Estimating Exposure to Information on Social Networks

BUDDHIKA NETTASINGHE, Tippie College of Business, The University of Iowa, Iowa City, United States
KOWE KADOMA, Cornell University, Ithaca, United States
MOR NAAMAN, Cornell Tech, New York, United States
VIKRAM KRISHNAMURTHY, Cornell University, Ithaca, United States

Estimating exposure to information on a social network is a problem with important consequences for our society. The exposure estimation problem involves finding the fraction of people on the network who have been exposed to a piece of information (e.g., a URL of a news article on Facebook, a hashtag on Twitter). The exact value of exposure to a piece of information is determined by two features: the structure of the underlying social network and the set of people who shared the piece of information. Often, both features are not publicly available (i.e., access to the two features is limited only to the internal administrators of the platform) and are difficult to estimate from data. As a solution, we propose two methods to estimate the exposure to a piece of information in an unbiased manner: a vanilla method that is based on sampling the network uniformly and a method that non-uniformly samples the network motivated by the Friendship Paradox. We provide theoretical results that characterize the conditions (in terms of properties of the network and the piece of information) under which one method outperforms the other. Further, we outline extensions of the proposed methods to dynamic information cascades (where the exposure needs to be tracked in real time). We demonstrate the practical feasibility of the proposed methods via experiments on multiple synthetic and real-world datasets.

CCS Concepts: • **Information systems** → **Social networks**; **Information retrieval;**

Additional Key Words and Phrases: Social networks, exposure to information, friendship paradox, information diffusion, information cascades

ACM Reference Format:
Buddhika Nettasinghe, Kowe Kadoma, Mor Naaman, and Vikram Krishnamurthy. 2024. Estimating Exposure to Information on Social Networks. *ACM Trans. Soc. Comput.* 7, 1-4, Article 8 (September 2024), 24 pages. https://doi.org/10.1145/3688599

1 Introduction

Online social networks are an important mechanism through which people are exposed to information. Estimating the total number of people who are exposed by their friends to a piece of

B. Nettasinghe and V. Krishnamurthy were funded by National Science Foundation Grant No. CCF-2112457 and Army Research Office Grant No. W911NF-24-1-0083.
Authors' Contact Information: Buddhika Nettasinghe, Tippie College of Business, The University of Iowa, Iowa City, Iowa, United States; e-mail: buddhika-nettasinghe@uiowa.edu; Kowe Kadoma, Cornell University, Ithaca, New York, United States; e-mail: kk696@cornell.edu; Mor Naaman, Cornell Tech, New York, New York, United States; e-mail: mor.naaman@cornell.edu; Vikram Krishnamurthy, Cornell University, Ithaca, New York, United States; e-mail: vikramk@cornell.edu.

This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2024 Copyright held by the owner/author(s).
ACM 2469-7818/2024/09-ART8
https://doi.org/10.1145/3688599

information[1] on an online social network (e.g., the URL of an article on Facebook, a hashtag on Twitter, etc.) is an important problem with key societal implications. Such estimates of exposure can, for example, help researchers and the public track the prevalence and reach of election misinformation [4, 10] or improve the public health response to the Coronavirus [37, 39].

To measure the exposure to a piece of information on a network, one needs access to two features: the set of people who shared the piece of information, and the structure of the underlying social network. Of these two features, the structure of the underlying social network is often not publicly available, and fully or partially estimating it from data is not a practically feasible task due to the networks' massive size (e.g., billions of nodes and edges in Facebook), constantly evolving nature [26, 27] and limits placed by corporations on data collection. Similarly, the set of people who shared the piece of information is often also not publicly known and difficult to estimate from data, since it evolves as the piece of information spreads through the social network in the form of an information cascade, e.g., a URL of a news article that is being shared on Facebook. As a result, calculating exposure to a piece of information on a social network in a data-driven manner remains a challenging task.

This state of affairs is unfortunate as efficient (in terms of computation and resources) and accurate (in terms of statistical properties) estimates of exposure to information can provide two important benefits:

(i) from an *analytic perspective*, to identify widely consumed pieces of information and better understand their effect on the outcomes of various high-stake events such as elections, COVID vaccine acceptance, and so on [4, 39], and
(ii) from an *intervention perspective*, to prioritize various pieces of information in the content moderation and fact checking process of a social media platform to prevent large numbers of users from being exposed to harmful or misleading information.

Surprisingly, the problem of estimating exposure to information has received relatively little attention in the literature despite its importance. Formally, this problem can be stated as follows (in the context of an undirected social network such as Facebook).

Problem of estimating exposure to information: Consider an undirected social network $G = (V, E)$, and let $s(v) = 1$ if the node $v \in V$ shared (with the set of their neighbors $\mathcal{N}(v) \subset V$) a piece of information and $s(v) = 0$, otherwise. Assuming the graph G and the sharing function $s : V \to \{0, 1\}$ are unknown, estimate the fraction of nodes exposed to the piece of information,

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

where $f(v) = 1$ if the node $v \in V$ has been exposed to the piece of information by one of their neighbors and $f(v) = 0$ otherwise, i.e., $f(v) = \mathbb{1}_{\{\exists\, u \in \mathcal{N}(v) \text{ such that } s(u)=1\}}(v)$.

In the above formulation, the value $s(v) \in \{0, 1\}$ indicates whether the user (i.e., node) $v \in V$ shared the piece of information in concern and $f(v) \in \{0, 1\}$ indicates whether the user $v \in V$ has been exposed to it by one of their neighbours. Thus, the parameter of interest $\bar{f}$ denotes the average exposure to the piece of information in the social network, i.e., $\bar{f} = \mathbb{E}\{f(X)\}$ where X denotes a uniformly sampled node from the set of all nodes V. In this setting, we are tasked with devising a method to estimate the average exposure $\bar{f}$. Observe that the average exposure $\bar{f}$ defined

[1] A "piece of information" refers to any uniquely identifiable message, such as a URL or a hashtag, that is shared by the users on a social network with their contacts. Depending on the application and context, one can even define the "piece of information" to be a collection of such uniquely identifiable messages of similar nature (e.g., a set of URLs and hashtags about a particular incident or a cause).

in Equation (1) depends explicitly on the exposure $f(v)$ of each node $v \in V$ in the graph. Since the sharing function $s : V \to \{0, 1\}$ and the network $G = (V, E)$ are both unknown, the exposure function $f : V \to \{0, 1\}$ is also unknown (as it depends on both $s : V \to \{0, 1\}$ and $G = (V, E)$). Hence, computing the exact average exposure $\bar{f}$ is not practically feasible. However, the exposure $f(v_i) \in \{0, 1\}$ of a small number of sampled nodes $v_i, i = 1, 2, \ldots, n$ (where $n \ll |V|$) can often be found by looking at whether at least one neighbor of v_i (for each $i = 1, 2, \ldots, n$) shared the piece of information.

Main results:

(1) We propose two intuitive and practically feasible methods for estimating the exposure to a piece of information on an undirected social network: a vanilla method based on uniform sampling and a friendship paradox-based method. The vanilla method is a naive solution to the problem based on the well-known uniform sampling, and hence we treat it as a baseline. In contrast, the friendship paradox-based method utilizes a non-uniform sampling approach that is suitable for most practical settings. Both methods produce unbiased estimates of the average exposure to information $\bar{f}$.
(2) Via theoretical analysis and numerical experiments, we characterize the conditions under which the friendship paradox-based method outperforms the vanilla method and vice-versa. These conditions depend only on network properties that are typically known *a priori* based on the context, such as assortativity of the network and whether the sharers are more likely to be highly popular nodes or less popular fringe nodes. Hence, the characterizing conditions help choose the most accurate method for estimating exposure to information depending on the context of the problem.
(3) We extend the two proposed methods to the setting of a dynamic information cascade (where the time-varying exposure needs to be tracked in real time) as well as to the context of directed networks (e.g., Twitter).
(4) We provide detailed numerical simulations (based on synthetic data) as well as empirical experiments (based on real-world data) to illustrate the usefulness and feasibility of the proposed methods under various practical settings.

2 Related Work and Preliminaries

Our problem definition and methods are motivated by recent literature and events, which we expand on below. We then present background preliminaries on the main approach that we use in this work, the friendship paradox.

2.1 Motivation

The need for a principled method for estimating exposure to information in social media has intensified recently.

One reason for this increased interest in exposure estimation is the role that exposure to information on social networks can play in affecting the outcomes of high-stake events that define the course of our society. In particular, the question of reach of "fake news" on Facebook has become a major public concern following the 2016 U.S. presidential election [4, 10]. Similarly, large-scale exposure to false information on social media has complicated the public health response to the Coronavirus [37, 39]. Both examples highlight the need for tracking and quantifying exposure, for example, to prioritize fact-checking of trending coronavirus and election-related information content.

It has also become clear that the platforms cannot be trusted to reliably provide this information, in real time or retroactively [11, 16]. For example, Facebook recently acknowledged serious problems in the data provided to academic researchers in 2020 [1, 40] and reportedly shelved

"most-viewed pages" reports when those conflicted with the company's publicity goals [2]. Such incidents have given rise to the need of methods that can accurately estimate the exposure to various pieces of information independently without the involvement of the social media companies.

Researchers looking to estimate exposure to information in social networks are largely limited to survey-based offline methods. In such methods, a question is presented to a set of respondents to gather information on the frequency and pattern of their social media usage [17]. For example, researchers had used a post-election online survey to assess how exposure to fake news affected the 2016 U.S. election [4]. Survey-based methods have several limitations. First, they cannot be implemented in real time, and they are implemented after an event has occurred (e.g., an election). Consequently, survey-based methods can only be used for post-event type studies to collect retrospective data and are not suitable for real-time tasks such as identifying trending information for independent fact-checking, and so on. Second, as respondents self-report their usage frequencies and patterns, the outcomes of surveys are prone to over- and under- reporting errors as well as human cognitive biases [22]. As a result, the estimates of exposure to information based on surveys alone may not have rigorous theoretical guarantees on the accuracy. Other research had used panels of users who provided access to their web traffic to assess such exposure [19]. Closer to our baseline method here, researchers had used a panel of Twitter users to track their exposure to specific "fake news" URLs and domains shared by people they follow, but did not offer network-wide measures [18].

In this context, our definition of exposure to information fills the lack of a formal definition in literature and also assists in the development of rigorous algorithmic estimation methods. Further, as opposed to the post-event survey-based approach, the social network sampling-based methods proposed in this work can be implemented in real time (to track the progression of exposure as a piece of information spreads over time). These methods yield unbiased estimates of the exposure to information across the entire social network. In addition, these techniques can be implemented in a practically feasible manner without the full knowledge of the network (e.g., via a random walk) as well as the set of people who shared the piece of information.

