodes in the network represent the Facebook pages of politicians and the edges represent mutual likes among them. Total number of nodes and edges in the network constructed from this dataset are 5908 and 41729 respectively. The neighbor degree correlation coefficient r_{kk} (defined in (28)) of the network is 0.02.
5) *Company Network* [47]: This dataset contains Facebook page networks of different companies. The nodes in the network represent the Facebook pages of companies and the edges represent mutual likes among them. Total number of nodes and edges in the network constructed from this dataset are $14,113$ and $52,310$ respectively. The neighbor degree correlation coefficient r_{kk} (defined in (28)) of the network is 0.01.

Label swapping procedure for modifying degree-label correlation: Given a graph $G = (V, E)$, we first assign labels $f(v)$ to each node $v \in V$ with a fixed probability. Then, to set the degree-label correlation coefficient defined in (29) to a desired value, we utilize the label swapping procedure followed in [20]: a node v_0 with a label $f(v_0) = 0$ and a node v_1 with a label $f(v_1) = 1$ are selected randomly and their labels are swapped if $d(v_0) < d(v_1)$ (respectively, $d(v_0) > d(v_1)$) to increase (respectively, decrease) the degree-label correlation coefficient ρ_{kf} to the desired value (or until it no longer changes). We consider $\rho_{kf} = -0.1, 0, 0.1$ in our experiments to study the effect of negative and positive degree-label correlations on the accuracy of the polling algorithms.

Empirical Results: The MSE and variance of the four polling methods (Algorithm 1, Algorithm 2, intent polling and naive NEP) were estimated using Monte-Carlo simulation over 600 independent iterations for each value of the sampling budget $|S|$ from 1 to approximately 1% of the total number of nodes in the network. The results are displayed in Fig. 2. The conclusions and insights that can be drawn from these empirical results and how they relate to the analytical results are discussed in Sec. 6.

5.2 Numerical Examples

Generative Models for Graphs: We use the following two generative models to yield two different types of degree distributions: power-law degree distribution and exponential degree distribution. In all experiments below, we consider graphs with $n = 5000$ nodes.

- **Configuration Model [42]:** Generate k half-edges for each of the n nodes where $k \sim ck^{-\alpha}$ (where c is a normalizing constant) and then, connect each half-edge to the another randomly selected half-edge avoiding self loops. This model yields a power-law degree distribution[9] $p(k) = ck^{-\alpha}$. We consider two cases: $\alpha = 2.1$ and $\alpha = 2.4$.
- **Erdős-Rényi (G(n,p)) model [43]:** Any two (distinct) nodes are connected by an edge with probability p. This model results in a Binomial degree distribution which can be approximated by a Poisson distribution for large n. We choose $p = 0.01, n = 5000$ to ensure that the graph has no isolated nodes with high probability.

Newman's edge-rewiring procedure for modifying neighbor degree correlation: We utilize the edge-rewiring procedure proposed in [51] to change the assortativity coefficient r_{kk} (28) of the graphs generated using the above models to a desired value while preserving the degree distribution. In the edge-rewiring procedure, two uniformly chosen links $(v_1, v_2), (u_1, u_2) \in E$ at each iteration are replaced with new links $(v_1, u_1), (v_2, u_2)$ if it increases (respectively, decreases) the value of the assortativity coefficient r_{kk}. The process is repeated until the desired value of the assortativity coefficient r_{kk} is achieved (or until it no longer changes).

Simulation Results: The four polling methods (Algorithm 1, Algorithm 2, intent polling (3) and naive NEP (7)) were evaluated on the networks obtained using the simulation setup described above. The MSE of the polling methods were estimated using Monte-Carlo simulation over 600 independent iterations. The resulting empirical MSE values for the configuration model (power-law degree distribution) are shown in Fig. 3 and Fig. 4 for power-law coefficient values $\alpha = 2.4$ and $\alpha = 3.1$ respectively. Similarly, results obtained for Erdős-Rényi graphs (Poisson degree distribu-

[9]The power-law degree distribution is generally accepted as a key feature of many real world networks such as World Wide Web, Internet and social networks [37], [43], [48], [49] with a power-law exponent $2 < \alpha < 3$ [50]. Further, it has been shown that friendship paradox and some of its effects are amplified in the presence of such power-law degree distributions [15], [20].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2940914, IEEE Transactions on Knowledge and Data Engineering