2.2 Friendship Paradox

Our work here is motivated by the graph theoretic consequence named *friendship paradox*, which states, "On average, the number of friends of a random friend is always greater than or equal to the number of friends of a random individual." Formally:

Theorem 1 (Friendship Paradox [14]). *Consider an undirected graph $G = (V, E)$. Let X be a node sampled uniformly from V and Y be a uniformly sampled end-node from a uniformly sampled edge $e \in E$. Then,*

$$\mathbb{E}\{d(Y)\} \geq \mathbb{E}\{d(X)\}, \tag{2}$$

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. This is because it is a random person from a uniformly sampled pair of friends.[2] The intuition behind the friendship paradox (Theorem 1) is that individuals with large number of friends (i.e., high-degree nodes) appear as the friends of many others. Therefore, a random end of a random link (i.e., the random variable Y) is more likely to yield a high degree node than a uniformly sampled node (i.e., the random variable X). Consequently, sampling random friends (i.e., Y) allows us to reach high-degree nodes in the network without the full knowledge of the network.

[2]A random friend Y on an undirected graph $G = (V, E)$ has the distribution $\mathbb{P}\{Y = v\} \propto d(v)$ for all $v \in V$. In other words, a random friend is a node sampled with a probability proportional to their degrees.

Previous work that rely on the friendship paradox for statistical inference: The friendship paradox has been exploited in many statistical inference methods [23, 29], e.g., to reduce the variance in survey-based polling methods [30], to efficiently estimate power-law degree distributions [13, 31], and to quickly detect the outbreak of a disease [15]. Such work in literature aim to exploit the friendship paradox to sample individuals with larger number of friends on average, similar to one of the methods discussed in our work (i.e., the friendship paradox-based estimator). However, while relying on the friendship paradox for estimation remains a common feature between our work and previous works, the quantity being estimated is different. For example, Reference [31] exploits the friendship paradox to estimate the exponent of a power-law degree distribution by sampling the tail of the distribution more efficiently, and Reference [15] exploits the fact that random-friends are better sensors to be monitored for detecting disease outbreaks as they are likely to catch the disease earlier due to their larger social circles. Similarly, in polling [29], the quantity of interest is the fraction of nodes in the graph with a certain attribute (e.g., the fraction of nodes who will vote for a certain political candidate), and the friendship paradox-based methods exploit the fact that random friends can be more useful for this purpose by summarizing the voting intentions of their larger social circles. Unlike such problems and methods in References [15, 29, 31], the quantity we are interested in estimating is the *exposure to information* defined in Equation (1), which is the union of the social circles (neighborhoods) of all individuals that have shared the piece of information. Although the friendship paradox is still exploited in the proposed method, the quantity being estimated, the form of the estimator as well as its statistical properties are different from the previous works.

Besides its applications in statistical inference tasks, other implications (e.g., perception bias [3, 21, 24, 25], information diffusion [32]) and generalizations of the friendship paradox (e.g., References [7, 12, 20]) have been widely studied in the context of social networks.

3 Algorithmic Approach

In this section, we present two methods for estimating the average exposure to information: a vanilla method based on uniform sampling and a friendship paradox-based method. The conditions under which each method yields more accurate results than the other are derived subsequently.

3.1 Vanilla Method Based on Uniform Sampling

The vanilla approach for estimating the average exposure $\bar{f}$ works by obtaining a set of random nodes and checking whether these nodes have been exposed to the piece of information via their contacts.

Vanilla method for estimating exposure to information

Step 1: Sample n random nodes $X_1, \ldots, X_n$ uniformly and independently from the set of all nodes V.

Step 2: Use

$$\hat{\bar{f}}_{\text{VI}} = \frac{\sum_{i=1}^{n} f(X_i)}{n} \tag{3}$$

as the estimate of the average exposure $\bar{f}$.

The vanilla estimator $\hat{\bar{f}}_{\text{VI}}$ given in Equation (3) is unbiased, i.e., $\mathbb{E}\{\hat{\bar{f}}_{\text{VI}}\} = \bar{f}$. However, the vanilla estimator $\hat{\bar{f}}_{\text{VI}}$ would intuitively yield a larger variance, since random nodes are not likely exposed to the piece of information when $s(\cdot)$ is a very sparse function (i.e., only very few people shared the information).

3.2 Friendship Paradox-based Method

To reduce the variance in the estimate of average exposure, we can exploit the friendship paradox-based sampling (instead of vanilla uniform sampling) as follows:

Friendship paradox-based method for estimating exposure to information

Step 1: Sample n random friends $Y_1, \ldots, Y_n$ from the network independently (a random friend Y_i is a random end of a random link, i.e., a link is sampled uniformly from the network and one end of that link is taken with an unbiased coin flip).

Step 2: Use

$$\hat{\bar{f}}_{\text{FP}} = \frac{\bar{k}}{n} \sum_{i=1}^{n} \frac{f(Y_i)}{d(Y_i)} \tag{4}$$

as the estimate of average exposure $\bar{f}$, where $d(v)$ denotes the degree of $v \in V$ and $\bar{k}$ the average degree of the graph $G = (V, E)$.

The friendship paradox-based estimator $\hat{\bar{f}}_{\text{FP}}$ can be viewed as an application of importance sampling in social networks where the samples are generated from a different distribution (i.e., random friends Y) than the distribution that is directly related to the parameter of interest $\bar{f} = \mathbb{E}\{f(X)\}$ (i.e., random nodes X). In particular, random friends are more popular than random nodes in expectation (according to Theorem 1) and thus, sampling random friends will lower the variance by accessing individuals who are more likely to be exposed to the piece of information due to their large popularity (even when the sharing function $s(\cdot)$ is sparse). The additional terms $d(Y_i), \bar{k}$ that appear in $\hat{\bar{f}}_{\text{FP}}$ (compared to $\hat{\bar{f}}_{\text{VI}}$) correct for the bias resulting from sampling the more popular random friends Y_i instead of random nodes X_i. Further, the average degree $\bar{k}$ in Equation (4) is a known parameter for most social networks such as Facebook [41], which makes the implementation of the friendship paradox-based estimator $\hat{\bar{f}}_{\text{FP}}$ practically feasible. Section 5.3 discusses alternative implementations for situations where the edges cannot be sampled uniformly from the network.

To summarize, Section 3 presented two methods to estimate the average exposure to information based on uniform (vanilla) and friendship paradox-based sampling. Extensions of the two proposed methods will be discussed in Section 5.

4 Comparison of Statistical Properties of the Two Methods

In this section, we analyze and compare the statistical properties of the two proposed estimators (the vanilla estimator $\hat{\bar{f}}_{\text{VI}}$ given in Equation (3) and the friendship paradox-based estimator $\hat{\bar{f}}_{\text{FP}}$ given in Equation (4)). The aim of this analysis is to identify the conditions under which one estimator may be more accurate than the other for estimating the average exposure to information $\bar{f}$.

The following result (see Appendix A.1 for proof) characterizes the bias and variance of the two exposure estimators.

Theorem 2. *Consider the vanilla estimator $\hat{\bar{f}}_{VI}$ given in Equation (3) and the friendship paradox-based estimator $\hat{\bar{f}}_{FP}$ given in Equation (4).*

(1) Both the vanilla estimator $\hat{\bar{f}}_{VI}$ and the friendship paradox-based estimator $\hat{\bar{f}}_{FP}$ are unbiased estimators of the fraction of people exposed to a piece of information $\bar{f}$ (defined in Equation (1)), i.e.,

$$\mathbb{E}\left\{\hat{\bar{f}}_{VI}\right\} = \mathbb{E}\left\{\hat{\bar{f}}_{FP}\right\} = \bar{f}. \tag{5}$$

(2) The variances of vanilla estimator $\hat{\bar{f}}_{VI}$ and the friendship paradox-based estimator $\hat{\bar{f}}_{FP}$ are

$$\text{Var}\left\{\hat{\bar{f}}_{VI}\right\} = \frac{1}{n}\bar{f}\left(1 - \bar{f}\right), \quad \text{Var}\left\{\hat{\bar{f}}_{FP}\right\} = \frac{1}{n}\left(\bar{k}\,\mathbb{E}\left\{\frac{f(X)}{d(X)}\right\} - \bar{f}^2\right), \tag{6}$$

where $\bar{k}$ is the average degree of the graph.

As stated in the first part of Theorem 2, both the vanilla estimator $\hat{f}_{\text{VI}}$ and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ are unbiased estimators of the average exposure $\bar{f}$. Therefore, the method that has the smaller variance in the given setting should be used for estimating the average exposure $\bar{f}$. To do this, the rest of this section aims to identify the conditions under which one method outperforms the other in terms of the variance.

To theoretically compare the variances, we consider the class of undirected Markovian random networks that are completely characterized by their degree distribution $P(k)$ (which gives the probability that a uniformly sampled node from the network has degree k) and the conditional degree distribution $P(k'|k)$ (which gives the conditional probability that an edge from a degree k node connects to a degree k' node). The term "Markovian" here refers to the fact that all higher-order correlations can be expressed only in terms of the two functions $P(k)$ and $P(k'|k)$. We can derive a joint degree distribution using $P(k)$ and $P(k'|k)$ as

$$P(k,k') = \frac{kP(k)}{\bar{k}} P(k'|k), \tag{7}$$

which gives the probability that a uniformly sampled link connects two nodes with degrees k and k'. The correlation coefficient corresponding to this joint degree distribution $P(k,k')$ is called the *assortativity coefficient*, and we denote it with $r \in [-1,1]$. Networks for which $r > 0$ are called assortative networks, since high-degree nodes are more likely to be connected to other high-degree nodes and low-degree nodes are more likely to be connected to other low-degree nodes. However, networks for which $r < 0$ are called disassortative networks, since high-degree nodes and low-degree nodes are more likely to be connected with each other. A detailed description of Markovian random networks and assortativity can be found in References [5, 6].

In addition to $P(k)$ and $P(k'|k)$, which characterize the Markovian random network, we also define $P_s(1|k)$, which is the conditional probability that a node with degree k shares the piece of information. Consequently, $P_s(0|k) = 1 - P_s(1|k)$ is the probability that a node with degree k does not share the piece of information. Intuitively, if the $P_s(1|k)$ is closer to 1 for larger (respectively, smaller) values of k, then high-degree (respectively, low-degree) nodes are more likely to share the piece of information. In particular, if the sharing happens independently of the node popularity (i.e., high-degree and low-degree nodes are equally likely to share the piece of information), then $P_s(1|k)$ would be a constant that does not depend on degree k. Such relations (between sharing and degree) can be captured using the correlation coefficient between the sharing and the degree, which we refer to as the *degree-sharing correlation coefficient* and denote by $\rho \in [-1,1]$.

The following result (see Appendix A.2 for proof) compares the variances of the two estimators $\hat{f}_{\text{FP}}$, $\hat{f}_{\text{VI}}$ (given in Equations (3) and (4), respectively) in terms of the degree distribution $P(k)$, the conditional degree distribution $P(k'|k)$ and the conditional sharing probability $P_s(1|k)$ in the context of Markovian random networks.