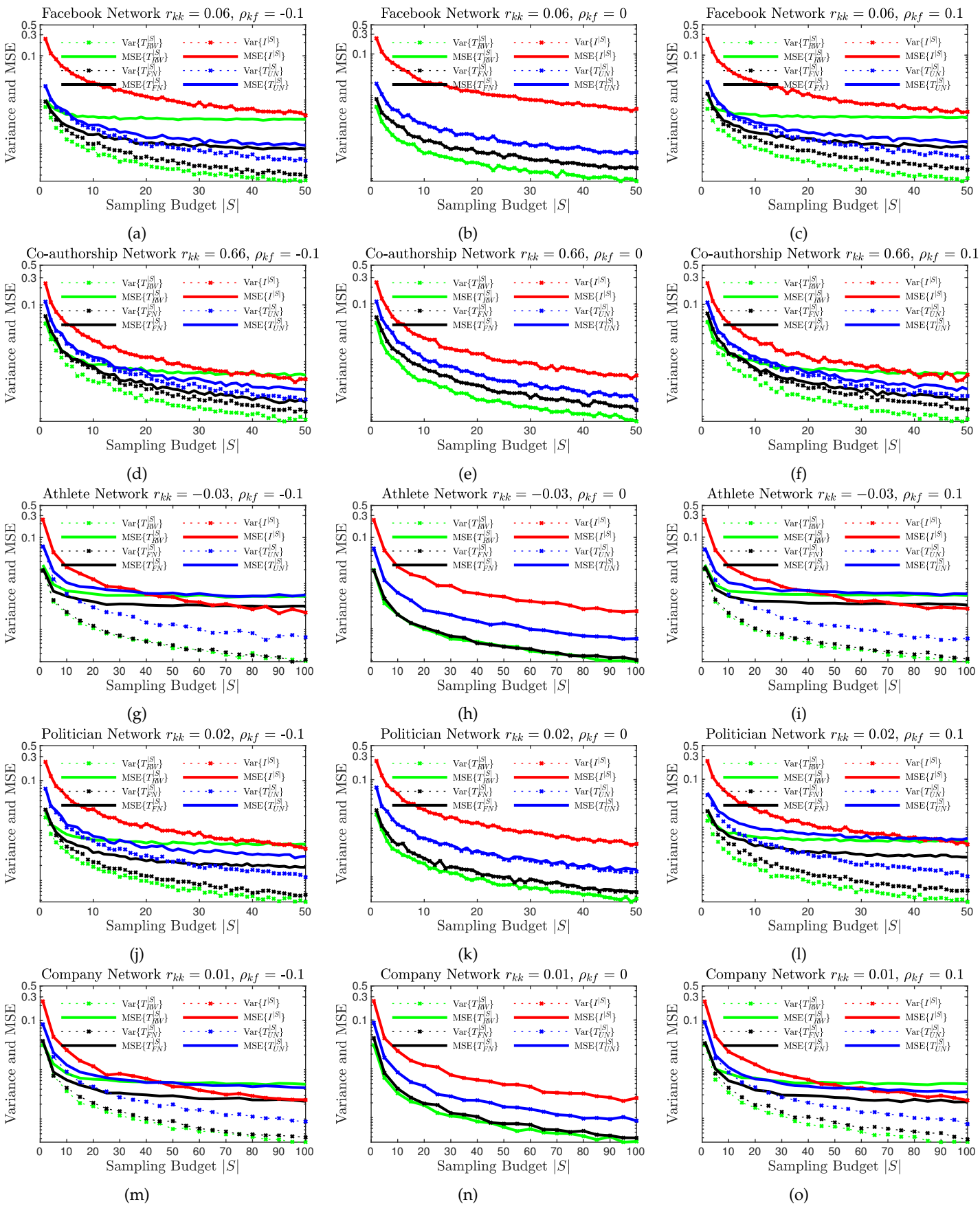


Fig. 2: Empirical MSE and Variance of estimates $T_{RW}^{|S|}$ (Algorithm 1), $T_{FN}^{|S|}$ (Algorithm 2), $I^{|S|}$(intent polling) and $T_{UN}^{|S|}$ (naive NEP) on five real world datasets (described in Sec. 5.1). The subplots show that friendship paradox based NEP methods (Algorithm 1 and Algorithm 2)) are more statistically efficient compared to intent polling and naive NEP and, achieves a bias-variance trade-off based on the length of the random walk.

tion)[10] are shown in Fig. 5. The conclusions and insights that can be drawn from these simulation results and how they relate to the analytical results are discussed in Sec. 6.

6 Discussion of Empirical and Simulation Results

This section discusses the insights and conclusions that can be drawn from the empirical and simulation results (Sec. 5) and, how they relate to the analytical results (Sec. 4). The main aim is to highlight how the analytical and experimental results help identify the contexts for which each polling algorithm is suitable.

6.1 Power-law Graphs

Intent Polling vs. Friendship Paradox Based NEP: Corollary 5 stated that the friendship paradox based NEP Algorithm 1 outperforms the classical intent polling in terms of the mean-squared error for small sampling budget $|S|$ values. The empirical results (Fig. 2) are consistent with Corollary 5; it can be seen that the MSE of the intent polling estimate $I^{|S|}$ is larger than the MSE of the estimates $T_{RW}^{|S|}, T_{FN}^{|S|}$ obtained via the friendship paradox based NEP methods for smaller (less than 50) sampling budget $|S|$ values. Further, MSE of estimates $T_{RW}^{|S|}, T_{FN}^{|S|}$ are smaller for all considered sampling budget $|S|$ values when the degree-label correlation coefficient ρ_{kf} is zero (and hence, the friendship paradox based polling produces an unbiased estimate according to Theorem 4). Hence, both analytical and empirical results indicate that friendship paradox based NEP methods outperform the classical intent polling method when the sampling budget $|S|$ is constrained to be smaller or, the node labels are uncorrelated with the node degrees ($\rho_{kf} = 0$).

Effect of degree-label correlation (ρ_{kf}): Fig. 2 shows that the friendship paradox based polling Algorithms 1 and 2 outperform both intent polling and naive NEP (7) for all considered sampling budget $|S|$ values when the node labels and node degrees are uncorrelated ($\rho_{kf} = 0$). When the node degree and node labels are correlated ($\rho_{kf} \neq 0$), Algorithm 2 still outperforms (in terms of MSE) the both intent polling and naive NEP methods for all considered sampling budget $|S|$ values whereas naive NEP method outperforms Algorithm 1 when $|S|$ becomes large due to the bias variance trade-off that is discussed next.

Friendship paradox based bias variance trade-off optimization: Note that the naive NEP estimate $T_{UN}^{|S|}$, NEP estimate $T_{FN}^{|S|}$ based on version 2 of friendship paradox (Theorem 2) and NEP estimate $T_{RW}^{|S|}$ based on version 1 of friendship paradox (Theorem 1) correspond to random walks of length $N = 0$ ($T_{UN}^{|S|}$), $N = 1$ ($T_{FN}^{|S|}$) and $N \to \infty$ ($T_{RW}^{|S|}$). As such, $T_{UN}^{|S|}$ is based on responses of individuals sampled independent of their degree, $T_{RW}^{|S|}$ is based on responses of individuals sampled with probabilities proportional to their degrees and $T_{FN}^{|S|}$ achieves a trade-off by taking only a single step random walk. Therefore, it is intuitive that the variance of the estimates should satisfy $\mathrm{Var}\{T_{RW}^{|S|}\} \leq \mathrm{Var}\{T_{FN}^{|S|}\} \leq \mathrm{Var}\{T_{UN}^{|S|}\}$, agreeing with Corollary 7 and the empirical variances plotted in Fig. 2. However, in terms of the mean-squared error (which takes the bias of the estimates into account), $T_{FN}^{|S|}$ outperforms both $T_{UN}^{|S|}, T_{RW}^{|S|}$ (in terms of MSE) for all $|S|$ values considered in the empirical results. This observation suggests that the length of random walk (e.g. $N = 1$) can be used to control the bias-variance trade-off of the friendship paradox based NEP methods. For example, if it is apriori known to the pollster that the labels have negligible correlation with the degrees (i.e. $\rho_{kf} \approx 0$ and hence, the bias of both $T_{RW}^{|S|}, T_{FN}^{|S|}$ will be negligible), she can choose to use $T_{RW}^{|S|}$ to minimize the variance of the estimate.

Effect of the heavy-tails: Comparing Fig. 3 with Fig. 4 shows that the MSE of Algorithm 1 and Algorithm 2 are smaller in the network with power-law coefficient $\alpha = 2.1$ compared to that with $\alpha = 2.4$. The difference in MSE is more pronounced for Algorithm 2 compared to Algorithm 1. This suggests that friendship paradox based algorithms are more suitable when the underlying network has a heavy tailed degree distribution.