Theorem 3. *The variance of the friendship paradox-based estimator $\hat{f}_{FP}$ given in Equation (4) is less than or equal to the variance of the vanilla estimator $\hat{f}_{VI}$ given in Equation (3) (i.e.,* $\text{Var}\{\hat{f}_{FP}\} \leq \text{Var}\{\hat{f}_{VI}\}$*) if and only if*

$$\mathbb{E}_{k\sim P(k)}\left\{\left(1-\frac{\bar{k}}{k}\right)\mathbb{P}\left\{f(X)=1|d(X)=k\right\}\right\} \geq 0, \tag{8}$$

where X is a uniformly sampled node from the network, $\mathbb{E}_{k\sim P(k)}$ denotes the expectation with respect to the degree distribution $P(k)$, and

$$\mathbb{P}\left\{f(X)=1|d(X)=k\right\} = 1-\left(\sum_{k'} P(k'|k)\, P_s(0|k')\right)^k. \tag{9}$$

Discussion of Theorem 3: Theorem 3 yields insights that help identify the settings where one method is more accurate (in terms of variance) compared to the other for estimating the average exposure to information $\bar{f}$. In particular, these insights relate the variance of the methods to important network parameters such as the degree distribution $P(k)$, assortativity coefficient r and the degree-sharing correlation coefficient ρ, as we discuss in detail below.

1. Choosing the best method based on assortativity coefficient r and degree-sharing correlation coefficient ρ: Due to the term $\left(1 - \bar{k}/k\right)$, the condition Equation (8) is more likely to be satisfied when the value $\mathbb{P}\{f(X) = 1|d(X) = k\}$ is closer to 1 for larger values of the degree k and the value $\mathbb{P}\{f(X) = 1|d(X) = k\}$ is closer to 0 for smaller values of the degree k. According to Equation (9), this happens when

(i) $P(k'|k)$ is closer to 1 for $k' \gg \bar{k} \gg k$ and $P_s(1|k')$ is closer to 0 for $k' \gg \bar{k}$, or
(ii) $P(k'|k)$ is closer to 1 for $k, k' \ll \bar{k}$ and $P_s(1|k')$ is closer to 0 for $k' \ll \bar{k}$.

Consequently, friendship paradox-based estimator $\hat{\bar{f}}_{\text{FP}}$ has a smaller variance compared to the vanilla estimator $\hat{\bar{f}}_{\text{VI}}$ when $r, \rho > 0$ (i.e., the network is assortative and the high-degree individuals are more likely to share the piece of information) or $r, \rho < 0$ (i.e., the network is disassortative and the low-degree individuals are more likely to share the piece of information). The arithmetic signs of the assortativity coefficient r and the degree-sharing correlation coefficient ρ are typically known based on the context. For example, it is known that $r > 0$ for most social networks whereas $r < 0$ for most technological networks [33]. Similarly, for pieces of information that tend to originate from highly influential individuals (e.g., social media influencers, celebrities, politicians), ρ is typically positive [43, 44]. Hence, the first insight helps us choose the best estimator for the given context based only on arithmetic signs of r and ρ.

2. When the sharing is independent from the popularity: If the node degree and the sharing are statistically independent, then Equation (9) yields that $\mathbb{P}\{f(X) = 1|d(X)\} = 1 - P_s(0)^k$, where $P_s(0) = 1 - P_s(1)$ is the probability that any node (independent of the degree k) does not share the piece of information. Consequently, when the node degree and the sharing are statistically independent, the friendship paradox-based estimator $\hat{\bar{f}}_{\text{FP}}$ has a smaller variance compared to the vanilla estimator $\hat{\bar{f}}_{\text{VI}}$ if and only if

$$\mathbb{E}_{k\sim P(k)}\left\{\left(1 - \frac{\bar{k}}{k}\right)\left(1 - P_s(0)^k\right)\right\} \geq 0, \tag{10}$$

according to Equation (8). We numerically evaluated the expected value in the left-hand side of Equation (10) for two types of degree distributions: a power-law distribution $P(k) \propto k^{-\alpha}$ (where $\alpha > 2$ is the power-law exponent) and an exponential distribution $P(k) = \frac{1}{\lambda}e^{-\lambda k}$ (where $\lambda > 0$ is a constant parameter). Our results suggest that the condition Equation (10) is not satisfied for all values of the power-law exponent $\alpha > 2$, the exponential parameter $\lambda > 0$, and the probability $P_s(0) \in [0, 1]$. Therefore, the vanilla estimator $\hat{\bar{f}}_{\text{VI}}$ produces the theoretically better estimate when the node degree and the sharing are statistically independent (implying that $\rho = 0$) and the network has either a power-law or an exponential degree distribution.

3. A widely shared piece of information versus a less widely shared piece of information: According to Equation (10), smaller values of $P_s(0)$ lead to a bigger disparity in the performance between the friendship paradox-based estimator $\hat{\bar{f}}_{\text{FP}}$ and the vanilla estimator $\hat{\bar{f}}_{\text{VI}}$. This implies that choosing the right method is of particular importance when the fraction of people who share a piece of information is small (compared to the size of the network). An important example for this situation is the starting phase of an information cascade where a piece of information has been less widely shared. Since identifying the correct exposure at this beginning stage is crucial for purposes such as

fact-checking before many people are exposed, this further highlights the importance of choosing the right method by using the insight (1) discussed earlier.

In summary, Section 4 theoretically analyzed the bias and variance of the two estimators presented in Section 3 (vanilla estimator $\hat{f}_{\text{Vl}}$ and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$) in terms of properties of the underlying network and the piece of information. Next, we present some extensions of the proposed methods (in Section 5), and then verify and complement the theoretical insights using numerical experiments (in Section 6).

5 Extensions of Proposed Methods and Practical Considerations

In this section, we extend the vanilla estimator $\hat{f}_{\text{Vl}}$ in Equation (3) and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ Equation (4) to directed networks (Section 5.1) and dynamic information cascades (Section 5.2). We also discuss details related to the implementation of the proposed methods in various practical settings.

5.1 Directed Networks

In a directed network $G = (V, E)$ (e.g., Twitter), a link $u \rightarrow v$ (pointing from a node $u \in V$ to $v \in V$) indicates that the node v follows the node u, i.e., u is the friend and v is the follower. Hence, the out-degree $d_o(v)$ and in-degree $d_i(v)$ of a node $v \in V$ denote the number of followers and friends of v, respectively. We say that a node $v \in V$ in a directed network is exposed to a piece of information (i.e., $f(v) = 1$) if at least one friend of v shared it. In this context, our aim is to estimate the average exposure to the piece of information that is denoted by $\bar{f}$.

To estimate the average exposure $\bar{f}$, a directed network can be sampled in three different ways: a random node $\mathcal{X}$ (sampled uniformly from V), a random friend $\mathcal{Y}$, which is the tail end of a uniformly sampled link (i.e., $\mathbb{P}(\mathcal{Y} = v) \propto d_o(v)$), a random follower $\mathcal{Z}$, which is the source end of a uniformly sampled link (i.e., $\mathbb{P}(\mathcal{Z} = v) \propto d_i(v)$). Consequently, we can construct three estimators of the average exposure $\bar{f}$ as follows:

$$\hat{f}_{\text{Vl}} = \frac{\sum_{i=1}^{n} f(\mathcal{X}_i)}{n}, \quad \hat{f}_{\text{Fr}} = \frac{\bar{k}}{n} \sum_{i=1}^{n} \frac{f(\mathcal{Y}_i)}{d_o(\mathcal{Y}_i)}, \quad \hat{f}_{\text{Fo}} = \frac{\bar{k}}{n} \sum_{i=1}^{n} \frac{f(\mathcal{Z}_i)}{d_i(\mathcal{Z}_i)}, \tag{11}$$

where $\bar{k}$ corresponds to the average in-degree $\mathbb{E}\{d_i(\mathcal{X})\}$ (which is also same as the average out-degree $\mathbb{E}\{d_o(\mathcal{X})\}$) and indices $i = 1, \ldots, n$ denote iid samples of each sampling method. The three estimators given in Equation (11) are motivated by the four versions of the friendship paradox that can exist on directed networks [3, 20]. In particular, one version says that a random follower on average has more friends than a random node (i.e., $\mathbb{E}\{d_i(\mathcal{Z})\} \geq \mathbb{E}\{d_i(\mathcal{X})\}$), implying that random followers ($\mathcal{Z}_i$) are more likely to be exposed to a piece of information. Hence, $\hat{f}_{\text{Fo}}$ in Equation (11) can reduce the variance by incorporating more exposed individuals into the sample.

5.2 Dynamic Information Cascades

The vanilla estimator $\hat{f}_{\text{Vl}}$ in Equation (3) and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ in Equation (4) assume that the function $s(\cdot)$ indicating the set of people who have shared the piece of information is static. However, the set of people who have shared a piece of information typically grows over time as it gets reshared and reposted by the users who were exposed it, leading to an information cascades. This subsection extends the vanilla estimator $\hat{f}_{\text{Vl}}$ and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ to track the increasing average exposure to such information cascades in real time. The key idea is to use a stochastic approximation algorithm with a constant step-size.

To simplify the notation, let us assume that only one sample can be collected at each time instant and there are no samples at time 0, allowing us to use the same variable n for discrete time and

the number of samples available. Further, let the vanilla and friendship paradox-based estimators at time n be denoted by $\hat{f}_{\text{VI}}^{(n)}$ and $\hat{f}_{\text{FP}}^{(n)}$, respectively. Then, note that

$$\begin{aligned}\hat{f}_{\text{VI}}^{(n)} &= \hat{f}_{\text{VI}}^{(n-1)} + \frac{1}{n}\left(f(X_n) - \hat{f}_{\text{VI}}^{(n-1)}\right),\\ \hat{f}_{\text{FP}}^{(n)} &= \hat{f}_{\text{FP}}^{(n-1)} + \frac{1}{n}\left(\bar{k}\frac{f(Y_n)}{d(Y_n)} - \hat{f}_{\text{FP}}^{(n-1)}\right),\end{aligned} \tag{12}$$

where X_n, Y_n denote a random node and a random friend at time n, respectively. The recursions in Equation (12) are obtained under the assumption that the average exposure $\bar{f}$ is time-invariant and, therefore, the update term decays with time (due to the decreasing step-size $1/n$) and converges to zero. Intuitively, this means that one new sample would not make a significant difference to an estimate (of a time-invariant parameter) derived with a relatively large number of samples (i.e., $n \gg 1$). In particular, it can be shown that the recursions in Equation (12) converge to the average exposure $\bar{f}$ with probability 1 under mild conditions.