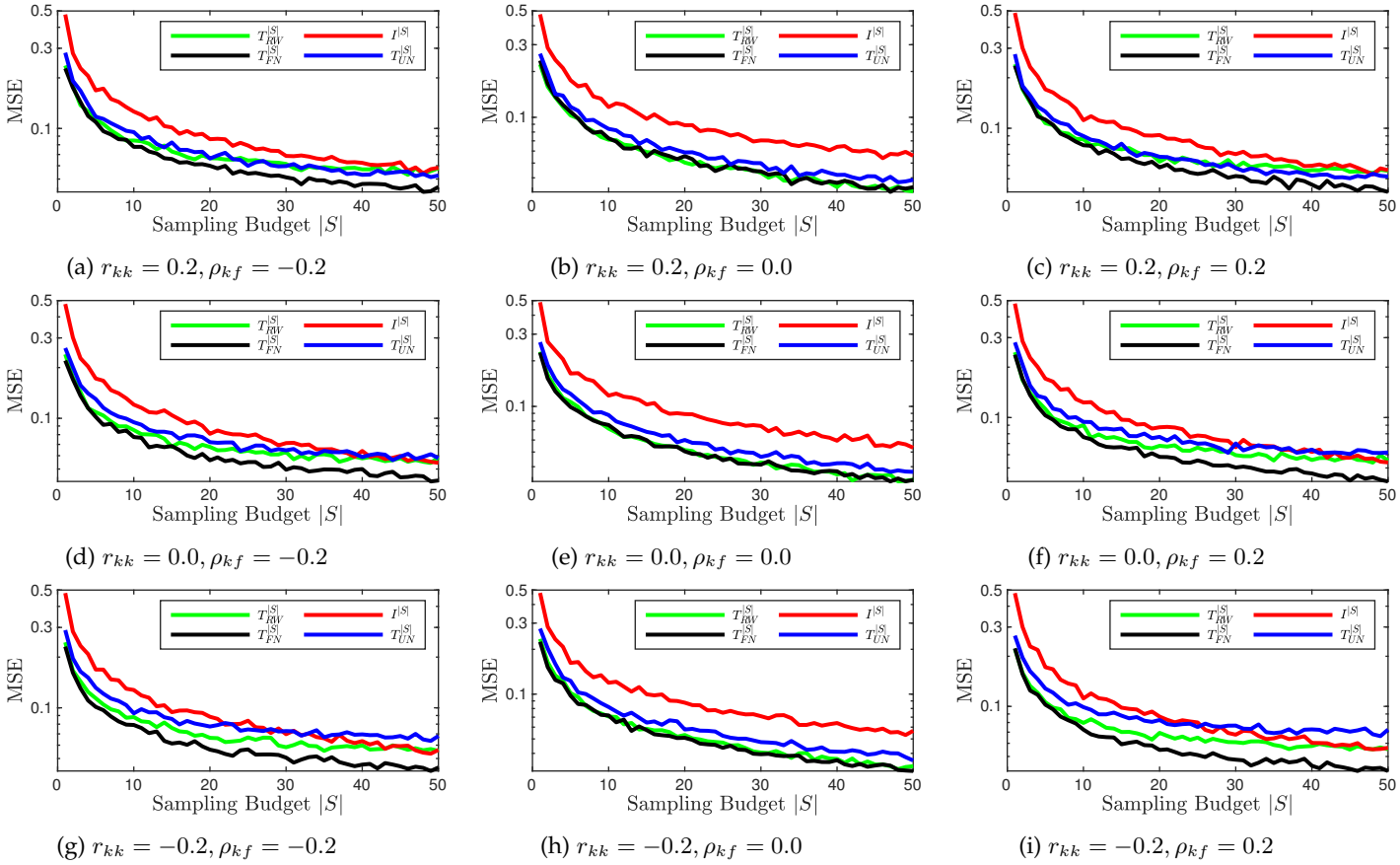
Effect of the Assortativity of the Network: Different joint degree distributions $e(k, k')$ can yield the same neighbor degree distribution $q(k)$ (explained in Sec. 5). Naturally, this marginal distribution $q(k)$ does not capture the joint variation of the degrees a random pair of neighbors. In Algorithm 1 (which samples neighbors uniformly), the degree distribution of the samples (i.e. queried nodes) is the neighbor degree distribution $q(k)$. Hence, the performance is not affected by the assortativity coefficient r_{kk}, which captures the joint variation of the degrees of a random pair of neighbors. This is seen in Fig. 3 where, each column (corresponding to different r_{kk} values) has approximately same MSE for Algorithm 1. However, the MSE of Algorithm 2 (that samples random friends Z of random nodes) increases with assortativity r_{kk} due to the fact that the distribution of degree $d(Z)$ of a random friend Z of a random node is a function of the joint degree distribution. In order to highlight this further, Fig. 6 illustrates the effect of the neighbor degree correlation r_{kk} on the distribution of $d(Z)$ (and the invariance of the distribution of $d(Y)$ to r_{kk}). This result indicates that, if the network is disassortative ($r_{kk} < 0$), Algorithm 2 is a more suitable choice for polling compared to Algorithm 1.

6.2 Erdős-Rényi Graphs

The Erdős-Rényi ($G(n, p)$) model constructs a random graph as follows: start with n vertices and then connect any two vertices with probability p. Therefore, the average degree of the resulting graph is $(n-1)p$. From Fig. 5, it can be seen that both Algorithm 1 and Algorithm 2 yield a smaller MSE than the intent polling method for the Erdős-Rényi network with $p = 0.01$ and $n = 5000$. Also, Algorithm 1 and Algorithm 2 have approximately equal MSE. This is due to the fact that in an Erdős-Rényi network, the neighbor degree correlation is approximately zero and therefore, distributions of the degree $d(Y)$ of a random neighbor Y and the distribution of

[10] In the case of Erdős-Rényi graphs, we only consider assortativity coefficient $r_{kk} = 0$ since it cannot be changed significantly due to the homogeneity in the degree distribution.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2940914, IEEE Transactions on Knowledge and Data Engineering



(a) $r_{kk} = 0.2, \rho_{kf} = -0.2$ (b) $r_{kk} = 0.2, \rho_{kf} = 0.0$ (c) $r_{kk} = 0.2, \rho_{kf} = 0.2$

(d) $r_{kk} = 0.0, \rho_{kf} = -0.2$ (e) $r_{kk} = 0.0, \rho_{kf} = 0.0$ (f) $r_{kk} = 0.0, \rho_{kf} = 0.2$

(g) $r_{kk} = -0.2, \rho_{kf} = -0.2$ (h) $r_{kk} = -0.2, \rho_{kf} = 0.0$ (i) $r_{kk} = -0.2, \rho_{kf} = 0.2$

Fig. 3: MSE of the estimates obtained using the four polling algorithms for a power-law graph with parameter $\alpha = 2.4$ and different values of assortativity coefficient r_{kk} and degree-label correlation coefficient ρ_{kf}. Subplots show that, for power-law networks, proposed polling methods have smaller MSE compared to alternative methods under general conditions.

the degree $d(Z)$ of a random neighbor Z of a random node are approximately equal.