However, the decreasing step-size $1/n$ in Equation (12) is not suitable when the average exposure is evolving over time (denoted by $\bar{f}^{(n)}$). This is because the decreasing step-size $1/n$ will stop updating eventually even though the average exposure $\bar{f}^{(n)}$ will keep changing. As a solution, the decreasing step-size in Equation (12) can be replaced with a constant step-size $\epsilon > 0$ for the case of time-evolving average exposure $\bar{f}^{(n)}$. Then, the new recursive methods for tracking the time-evolving average exposure $\bar{f}^{(n)}$ using the vanilla and friendship paradox-based methods will be as follows:

$$\hat{f}_{\text{VI}}^{(n)} = \hat{f}_{\text{VI}}^{(n-1)} + \epsilon\left(f(X_n) - \hat{f}_{\text{VI}}^{(n-1)}\right), \tag{13}$$

$$\hat{f}_{\text{FP}}^{(n)} = \hat{f}_{\text{FP}}^{(n-1)} + \epsilon\left(\bar{k}\frac{f(Y_n)}{d(Y_n)} - \hat{f}_{\text{FP}}^{(n-1)}\right). \tag{14}$$

The above two methods can track the progression of average exposure $\bar{f}^{(n)}$ when it is evolving on a slower timescale compared to the collection of samples. In other words, $\bar{f}^{(n)}$ is assumed to remain approximately constant for every $c > 1$ samples being collected; if $c \approx 1$ (respectively, $c \gg 1$), we say the piece of information is spreading rapidly (respectively, slowly). The value of the step-size parameter $\epsilon > 0$ in Equations (13) and (14) determines the effect of the update at each time. In particular, the value of ϵ should be relatively large (respectively, small) to track the average exposure to a piece of information that is spreading rapidly (respectively, slowly) through the social network.

5.3 Details on Practical Implementation

In this subsection, we discuss how several assumptions used for deriving the estimators proposed in Section 3 can be relaxed to increase the practical feasibility.

1. Implementing the friendship paradox-based estimator $\hat{f}_{FP}$ when the average degree $\bar{k}$ is unknown: The expression for the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ given in Equation (4) involves the average degree $\bar{k}$ of the underlying graph. Although average degree $\bar{k}$ is a known parameter for most widely used social networks such as Facebook and Twitter, there may be other real-world social networks where it is unknown, such as Tiktok and Mastodon [45]. Even mature social networks, such as Twitter and Facebook, may need to have the average degree estimated again after sudden changes in usage patterns and structure following disruptive events such as COVID pandemic [36]. To implement the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ (given in Equation (4)) in such settings, the average degree $\bar{k}$ can be estimated in a statistically efficient manner using the same set of randomly sampled friends $Y_1, Y_2, \ldots, Y_n$ assuming that the underlying degree distribution

has a specific parametric form. In particular, when the underlying degree distribution is assumed to have a power-law (heavy-tailed) degree distribution of the form $P(k) \propto k^{-\alpha}$ (where $\alpha > 2$ is the power-law exponent), Reference [31] shows that the maximum-likelihood estimator of α given by

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \ln(\frac{d(Y_i)}{k_{\min}})} + 2 \tag{15}$$

has a mean-squared error less than 0.01 for most synthetic and real-world networks. Then $\hat{\alpha}$ can be used to compute an estimate of the average degree. Reference [30] shows that a similar method can be used for exponential degree distributions as well. Therefore, when the underlying network is assumed to have a power-law or an exponential degree distribution, the unknown average degree $\bar{k}$ can be estimated without using any additional samples.

2. Implementing the friendship paradox-based estimator $\hat{f}_{FP}$ when the edges cannot be sampled uniformly: The implementation of the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ given in Equation (4) requires the uniform sampling of links from the underlying social network to obtain random friends $Y_1, Y_2, \ldots, Y_n$. This sampling approach is feasible in situations where links have unique IDs from a range of integers and the link corresponding to a given integer can be accessed. In settings where such uniform edge sampling is not possible (e.g., a fully unknown social network), the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ can be implemented via the use of random walks, since a stationary distribution of a random walk on an undirected, connected, non-bipartite graph samples nodes with probabilities proportional to their degrees (Reference [9], p. 298). Hence, the random variables $Y_1, Y_2, \ldots, Y_n$ could be replaced with samples from a sufficiently long random walk. Alternatively, one can also use a second version of the friendship paradox, which states that "uniformly sampled friend of a uniformly sampled node has more friends than a uniformly sampled node, on average" [8]. Hence, taking n uniformly sampled nodes and then taking one random friend of each of them would also be an alternative approach for friendship paradox-based sampling in undirected networks.

3. When the set of sharers is known: The set of sharers $S = \{v \in V : s(v) = 1\}$ maybe publicly known in some contexts (e.g., Twitter users who shared a particular hashtag). In such cases, several improvements can be made to the proposed methods. First, S is typically an array that can be ordered (e.g., a set of unique Twitter handles, a set of integer node IDs, etc.). Thus, for each sampled node $v \in V$, calculating $f(v) \in \{0, 1\}$ becomes equivalent to the problem of finding out whether two ordered arrays (the node v's neighbors $\mathcal{N}(v)$ and the set of sharers S) intersect or not. Hence, it is computationally easier to calculate $f(v)$ when the set of sharers $S \subset V$ is known. Second, when $S \subset V$ is known, the average degree of the sharers $\mathbb{E}\{d(X)|s(X) = 1\} = \frac{\sum_{v \in S} d(v)}{|S|}$ (which is typically hard to estimate if the set S is small compared to V) can be calculated. Further, the average degree of the people who have not shared (i.e., $\mathbb{E}\{d(X)|s(X) = 0\}$) can be estimated by sampling. Then, comparing the two values can be used as a heuristic estimate of the sign of the degree-sharing correlation coefficient ρ.

In summary, Section 5 extended the vanilla and friendship paradox-based estimators proposed in Section 3 to two settings: directed networks and dynamic information cascades. These extensions are numerically and empirically evaluated in the subsequent sections. We also discussed how some of the assumptions used for deriving the proposed estimators can be relaxed to increase practical feasibility.

6 Numerical Experiments

In this section, we numerically compare the vanilla estimator $\hat{f}_{\text{VI}}$ given in Equation (3) and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ given in Equation (4) using detailed simulation

experiments. The aim is to verify and complement the theoretical analysis in Section 4 and obtain additional insight into the performance of the two methods in various different settings.

Simulation setup: The configuration model [34] is used to synthetically generate 10,000 node power-law networks with exponents $\alpha = 2.5$ and $\alpha = 2.2$.[3] The assortativity coefficient of each network is then changed to three values ($r < 0, r = 0, r > 0$) using the attribute swapping method proposed in Reference [42].[4] Next, the sharing function values $\{s(v), v \in V\}$ are assigned as iid Bernoulli random variables and are then swapped using the method used in Reference [25] to generate three degree-label correlation coefficient values ($\rho < 0$, $\rho = 0$ and $\rho > 0$). The absolute error values for various sample sizes were then estimated using a Monte Carlo average of 5,000 iterations. The results obtained using this simulation setup are shown in Figure 1. To compare the recursive algorithms (given in Equations (13) and (14), respectively), we use two well-known information diffusion models: the **Independent Cascade model (ICM)** and the **Linear Threshold Model (LTM)**. The results for the ICM are shown in Figure 2 and the results for the LTM are given in Appendix B. A more detailed description of the simulation setup is given in Appendix B.

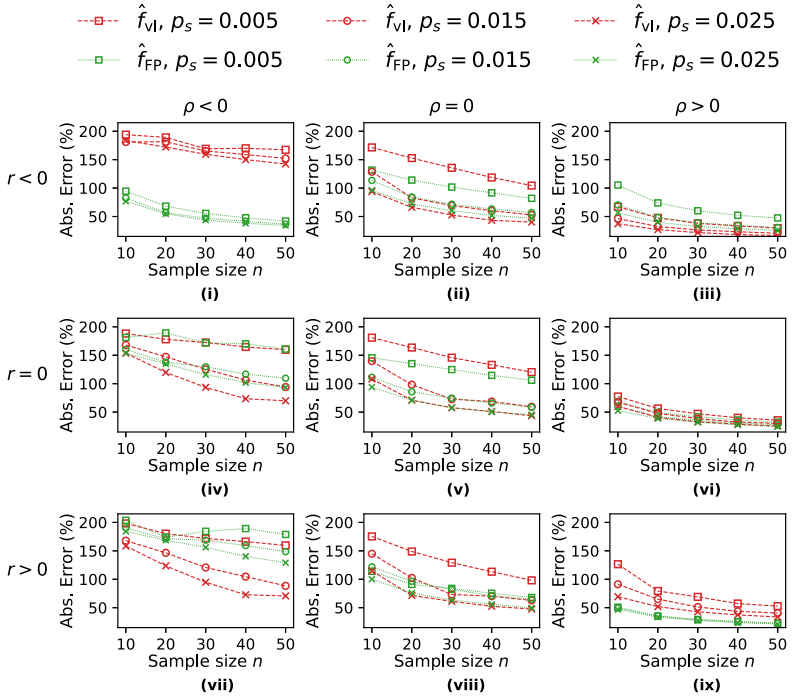
Discussion of the Numerical Results (Figures 1 and 2): The numerical results verify the theoretical conclusions (from Section 4) and yield additional insight as we discuss below.

1. Choosing the method that is best for the context: Figure 1(a) shows that the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ is more accurate (compared to the vanilla estimate $\hat{f}_{\text{VI}}$) when the assortativity coefficient r and the degree-sharing correlation coefficient ρ have the same signs (i.e., Figures 1(a)(i) and 1(a)(ix)). When the assortativity coefficient r and the degree-sharing correlation coefficient ρ have different signs (i.e., Figures 1(a)(iii) and 1(a)(vii)), the vanilla estimate $\hat{f}_{\text{VI}}$ is more accurate compared to the friendship paradox-based estimate $\hat{f}_{\text{FP}}$. In addition, the vanilla estimate $\hat{f}_{\text{VI}}$ has a smaller error when the degree and sharing are uncorrelated (i.e., $\rho = 0$ corresponding to middle column of Figures 1(a) and 1(b)) for when the sharing probability is not too small (so that both methods yield absolute errors smaller than 100% of the true parameter $\bar{f}$). Further, each subfigure of Figures 1(a) and 1(b) shows that the difference in the accuracy of the two estimates is larger when the unconditional sharing probability p_s (i.e., $p_s = \sum_k P_s(1|k) P(k)$) is smaller, highlighting that choosing the best method is crucial when the piece of information has been shared by only a smaller fraction of people. These numerical observations verify the theoretical expectations captured in the first and second points in the discussion related to Theorem 3 in Section 4, and emphasizes the importance of the choice for less widely shared pieces of information, per the third point.

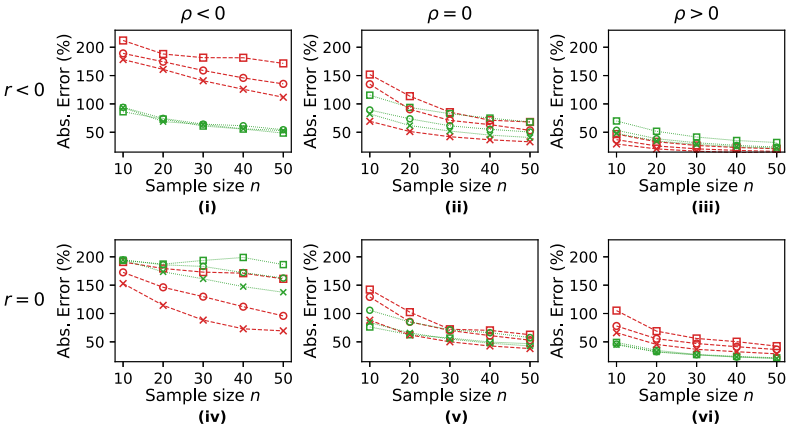
2. Implications of the heavy-tails: Comparing Figure 1(a) with Figure 1(b) indicates that the difference in the accuracy of the two methods is larger when the tail of the degree distribution is heavier (i.e., the power-law exponent α is smaller). For example, Figure 1(b)(vi) corresponding to $\alpha = 2.2$ shows that all red lines are above all green lines, indicating that the worst observed empirical accuracy of the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ is still better than the best observed empirical accuracy of the vanilla estimator $\hat{f}_{\text{VI}}$ (for the same sample size). However, Figure 1(a)(vi) corresponding to $\alpha = 2.5$ does not show such a clear separation between the two estimators.

[3]Previous empirical studies related to real-world networks have found that the power-law exponent α of real-world networks is in the interval (2, 3) [34, 35]. Due to this empirical finding, we choose to use two different values from within this range to explore the implication of the value of α on the performance of the proposed estimators of exposure to information. A larger exponent corresponds to a heavier tail in power-law distributions.

[4]For power-law graphs with smaller α (i.e., heavier tails), obtaining high-assortative graphs is theoretically impossible as explained in Reference [42]. As a consequence, the attribute swapping method does not converge when $r > 0$ and $\alpha = 2.2$. Therefore, we only consider the cases $r > 0$ and $r = 0$ for $\alpha = 2.2$.

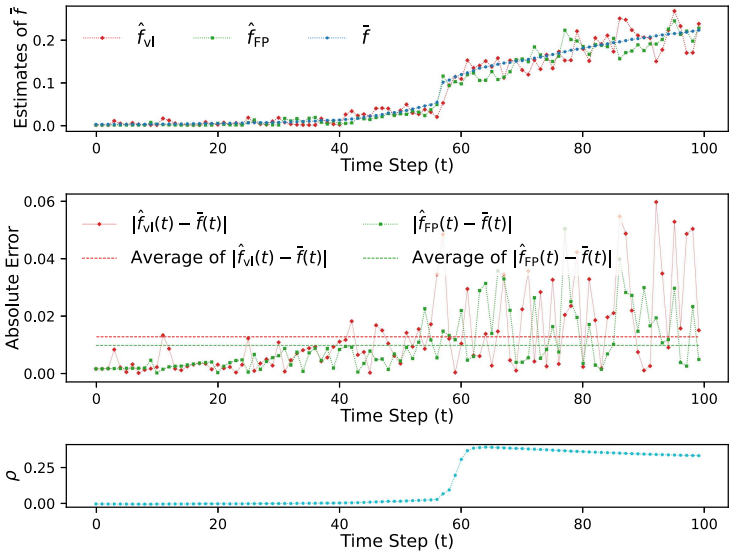


(a) Absolute error values for a power-law exponent $\alpha = 2.5$

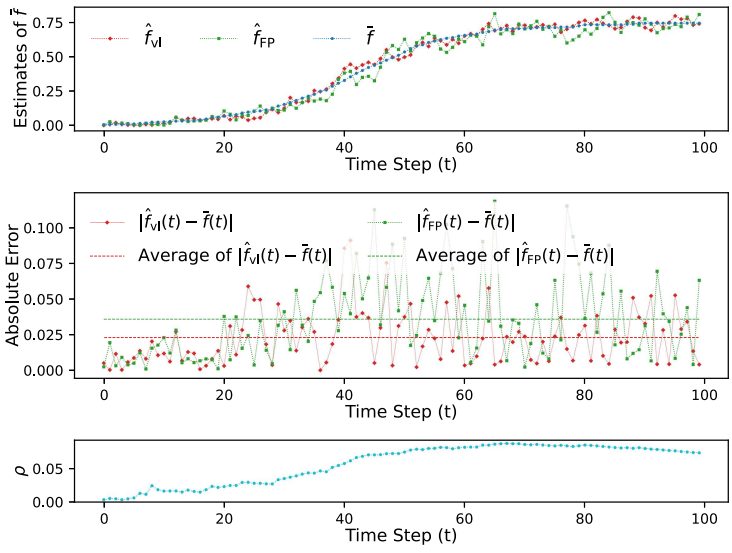


(b) Absolute error values for a power-law exponent $\alpha = 2.2$.

Fig. 1. Absolute error values of the vanilla estimate $\hat{f}_{\text{VI}}$ (given in Equation (3)) and the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ (given in Equation (4)) for two synthetically generated power-law networks (with exponents $\alpha = 2.5$ and $\alpha = 2.5$) with various values of the assortativity coefficient $\mathbf{r}$, degree-sharing correlation coefficient ρ, and the unconditional sharing probability p_s (i.e., p_s is the fraction of nodes that shared the piece of information). The plots show that the numerical results agree with the conclusions of the statistical analysis in Section 4. In particular, $\hat{f}_{\text{FP}}$ is the better choice when both $\mathbf{r}$ and ρ have the same sign and heavy-tails increase the disparity between the performances of the two estimators.



(a) Assortative network (i.e., assortativity coefficient $r > 0$)



(b) Disassortative network (i.e., assortativity coefficient $r < 0$)

Fig. 2. Performance of the vanilla and friendship paradox-based stochastic approximation algorithms (given in Equations (13) and (14), respectively) for tracking the exposure to an information cascade (in real-time) under the Independent Cascade model (ICM) on power-law networks with exponent $\alpha = 2.5$. In each sub-figure: the top-row shows the two estimates together with the true parameter $\bar{f}$, the middle-row shows the absolute errors corresponding to the two estimates at each time instant and the average error over all time instants, and the third row shows the variation of the degree-sharing correlation coefficient ρ. The friendship paradox-based algorithm works better for the assortative network (Figure 2(a)) while the vanilla algorithm works better for the disassortative network (Figure 2(b)). This observation agrees with the theoretical conclusions reached in Section 4.

Since real-world social networks have been empirically shown to have heavy-tails [34, 35], this observation highlights the importance of utilizing the theoretical insight to pick the best method for the social network using the insights discussed in the previous point.

3. Tracking the exposure to an information cascade in real time: Figure 2 shows the performance of the vanilla and the friendship paradox-based stochastic approximation algorithms (Equations (13) and (14), respectively) for tracking the exposure to an information cascade simulated from the ICM. In Figure 2(a), it can be clearly seen that the friendship paradox-based stochastic approximation Equation (14) outperforms the vanilla stochastic approximation Equation (13), especially after the time-step 40 where the diffusion process speeds up and the degree-sharing correlation coefficient starts increasing rapidly. In particular, the friendship paradox-based method closely detects the sudden phase transition of the diffusion process approximately at time-step 55 where the exposure suddenly jumps to a larger value. This result aligns with the theoretical conclusions in Section 4, since both assortativity coefficient r and the degree-sharing correlation coefficient ρ are both positive, leading to the friendship paradox-based method to outperform the vanilla method. However, the Figure 2(b) corresponds to the disassortative networks. Since the assortativity coefficient r and the degree-sharing correlation coefficient ρ have opposite signs, the vanilla method outperforms the friendship paradox-based method.

4. Implications of assortativity r and degree-sharing correlation ρ on the overall accuracy: It can be seen from Figure 1 that both estimates ($\hat{f}_{\text{VI}}$ and $\hat{f}_{\text{FP}}$) tend to be less accurate when the degree and sharing are negatively correlated (i.e., $\rho < 0$ in first column of Figure 1). Although the friendship paradox-based estimator performs better in Figure 1(a)(i), its accuracy decreases when moving to Figure 1(a)(iv,vii). This result is due to the fact that when $\rho < 0$, the nodes who share the piece of information are the less popular nodes, which makes the average exposure $\bar{f}$ smaller and more difficult to estimate. If $r < 0$ in addition to $\rho < 0$ (e.g., a star graph where outer nodes are sharing), then the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ can easily reach the core nodes that are more likely to be exposed due to their popularity, and thus reduce the variance as in Figure 1(a)(i). However, when $r > 0$ and $\rho < 0$ (i.e., Figure 1(a)(vii)), the less popular fringe nodes who share the piece of information are more likely to be separated from the core of the network so that the friendship paradox-based estimator cannot reach them. As such, both estimators tend to be the least accurate when $\rho < 0, r > 0$. An important example of this is the case where a piece of information originates with the less visible (i.e., fringe) nodes of the network. As such, special attention should be paid to choosing the best method when $\rho < 0$ to get the best possible accuracy.

In summary, Section 6 numerically compared the absolute errors of the estimates obtained using the two estimators presented in Section 3 (vanilla estimator $\hat{f}_{\text{VI}}$ and the friendship paradox-based estimator $\hat{f}_{\text{FP}}$). The numerical results agree with the theoretical conclusions provided in Section 4 and shed more light on the conditions under which one method outperforms the other.

7 Results on Real-world Networks

Evaluating the accuracy of the two estimators $\hat{f}_{\text{VI}}$, $\hat{f}_{\text{FP}}$ (proposed in Section 3) requires the true exposure $\bar{f}$ (i.e., the ground truth). However, as we stressed in Section 1, the exact value of the ground truth $\bar{f}$ depends on two features (the full network and the set of sharers), which are highly difficult to obtain, and our study was motivated by this difficulty in the first place. As such, comparing the estimates with the ground truth $\bar{f}$ in most real-world networks (e.g., Twitter, Facebook) is not feasible from a resource and computation viewpoint. Therefore, in this section, we first use real-world undirected networks with the sharing function generated synthetically (Section 7.1). We then evaluate our estimators on a real-world network with actual sharing data, using the directed ACM citation network (Section 7.2).

7.1 Undirected Networks

We tested the vanilla and friendship paradox-based estimators ($\hat{f}_{\text{Vl}}$, $\hat{f}_{\text{FP}}$) on four publicly available real-world network datasets in the SNAP database [28]. These networks include: a collaboration network between authors of papers submitted to Astrophysics and General Relativity in the Arxiv website, a network of Facebook pages of athletes, and a Facebook page network of different companies. For these four networks, the sharing function $s : V \to \{0, 1\}$ was synthetically generated using the methods in Section 6. The results obtained using these five real-world networks are shown in Figure 3.