7 Conclusion

This paper considered the problem of estimating the fraction of nodes in a graph that has a particular attribute (represented by a binary label) and, proposed a novel class of polling methods called Neighborhood Expectation Polling (NEP). In NEP, each sampled individual responds with information about the fraction of her neighbors in the social network that has label 1. We considered the cases where either: 1) the pollster has no knowledge about the social graph but, has the ability to perform random walks on the graph 2) uniformly sampled nodes from the unknown social graph are available. Two NEP algorithms were proposed (for case 1 and case 2) exploiting a form of network bias called friendship paradox. Theorems 3 to 8 characterized the bias, variance and mean-squared error of the estimate as well as how they depend on the properties of the underlying network (correlation between node labels and degree, expansion, average, minimum and maximum degree, etc.) were derived. These results are useful for a pollster to incorporate prior knowledge about the underlying network to choose the best algorithm (in terms of statistical efficiency) and guarantee its performance. Extensive empirical and simulation results are provided to illustrate the performance of the proposed methods under different network properties. These complement the theoretical analysis and provide insights into how the proposed algorithms would perform under different conditions. Both theoretical and experimental results indicate that the friendship paradox based NEP algorithms are capable of obtaining an estimate with a smaller mean-squared error with only a smaller (compared to alternative methods) number of respondents.

Acknowledgments

The authors thank Jon Kleinberg at Department of Computer Science of Cornell University for helpful suggestions.

References

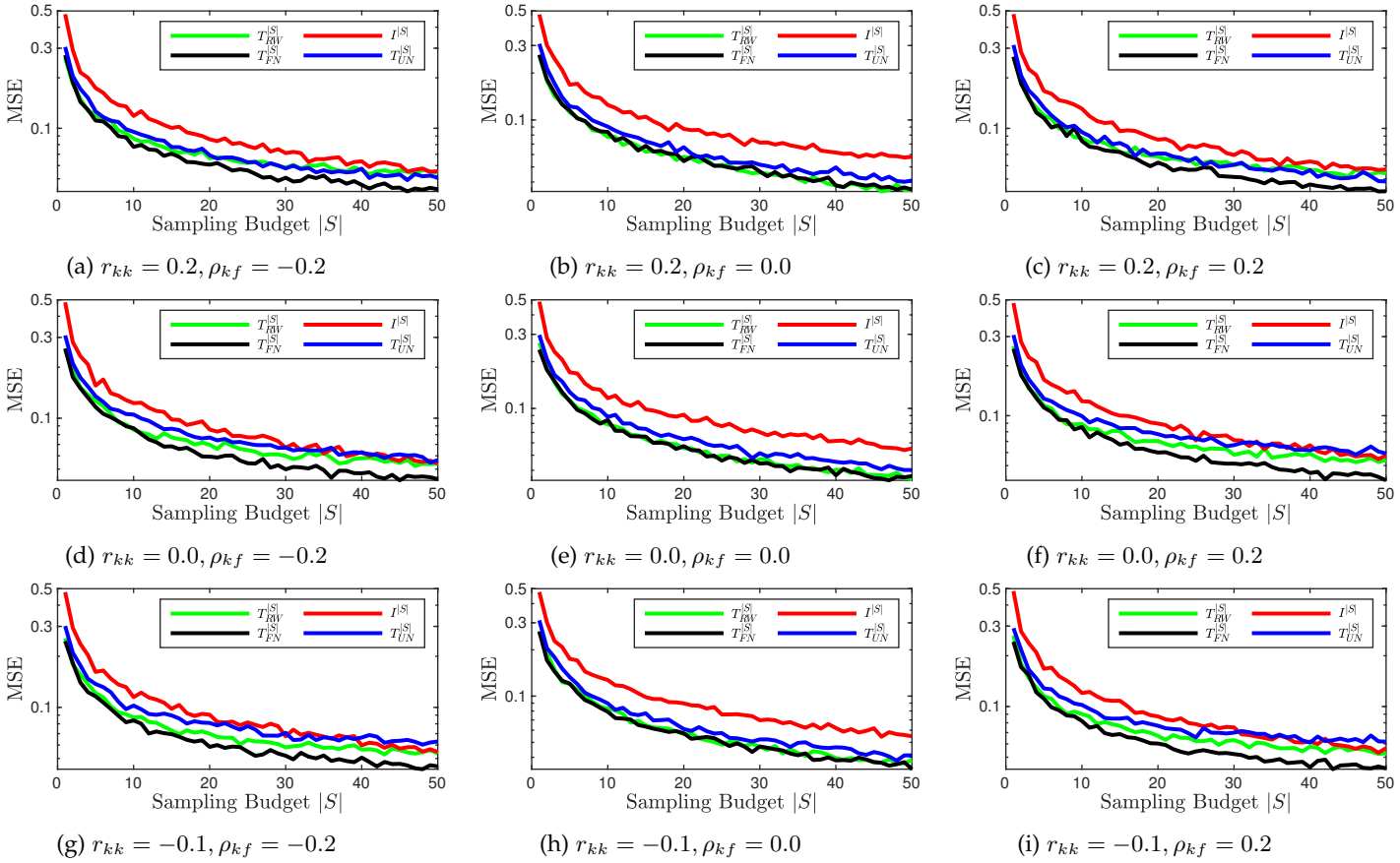
[1] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment." *ICWSM*, vol. 10, no. 1, pp. 178–185, 2010.

[2] K. J. Gile, "Improved inference for respondent-driven sampling data with application to HIV prevalence estimation," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 135–146, 2011.

[3] A. Dasgupta, R. Kumar, and D. Sivakumar, "Social sampling," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 235–243.

[4] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2940914, IEEE Transactions on Knowledge and Data Engineering



(a) $r_{kk} = 0.2, \rho_{kf} = -0.2$ (b) $r_{kk} = 0.2, \rho_{kf} = 0.0$ (c) $r_{kk} = 0.2, \rho_{kf} = 0.2$

(d) $r_{kk} = 0.0, \rho_{kf} = -0.2$ (e) $r_{kk} = 0.0, \rho_{kf} = 0.0$ (f) $r_{kk} = 0.0, \rho_{kf} = 0.2$

(g) $r_{kk} = -0.1, \rho_{kf} = -0.2$ (h) $r_{kk} = -0.1, \rho_{kf} = 0.0$ (i) $r_{kk} = -0.1, \rho_{kf} = 0.2$

Fig. 4: MSE of the estimates obtained using the four polling algorithms for a power-law graph with parameter $\alpha = 3.1$ and different values of assortativity coefficient r_{kk} and degree-label correlation coefficient ρ_{kf}. Subplots show that, for power-law networks, proposed polling methods have smaller MSE compared to alternative methods under general conditions.

[5] S. L. Feld, "Why your friends have more friends than you do," *American Journal of Sociology*, vol. 96, no. 6, pp. 1464–1477, 1991.

[6] D. M. Rothschild and J. Wolfers, "Forecasting elections: Voter intentions versus expectations," 2011.

[7] A. Graefe, "Accuracy gains of adding vote expectation surveys to a combined forecast of us presidential election outcomes," *Research & Politics*, vol. 2, no. 1, p. 2053168015570416, 2015.

[8] A. E. Murr, "The wisdom of crowds: Applying Condorcet's jury theorem to forecasting us presidential elections," *International Journal of Forecasting*, vol. 31, no. 3, pp. 916–929, 2015.

[9] A. Graefe, "Accuracy of vote expectation surveys in forecasting elections," *Public Opinion Quarterly*, vol. 78, no. S1, pp. 204–232, 2014.

[10] A. E. Murr, ""Wisdom of crowds"? a decentralised election forecasting model that uses citizens local expectations," *Electoral Studies*, vol. 30, no. 4, pp. 771–783, 2011.

[11] C. F. Manski, "Measuring expectations," *Econometrica*, vol. 72, no. 5, pp. 1329–1376, 2004.

[12] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

[13] V. Krishnamurthy and W. Hoiles, "Online reputation and polling systems: Data incest, social learning, and revealed preferences," *IEEE Transactions on Computational Social Systems*, vol. 1, no. 3, pp. 164–179, 2014.

[14] B. Nettasinghe and V. Krishnamurthy, "Maximum likelihood estimation of power-law degree distributions using friendship paradox based sampling," *arXiv preprint arXiv:1908.00310*, 2019.

[15] Y.-H. Eom and H.-H. Jo, "Tail-scope: Using friends to estimate heavy tails of degree distributions in large-scale complex networks," *Scientific reports*, vol. 5, p. 09752, 2015.

[16] M. Garcia-Herranz, E. Moro, M. Cebrian, N. A. Christakis, and J. H. Fowler, "Using friends as sensors to detect global-scale contagious outbreaks," *PloS one*, vol. 9, no. 4, p. e92413, 2014.

[17] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PloS one*, vol. 5, no. 9, p. e12948, 2010.

[18] N. Alipourfard, B. Nettasinghe, A. Abeliuk, V. Krishnamurthy, and K. Lerman, "Friendship paradox biases perceptions in directed networks," *arXiv preprint arXiv:1905.05286*, 2019.

[19] M. O. Jackson, "The friendship paradox and systematic biases in perceptions and social norms," *Journal of Political Economy*, vol. 127, no. 2, pp. 777–818, 2019.