For the network datasets corresponding to Figures 3(a)–3(c) (where $r > 0$), the friendship paradox-based estimator $\hat{f}_{\text{FP}}$ outperforms the vanilla estimator $\hat{f}_{\text{Vl}}$ when $\rho > 0$ while both methods have relatively large error values (above 100%) when $\rho < 0$. This result is as we expected, since $\hat{f}_{\text{FP}}$ works better when $r, \rho > 0$ (as per the first point in the discussion related to Theorem 3) and both $\hat{f}_{\text{Vl}}$, $\hat{f}_{\text{FP}}$ tend to be less accurate when $\rho < 0, r > 0$ (as per the fourth point in the discussion of numerical results). For the network dataset corresponding to Figure 3(d), $\hat{f}_{\text{FP}}$ (respectively, $\hat{f}_{\text{Vl}}$) works better when $\rho < 0$ (respectively, $\rho > 0$), since the network has $r < 0$ (as we theoretically expected). Therefore, the empirical findings align with both the theoretical and numerical results (Sections 4 and 6, respectively).

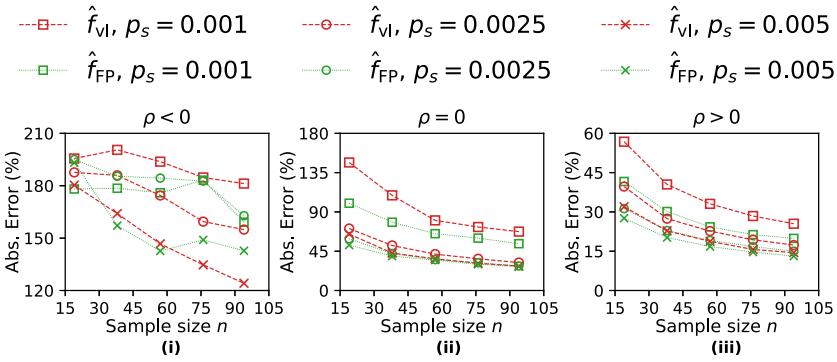
7.2 ACM Citation Network

In this analysis, we provide a full real-world network (i.e., without subsampling the network) as well as actual sharing data from that network. Specifically, we create a citation network of 217,335 academic papers using the data in Reference [38]. Any paper that contains a specific phrase in its title is then considered as a "sharer" of that phrase and all papers that cite that paper as the "exposed." The three estimators given in Equation (11) are then evaluated (via the same Monte Carlo averaging approach used in Section 6) for 25 popular phrases, 25 average phrases, and 25 unpopular phrases. Additional details on the experiments are given in Appendix B.

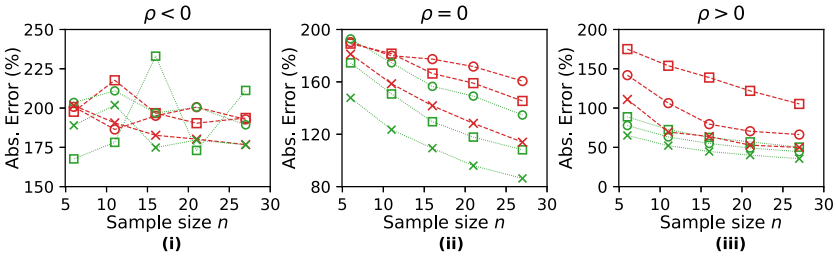
Figure 4 shows the average absolute errors of the vanilla estimate $\hat{f}_{\text{Vl}}$, friend-based estimate $\hat{f}_{\text{Fr}}$, and follower-based estimate $\hat{f}_{\text{Fo}}$ for: (a) popular phrases, (b) average phrases, and (c) unpopular phrases. In each case, the follower-based estimate $\hat{f}_{\text{Fo}}$ outperforms the other two as we expected from the directed versions of the friendship paradox mentioned in Section 5.1. In particular, Figure 4 shows that the absolute error of all three estimates increase as the popularity of the phrases decrease. However, the follower-based estimate $\hat{f}_{\text{Fo}}$ is more accurate compared to the other two estimates, especially for unpopular phrases (Figure 4(c)). This is because random followers are more likely to be exposed even to an unpopular piece of information due to their larger friend count (according to the directed versions of the friendship paradox as discussed in Section 5.1) and hence, sampling random friends lowers the variance of the estimate.

8 Discussion and Conclusion

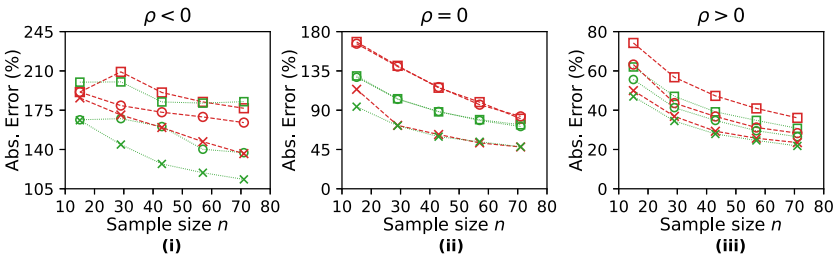
Summary: Knowing the exposure to various pieces of information in online social networks is crucial for understanding the impact of various types of content such as misinformation and news articles as well as for effective fact checking and content moderation. However exactly computing the exposure to a piece of information is practically difficult, since it requires the structure of the social network as well as the individuals who have shared the piece of information to be known. As a solution, we presented a practically feasible framework for estimating the fraction of people who have been exposed to a piece of information by their contacts (i.e., average exposure) in a social network. In particular, we proposed two methods to estimate the average exposure: a vanilla method that is based on uniform sampling and a method that samples random friends (random ends of



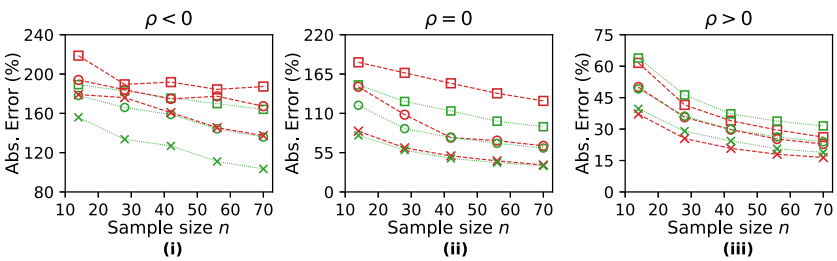
(a) For an astrophysics co-authorship network ($r = 0.21$)



(b) For a general relativity co-authorship network ($r = 0.66$)



(c) For a company Facebook page dataset ($r = 0.01$)



(d) For an athlete Facebook page dataset ($r = -0.03$)

Fig. 3. Absolute error values of the vanilla estimate $\hat{f}_{\text{VI}}$ (given in Equation (3)) and the friendship paradox-based estimate $\hat{f}_{\text{FP}}$ (given in Equation (4)) for four real-world network datasets. These results validate the theoretical insights (Section 4) and complement the numerical experiments (Section 6).

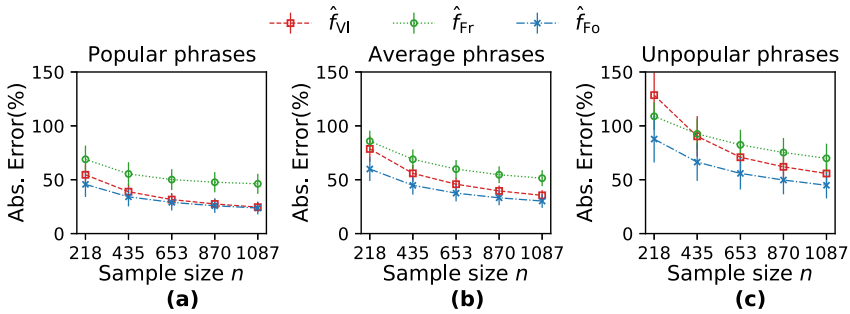


Fig. 4. Absolute error values of the vanilla estimate $\hat{f}_{\mathrm{Vl}}$, the friend-based estimate $\hat{f}_{\mathrm{Fr}}$ and the follower-based estimate $\hat{f}_{\mathrm{Fo}}$ given in Equation (11) on a real-world, directed network. It can be seen that estimate $\hat{f}_{\mathrm{Fo}}$ is more accurate compared to the other two estimates as we expected based on the directed versions of the friendship paradox (discussed in Section 5.1).

random links). The latter method can be thought of as a variance reduction method motivated by the friendship paradox, which incorporates more high-degree individuals (who are more likely to be exposed to the piece of information due to their large popularity) into the sample. Both methods are unbiased, and we provided theoretical results, which characterize the conditions where one method outperforms the other in terms of the variance. Specifically, the friendship paradox-based estimator has a smaller variance compared to the vanilla (uniform sampling-based) method when both the assortativity coefficient of the network and the correlation coefficient between the degree and information sharing have the same arithmetic sign. We also presented extensions of the proposed methods to directed networks and dynamic information cascades (where the average exposure needs to be tracked in real time). The proposed methods and the theoretical conclusions were verified numerically (via simulations) as well as via experiments on several real-world network datasets.

Ethical considerations: While the methods proposed in this article offer practical approaches to estimate exposure to information, attention must be paid to ethical aspects and potential misuses when implementing such approaches. For example, the individual level exposure to various pieces of information should be unidentifiable via the obtained estimates of population level exposure to protect the privacy of the users of the social network platform. Coupling machine learning frameworks such as differential privacy with the proposed methods could a important future research direction in this regard. Further, implementing the algorithms on online social networks should adhere to the rules and restrictions of the platform and should not constitute a privacy infringement.

Limitations and future research directions: While the proposed methods can be utilized to estimate the exposure to information, they are not without limitations. For instance, the definition of exposure to information that we utilize would consider an individual to be exposed if at least one of their social contacts has shared the piece of information. Instead, one can extend the proposed methods based on a definition that considers various degrees of exposure (e.g., highly exposed, moderately exposed, not exposed) determined by the fraction of contacts who have shared the piece of information. Another key limitation is the fact that the proposed algorithms need to look at all neighbors of each sampled individual to find the whether that person has been exposed. This is difficult when dealing with either incomplete data or settings where all neighbors of certain nodes cannot be identified. Extending the proposed methods to work with only a random subset of neighbors could be a useful future direction to avoid this limitation. Further, the insights on the performance of the two methods (vanilla method and friendship paradox-based method) relied

largely on the assumption that the network structure is dependent only on the assortativity and the degree distribution. Relaxing this assumption, analyzing the implications of real-world network properties such as community structure, clustering, and smaller diameters, and exploiting these to increase the accuracy of the proposed methods remain important future steps. Generalizing the proposed methods to simultaneously estimate the exposure to multiple pieces of information would also be useful. Similarly, generalizing the proposed methods to estimate the exposure to a piece of information to various groups of nodes instead of all nodes (e.g., fraction of conservatives and the fraction of liberals exposed to a specific piece information) would be highly useful in obtaining a deeper understanding of the information exposure patterns. Finally, incorporating the effects of recommendation systems and other such features of social media platforms into the process of estimation of exposure to information would make the proposed methods practically more useful.

Appendices

A Proofs of Theorems

A.1 Proof of Theorem 2

Part 1: For the vanilla estimate $\hat{f}_{\text{VI}}$, it follows that $\mathbb{E}\{\hat{f}_{\text{VI}}\} = \bar{f}$, since it is the average of n iid Bernoulli random variables (with parameter $\bar{f}$). For the friendship paradox-based estimate $\hat{f}_{\text{FP}}$,

$$\begin{aligned}
\mathbb{E}\left\{\hat{f}_{\text{FP}}\right\} &= \mathbb{E}\left\{\frac{\bar{k}}{n}\sum_{i=1}^{n}\frac{f(Y_i)}{d(Y_i)}\right\} \\
&= \bar{k}\,\mathbb{E}\left\{\frac{f(Y_1)}{d(Y_1)}\right\} \quad (\because Y_1,\ldots,Y_n \text{ are iid samples.}) \\
&= \bar{k}\sum_{v\in V}\frac{f(v)}{d(v)}\times\frac{d(v)}{\sum_{v\in V}d(v)} \quad \left(\because \mathbb{P}(Y_1=v)=\frac{d(v)}{\sum_{v\in V}d(v)}\right) \\
&= \bar{k}\sum_{v\in V}\frac{f(v)}{n\bar{k}} = \sum_{v\in V}\frac{f(v)}{n} = \bar{f}.
\end{aligned}$$

Therefore, both $\hat{f}_{\text{VI}}, \hat{f}_{\text{FP}}$ are unbiased estimates of $\bar{f}$.

Part 2: Consider the variance of the vanilla estimate $\hat{f}_{\text{VI}}$. Since the estimate is the average of n iid Bernoulli random variables (with parameter $\bar{f}$), their variance is given by $\bar{f}(1-\bar{f})/n$. For the friendship paradox-based estimate $\hat{f}_{\text{FP}}$,

$$\begin{aligned}
\text{Var}\left\{\hat{f}_{\text{FP}}\right\} &= \text{Var}\left\{\frac{\bar{k}}{n}\sum_{i=1}^{n}\frac{f(Y_i)}{d(Y_i)}\right\} \\
&= \frac{1}{n}\,\text{Var}\left\{\bar{k}\frac{f(Y_1)}{d(Y_1)}\right\} \quad (\because Y_1,\ldots,Y_n \text{ are iid samples.}) \\
&= \frac{1}{n}\left(\mathbb{E}\left\{\left(\bar{k}\frac{f(Y_1)}{d(Y_1)}\right)^2\right\} - \bar{f}^2\right) \quad \left(\because \mathbb{E}\left\{\hat{f}_{\text{FP}}\right\} = \bar{f}\right) \\
&= \frac{1}{n}\left(\bar{k}^2\sum_{v\in V}\frac{f^2(v)}{d^2(v)}\times\frac{d(v)}{\sum_{v\in V}d(v)} - \bar{f}^2\right)
\end{aligned}$$

$$\left(\because \mathbb{P}(Y_1 = v) = \frac{d(v)}{\sum_{v\in V} d(v)}\right)$$
$$= \frac{1}{n}\left(\bar{k}^2 \sum_{v\in V} \frac{f(v)}{d(v)} \times \frac{1}{n\bar{k}} - \bar{f}^2\right) = \frac{1}{n}\left(\bar{k}\,\mathbb{E}\left\{\frac{f(X)}{d(X)}\right\} - \bar{f}^2\right).$$

A.2 Proof of Theorem 3

Note that $P(k'|k)\,P_s(0|k')$ is the probability that a degree k node connects to a degree k' node that has not shared the piece of information. Therefore, averaging this term over the value k' (i.e., $\sum_{k'} P(k'|k)\,P_s(0|k')$) yields the probability that a degree k node having a neighbor that has not shared the piece of information. For a degree k node to not be exposed to the information, all k neighbors of that node must not have shared the piece of information. Hence, $(\sum_{k'} P(k'|k)\,P_s(0|k'))^k$ is the probability that a node with degree k has not been exposed to the piece of information, i.e.,

$$\mathbb{P}\{f(X) = 0|d(X) = k\} = \left(\sum_{k'} P(k'|k)\,P_s(0|k')\right)^k, \tag{16}$$

which yields Equation (9).

Next, using the expressions for the variance in Equation (6) and the fact that $\bar{f} = \mathbb{E}\{f(X)\}$ (where X is a uniformly sampled node), we get

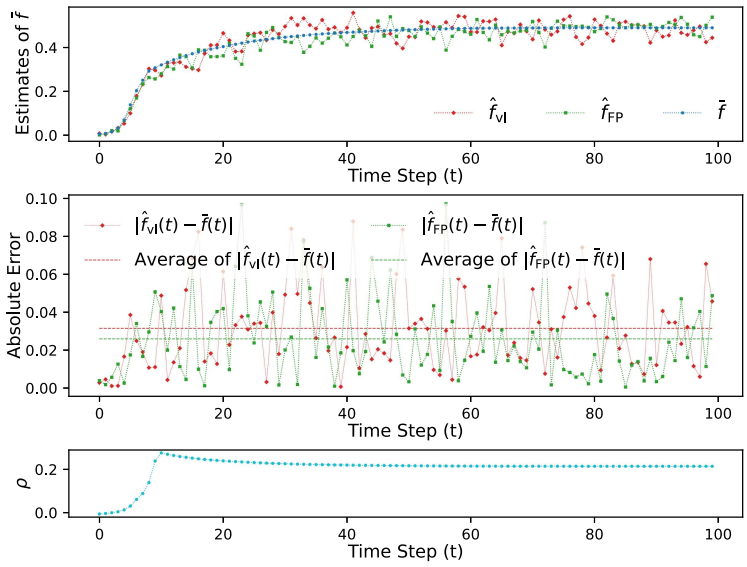
$$\mathrm{Var}\{\hat{f}_{\mathrm{VI}}\} \geq \mathrm{Var}\{\hat{f}_{\mathrm{FP}}\} \iff \frac{1}{n}\left(\bar{k}\,\mathbb{E}\left\{\frac{f(X)}{d(X)}\right\} - \bar{f}^2\right) \geq \frac{1}{n}\bar{f}\left(1-\bar{f}\right)$$
$$\iff \bar{f} - \bar{k}\,\mathbb{E}\left\{\frac{f(X)}{d(X)}\right\} \geq 0$$
$$\iff \mathbb{E}\left\{f(X)\left(1 - \frac{\bar{k}}{d(X)}\right)\right\} \geq 0 \quad (\because \mathbb{E}\{f(X)\} = \bar{f} \text{ from Equation (5)})$$
$$\iff \mathbb{E}_{k\sim P}\left\{\mathbb{E}\left\{f(X)\left(1 - \frac{\bar{k}}{d(X)}\right)\middle| d(X) = k\right\}\right\} \geq 0$$
(by conditioning on $d(X) = k$ and then averaging over k)
$$\iff \mathbb{E}_{k\sim P}\left\{\left(1 - \frac{\bar{k}}{k}\right)\mathbb{P}\{f(X) = 1|d(X) = k\}\right\} \geq 0.$$

B Additional Numerical Results and Details for Reproducibility

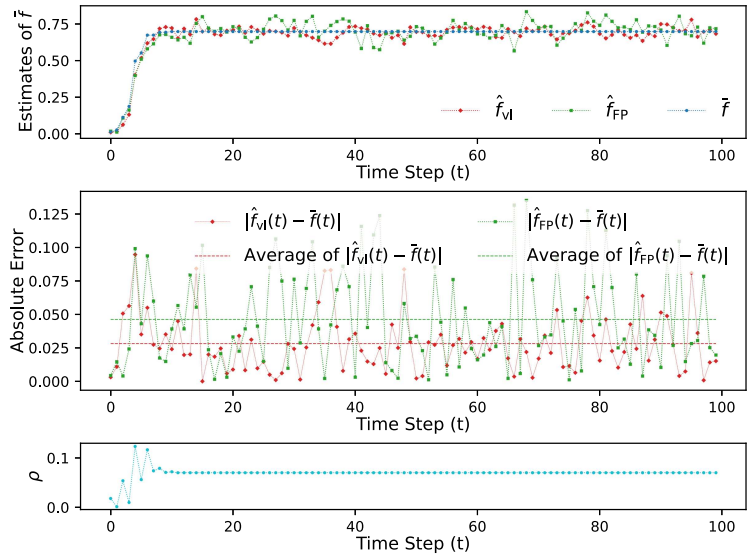
Below, we provide additional details related to the simulation setup used to generate the numerical results in Section 6.

Detailed Simulation setup for comparing the estimators $\hat{f}_{VI}$, $\hat{f}_{FP}$ (Figure 1): To generate the power-law networks, a sequence of 10K random variables from a power-law distribution with the required power-law exponent α were generated and rounded up to the nearest integer. Then, the first number in the sequence was altered by a value of 1 if needed to make sure that the sum of the numbers is even (to be valid sequence of degrees). Then, the configuration model (*configuration_model* function in the *networkx* package) was used to generate the networks with the given degree sequence.

To change the assortativity of the generated power-law networks, we first sample two edges from the network uniformly (without replacement) and rewire them to increase or decrease the assortativity coefficient r from the initial value. Lemma 1 of Reference [42], which orders the three



(a) Assortative network (i.e., assortativity coefficient $r > 0$)



(b) Disassortative network (i.e., assortativity coefficient $r < 0$)

Fig. 5. Performance of stochastic approximation algorithms based on the vanilla and friendship paradox-based estimates (given in Equations (13) and (14), respectively) for tracking the exposure to an information cascade under the Linear Threshold Model (LTM) on power-law networks with the exponent $\alpha = 2.5$. This result complements the result shown in Figure 2 for the Independent Cascade Model (ICM) and show that the conclusions reached using the ICM hold for the LTM as well. Thus, the proposed stochastic approximation algorithms can be used to track the exposure to information cascades with various dynamical properties.

possible ways to rewire the two selected edges based on the resulting assortativity coefficient values, is used to get the maximum increase or decrease in the assortativity coefficient when rewiring. This process is repeated until the required assortativity coefficient values ($r \in \{-0.2, 0, 0.2\}$) are reached or the maximum number of iterations (100K) is reached.