[20] K. Lerman, X. Yan, and X.-Z. Wu, "The "majority illusion" in social networks," *PloS one*, vol. 11, no. 2, p. e0147617, 2016.

[21] B. Nettasinghe, V. Krishnamurthy, and K. Lerman, "Diffusion in social networks: Effects of monophilic contagion, friendship paradox and reactive networks," *IEEE Transactions on Network Science and Engineering*, 2019.

[22] V. Krishnamurthy and B. Nettasinghe, "Information diffusion in social networks: friendship paradox based models and statistical inference," in *Modeling, Stochastic Control, Optimization, and Applications*, ser. The IMA Volumes in Mathematics and its Applications, 2019, vol. 164, pp. 369–406.

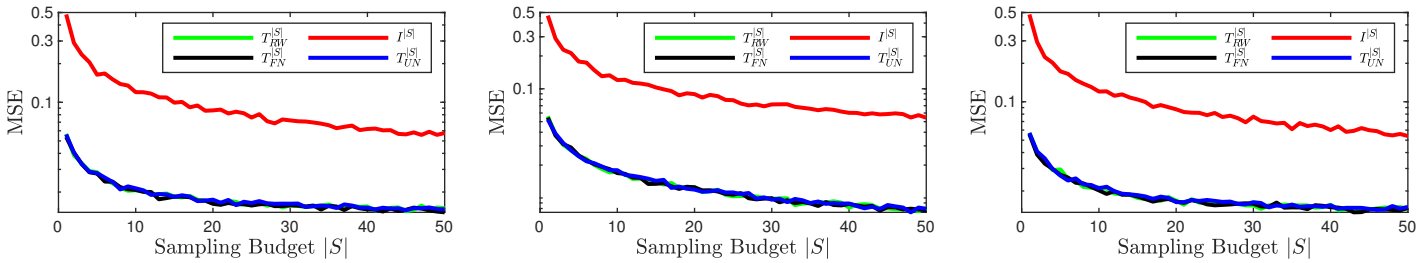
[23] E. Lee, S. Lee, Y.-H. Eom, P. Holme, and H.-H. Jo, "Impact of perception models on friendship paradox and opinion formation," *Physical Review E*, vol. 99, no. 5, p. 052302, 2019.

[24] J. P. Bagrow, C. M. Danforth, and L. Mitchell, "Which friends are more popular than you?: Contact strength and the friendship paradox in social networks," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM, 2017, pp. 103–108.

[25] A. Chin, D. Eckles, and J. Ugander, "Evaluating stochastic seeding strategies in networks," *arXiv preprint arXiv:1809.09561*, 2018.

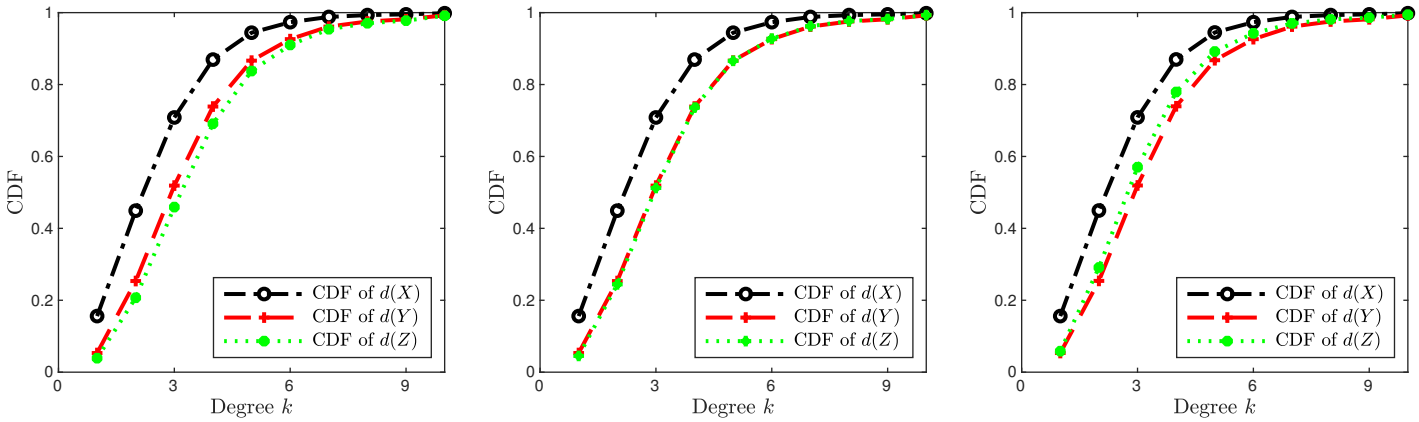
[26] V. Kumar, D. Krackhardt, and S. Feld, "Network interventions based on inversity: Leveraging the friendship paradox in unknown network structures," Yale University, Tech. Rep., 2018.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2940914, IEEE Transactions on Knowledge and Data Engineering



(a) $r_{kk} = 0.0, \rho_{kf} = -0.2$ (b) $r_{kk} = 0.0, \rho_{kf} = 0.0$ (c) $r_{kk} = 0.0, \rho_{kf} = 0.2$

Fig. 5: MSE of the estimates obtained using the four polling algorithms for a Erdős-Rényi (ER) graph with average degree 50 with assortativity coefficient $r_{kk} = 0$ and different values of degree-label correlation coefficient ρ_{kf}. The main conclusion is that, for ER graphs, the proposed friendship paradox based NEP method as well as the greedy deterministic sample selection method result in better performance compared to the intent polling method.