To change the degree-sharing correlation coefficient ρ, we follow the attribute swapping procedure used in Reference [25] as follows. First, the value $s(v) \in \{0, 1\}$ for each node $v \in V$ is first assigned as an iid Bernoulli random variable whose parameter determines the fraction of the people that share the piece of information. Then, we uniformly pick a node u from the set of nodes who shared the piece of information (i.e., $s(u) = 1$) and another node v from the set of the people who has not shared the piece of information uniformly (i.e., $s(v) = 0$). Next, to increase (respectively, decrease) degree-sharing correlation coefficient ρ, we swap $s(v)$ and $s(u)$ if $d(u) < d(v)$ (respectively, $d(u) > d(v)$). This process is repeated until the required degree-sharing correlation coefficient values ($\rho \in \{-0.2, 0, 0.2\}$) are reached or the maximum number of iterations (100K) is reached.

Simulation setup for comparing the vanilla and the friendship paradox-based stochastic approximations (given in Equations (13) *and* (14)*):* Under the ICM used to generate Figure 2, each neighbor of a node who shared a piece of information at a previous time instant shares in the current time instant with a pre-specified probability named the *infection probability*. For Figure 2, the diffusion was initialized with 10 uniformly chosen nodes and the infection probability is set to 0.05. Additionally, the step size ϵ of the stochastic approximations Equations (13) and (14) is set to 0.01. Further, it is assumed that the stochastic approximations Equations (13) and (14) are updated 100 times for each step of the diffusion process, i.e., the samples are collected 100 times faster than the evolution of the diffusion process.

Figure 5 shows analogous results obtained using LTM, where a node shares a piece of information at the current time instant if the fraction of its neighbors that have already shared it by the previous time instant exceeds a certain threshold. We choose the threshold value to be 5% for the Figure 5. The step-size of both stochastic approximations Equations (13) and (14) as well as the number of samples collected at each time instant are the same as the case for the ICM.

Filtering the ACM Citation Network: We obtained a dataset of 629,814 papers from DBLP, ACM, and MAG (Microsoft Academic Group) [38]. We filtered out papers that did not have references within the original dataset or were not referenced by another paper in the original dataset to create a final dataset of 217,335 papers. Afterward, we determined phrases of varying popularity. We first filter the papers' titles for stopwords and determine the frequency of each word to create a numerically sorted dictionary with word frequency pairs. Then, we use NLTK's bigram association measures to create word pairs (i.e., phrases) using a subset of words from the dictionary's beginning. We defined popular phrases as having more than 400 sharers (e.g., *data mining*, *information systems*), average phrases with between 200 and 400 sharers (e.g., *computer graphics*, *embedded systems*), and unpopular phrases with 100 to 200 sharers (e.g., *network design*, *optimization problems*).

Code and Data Availability: All codes and datasets used in this article are publicly available at: https://github.com/ComplexInfo/Estimating_Info_Exposure.

References

[1] Davey Alba. 2021. Facebook Sent Flawed Data to Misinformation Researchers. *New York Times.* Retrieved from https://www.nytimes.com/live/2020/2020-election-misinformation-distortions#facebook-sent-flawed-data-to-misinformation-researchers

[2] Davey Alba and Ryan Mac. 2021. Facebook, Fearing Public Outcry, Shelved Earlier Report on Popular Posts. *New York Times.* Retrieved from https://www.nytimes.com/2021/08/20/technology/facebook-popular-posts.html

[3] Nazanin Alipourfard, Buddhika Nettasinghe, Andrés Abeliuk, Vikram Krishnamurthy, and Kristina Lerman. 2020. Friendship paradox biases perceptions in directed networks. *Nature Commun.* 11, 1 (2020), 1–9.
[4] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *J. Econ. Perspect.* 31, 2 (2017), 211–36.
[5] Marián Boguñá and Romualdo Pastor-Satorras. 2002. Epidemic spreading in correlated complex networks. *Phys. Rev. E* 66, 4 (2002), 047104.
[6] Marián Boguñá, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2003. Epidemic spreading in complex networks with degree correlations. In *Statistical Mechanics of Complex Networks*. Springer, 127–147.
[7] George T. Cantwell, Alec Kirkley, and M. E. J. Newman. 2021. The friendship paradox in real and model networks. *J. Complex Netw.* 9, 2 (2021), cnab011.
[8] Yang Cao and Sheldon M. Ross. 2016. The Friendship Paradox. *Math. Sci.* 41, 1 (2016), 61–64.
[9] Rick Durrett. 2010. *Probability: Theory and Examples* (4th ed.). Cambridge University Press, Cambridge, UK.
[10] Elizabeth Dwoskin. 2021. Misinformation on Facebook Got Six Times More Clicks Than Factual News During the 2020 Election, Study Says. *The Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2021/09/03/facebook-misinformation-nyu-study/
[11] Gilad Edelman. 2021. What Social Media Needs to Learn From Traditional Media. *Wired*. Retrieved from https://www.wired.com/story/what-social-media-needs-to-learn-from-traditional-media/
[12] Young-Ho Eom and Hang-Hyun Jo. 2014. Generalized friendship paradox in complex networks: The case of scientific collaboration. *Sci. Rep.* 4, 1 (2014), 1–6.
[13] Young-Ho Eom and Hang-Hyun Jo. 2015. Tail-scope: Using friends to estimate heavy tails of degree distributions in large-scale complex networks. *Sci. Rep.* 5, 1 (2015), 1–9.
[14] Scott L. Feld. 1991. Why your friends have more friends than you do. *Amer. J. Sociol.* 96, 6 (1991), 1464–1477.
[15] Manuel Garcia-Herranz, Esteban Moro, Manuel Cebrian, Nicholas A. Christakis, and James H. Fowler. 2014. Using friends as sensors to detect global-scale contagious outbreaks. *PloS One* 9, 4 (2014), e92413.
[16] Shirin Ghaffary. 2021. Why No One Really Knows how Bad Facebook's Vaccine Misinformation Problem is. *Vox*. Retrieved from https://www.vox.com/22622070/facebook-data-covid-19-vaccine-misinformation-researchers-access-nyu-academics
[17] Seth K. Goldman and Stephen Warren. 2020. Debating how to measure media exposure in surveys. *The Oxford Handbook of Electoral Persuasion*. Oxford University Press, Oxford, UK.
[18] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378.
[19] Andrew M. Guess, Brendan Nyhan, and Jason Reifler. 2020. Exposure to untrustworthy websites in the 2016 U.S. election. *Nature Hum. Behav.* 4, 5 (2020), 472–480.
[20] D. J. Higham. 2019. Centrality-friendship paradoxes: when our friends are more important than us. *Journal of Complex Networks* 7, 4 (2019), 515–528.
[21] Matthew O. Jackson. 2019. The friendship paradox and systematic biases in perceptions and social norms. *J. Politic. Econ.* 127, 2 (2019), 777–818.
[22] Tobias Konitzer, Jennifer Allen, Stephanie Eckman, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2021. Comparing estimates of news consumption from survey and passively collected behavioral data. *Public Opin. Quart.* 85, S1 (2021), 347–370.
[23] Vikram Krishnamurthy and Buddhika Nettasinghe. 2019. Information diffusion in social networks: Friendship paradox based models and statistical inference. *Modeling, Stochastic Control, Optimization, and Applications* (2019), 369–406.
[24] Eun Lee, Sungmin Lee, Young-Ho Eom, Petter Holme, and Hang-Hyun Jo. 2019. Impact of perception models on friendship paradox and opinion formation. *Phys. Rev. E* 99, 5 (2019), 052302.
[25] Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. 2016. The "majority illusion" in social networks. *PloS One* 11, 2 (2016), e0147617.
[26] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. 2008. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 462–470.
[27] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2005. Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 177–187.
[28] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. Retrieved from http://snap.stanford.edu/data
[29] Buddhika Nettasinghe and Vikram Krishnamurthy. 2019. The friendship paradox: Implications in statistical inference of social networks. In *Proceedings of the IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP'19)*. IEEE, 1–6.
[30] Buddhika Nettasinghe and Vikram Krishnamurthy. 2021. "What Do Your Friends Think?": Efficient polling methods for networks using friendship paradox. *IEEE Trans. Knowl. Data Eng.* 33, 3 (2021), 1291– 1305.

[31] Buddhika Nettasinghe and Vikram Krishnamurthy. 2021. Maximum likelihood estimation of power-law degree distributions via friendship paradox-based sampling. *ACM Trans. Knowl. Discov. Data* 15, 6 (2021), 1–28.
[32] Buddhika Nettasinghe, Vikram Krishnamurthy, and Kristina Lerman. 2019. Diffusion in social networks: Effects of monophilic contagion, friendship paradox, and reactive networks. *IEEE Trans. Netw. Sci. Eng.* 7, 3 (2019), 1121–1132.
[33] Mark E. J. Newman. 2002. Assortative mixing in networks. *Phys. Rev. Lett.* 89, 20 (2002), 208701.
[34] Mark E. J. Newman. 2003. The structure and function of complex networks. *SIAM Rev.* 45, 2 (2003), 167–256.
[35] Mark E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* 46, 5 (2005), 323–351.
[36] Kerstin Paschke, Maria Isabella Austermann, Kathrin Simon-Kutscher, and Rainer Thomasius. 2021. Adolescent gaming and social media usage before and during the COVID-19 pandemic. *Sucht* (2021).
[37] Erin Simpson and Adam Conner. 2020. Fighting coronavirus misinformation and disinformation. *Center for American Progress* (2020). https://www.americanprogress.org/article/fighting-coronavirus-misinformation-disinformation/
[38] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. ArnetMiner: Extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. Association for Computing Machinery, New York, NY, 990–998. https://doi.org/10.1145/1401890.1402008
[39] Samia Tasnim, Md Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on COVID-19 in social media. *J. Prevent. Med. Public Health* 53, 3 (2020), 171–174.
[40] Craig Timberg. 2021. Facebook Made Big Mistake in Data it Provided to Researchers, Undermining Academic Work. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/
[41] Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. 2011. The anatomy of the facebook social graph. Retrieved from https://arXiv:1111.4503
[42] Piet Van Mieghem, Huijuan Wang, Xin Ge, Siyu Tang, and Fernando A. Kuipers. 2010. Influence of assortativity and degree-preserving rewiring on the spectra of networks. *Eur. Phys. J. B* 76, 4 (2010), 643–652.
[43] Wouter Vollenbroek, Sjoerd De Vries, Efthymios Constantinides, and Piet Kommers. 2014. Identification of influence in social media communities. *Int. J. Web Based Commun.* 10, 3 (2014), 280–297.
[44] Yun-Bei Zhuang, Zhi-Hong Li, and Yun-Jing Zhuang. 2021. Identification of influencers in online social networks: Measuring influence considering multidimensional factors exploration. *Heliyon* 7, 4 (2021).
[45] Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. 2018. Follow the "mastodon": Structure and evolution of a decentralized online social network. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12. 541–550.

Received 1 March 2023; revised 8 July 2024; accepted 11 July 2024