(a) $r_{kk} = -0.2$ (disassortative network) (b) $r_{kk} = 0.0$ (c) $r_{kk} = 0.2$ (assortative network)

Fig. 6: The cumulative distribution functions (CDF) of the degrees $d(X), d(Y), d(Z)$ of a random node (X), a random friend (Y) and a random friend of a random node (Z) respectively, for three graphs with the same degree distribution (power-law distribution with a coefficient $\alpha = 2.4$) but different neighbor-degree correlation coefficients r_{kk}, generated using the Newman's edge rewiring procedure. This illustrates that $\mathbb{E}\{d(Z)\} \geq \mathbb{E}\{d(Y)\}$ for $r_{kk} \leq 0$ (Fig. 6a) and vice-versa. This figure also shows how the distributions of $d(X), d(Y)$ remain invariant to the changes in the joint degree distribution $e(k, k')$ that preserve the degree distribution $P(k)$.

[27] S. Lattanzi and Y. Singer, "The power of random neighbors in social networks," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 77–86.

[28] D. J. Higham, "Centrality-friendship paradoxes: when our friends are more important than us," *Journal of Complex Networks*, vol. 7, no. 4, pp. 515–528, 11 2018.

[29] Y.-H. Eom and H.-H. Jo, "Generalized friendship paradox in complex networks: The case of scientific collaboration," *Scientific Reports*, vol. 4, Apr. 2014.

[30] N. Momeni and M. Rabbat, "Qualities and inequalities in online social networks through the lens of the generalized friendship paradox," *PloS one*, vol. 11, no. 2, p. e0143633, 2016.

[31] N. O. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you," in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[32] Y. Cao and S. M. Ross, "The friendship paradox." *Mathematical Scientist*, vol. 41, no. 1, 2016.

[33] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 631–636.

[34] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in Facebook: A case study of unbiased sampling of OSNs," in *Infocom, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[35] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. ACM, 2010, pp. 390–403.

[36] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical recommendations on crawling online social networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1872–1892, 2011.

[37] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM, 2007, pp. 29–42.

[38] D. Aldous and J. Fill, "Reversible Markov chains and random walks on graphs," 2002.

[39] J. B. Kramer, J. Cutler, and A. Radcliffe, "The multistep friendship paradox," *The American Mathematical Monthly*, vol. 123, no. 9, pp. 900–908, 2016.

[40] R. Durrett, *Probability: Theory and Examples*, 4th ed., ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[41] E. Estrada, "Network robustness to targeted attacks. the interplay of expansibility and degree distribution," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 52, no. 4, pp. 563–574, 2006.

[42] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random structures & algorithms*, vol. 6, no. 2-3, pp. 161–180, 1995.

[43] M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph

models of social networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. suppl 1, pp. 2566–2572, 2002.
[44] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.
[45] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.
[46] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 2, 2007.
[47] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "Gemsec: Graph embedding with self clustering," *arXiv preprint arXiv:1802.03997*, 2018.
[48] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.
[49] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," *Physical review E*, vol. 64, no. 4, p. 046135, 2001.
[50] M. Boguá, R. Pastor-Satorras, and A. Vespignani, "Epidemic spreading in complex networks with degree correlations," in *Statistical mechanics of complex networks*. Springer, 2003, pp. 127–147.
[51] M. E. Newman, "Assortative mixing in networks," *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.

PLACE PHOTO HERE

Buddhika Nettasinghe received the M.A.Sc. degree from the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, Canada in 2016. He is currently a Ph.D. student with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA. His current research interests include statistical inference and learning, network science, computational social science and complex systems.

PLACE PHOTO HERE

Vikram Krishnamurthy (F'05) received the Ph.D. degree from the Australian National University, Canberra, ACT, Australia, in 1992. He is currently a Professor in the School of Electrical and Computer Engineering, Cornell University and Cornell Tech, New York, NY, USA. From 2002 to 2016, he was the Canada Research Chair Professor in statistical signal processing at the University of British Columbia, Vancouver, BC, Canada. His current research interests include statistical signal processing and stochastic control with applications in social networks. He served as a Distinguished Lecturer for the IEEE signal processing society and Editor-in-Chief of IEEE Journal Selected Topics in Signal Processing. He received an honorary doctorate from KTH (Royal Institute of Technology), Stockholm, Sweden in 2013.