Quantifying Uncertainty in Node Feature Prediction of Large Social Networks

Cuneyt G. Akcora
University of Texas at Dallas
800 West Campbell Rd.
Richardson, Texas 75080
cuneyt.akcora@utdallas.edu

Yulia R. Gel
University of Texas at Dallas
800 West Campbell Rd.
Richardson, Texas 75080
ygl@utdallas.edu

Murat Kantarcioglu
University of Texas at Dallas
800 West Campbell Rd.
Richardson, Texas 75080
muratk@utdallas.edu

Vyacheslav Lyubchich
University of Maryland
2020 Horns Point Rd.
Cambridge, Maryland 21613
lyubchich@umces.edu

Bhavani Thuraisingham
University of Texas at Dallas
800 West Campbell Rd.
Richardson, Texas 75080
bxt043000@utdallas.edu

ABSTRACT

In recent years, as online social networks continue to grow in size, estimating node features such as sociodemographics, preferences and health status in a scalable and reliable way has become a primary research direction in social network mining.

To our knowledge, this paper is the first work that provides an insight into quantification and propagation of uncertainty in feature graph mining. That is, this paper sheds light on key questions in social network analysis such as: given a limited availability of social network data, how much data should be queried from the network, and which node features can be estimated reliably? More importantly, how can we evaluate the uncertainty of our estimators? To address these challenges, we propose a novel bootstrap method for uncertainty analysis of node feature prediction in social network mining, derive its asymptotic properties, and demonstrate its effectiveness with extensive experiments.

CCS CONCEPTS

•Mathematics of computing →Random graphs; Graph algorithms; Bootstrapping; *Nonparametric statistics;* **•Computing methodologies →**Natural language processing; Feature selection;

KEYWORDS

Node feature prediction, bootstrap, sampling, uncertainty quantification

ACM Reference format:
Cuneyt G. Akcora, Yulia R. Gel, Murat Kantarcioglu, Vyacheslav Lyubchich, and Bhavani Thuraisingham. 2018. Quantifying Uncertainty in Node Feature Prediction of Large Social Networks. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, London, United Kingdom, August 2018 (KDD'18),* 9 pages.

KDD'18, London, United Kingdom

DOI: 10.1145/nnnnnnn.nnnnnnn

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent advances in social network analysis allow us to build scalable and efficient models for estimating many global statistics of large complex networks, ranging from node centralities [27] to degrees of separation among nodes [3]. At the same time, many scenarios require estimating such properties for a subnetwork only, where we may not know all the nodes belonging to the network. For example, to study social interaction patterns of drug addicts in drug abuse recovery and online drug dealing, we can start by surveying a few known drug addicts and then move over the network, while manually classifying nodes as drug abusers or not. Interaction patterns of "hard-to-reach" populations (e.g., drug addicts) are of crucial importance for solving a slew of problems, including drug abuse and crime. However, such patterns are often studied using convenience or haphazard samples emanating from the known instances (e.g., using the samples derived from known drug addicts). This results in a set of (manually) selected nodes, with inadequate information on how well the sampled nodes represent the population. Even when a large number of nodes (i.e., close to the population size) can be sampled, accessing the data can be prohibitive in terms of time, resources, and storage. Furthermore, most proprietary networks deny wide scale access to their data through restrictive API rates, and limit the usage of data to only a few approved scenarios.

Despite these challenges, from sensing radical groups and their supporters on Facebook to peer-driven drug abuse prevention campaigns on Twitter, many applications of social network analysis involve constructing subnetworks and then using the data within the subnetworks for population estimates, inference, and feature mining.

Such feature estimation with subnetworks raises a number of important questions that are tightly linked to reliability of any conclusions and decisions based on incomplete or noisy data. That is, are the sampled subnetwork data representative of the target population? If we have to enlarge the subnetwork, how should we do it? Which node features can be estimated confidently and which cannot be, even with large samples? Finally, considering all

these challenges, what do we know about the uncertainty of our estimates and how reliable are they?

Although social network analysis is a vast multi-disciplinary research area at the intersection of computer science, statistics, and social studies (see [1, 33, 40] for recent reviews), and despite the fact that uncertainty quantification is rapidly gaining attention in various facets of data analytics, understanding of uncertainty and its dynamics in social network analysis is very limited. The majority of recent studies in graph mining focus on efficiency and scalability of sampling strategies in estimating network statistics, while uncertainty analysis of node feature estimates remains virtually unexplored. The first and, to our knowledge, only steps toward uncertainty quantification on graphs have been taken by [5, 15, 35, 36]. However, these algorithms are restricted to only very small networks and do not assess uncertainty in estimation of node features on very large networks.

In this paper, we address the challenge of uncertainty quantification in node features prediction/estimation with feature graph GraphBoot. This new scalable bootstrap approach yields sampling estimates of network features under limited data availability and, most importantly, *quantifies the estimation uncertainty* of node features. Working with a user-specified level of confidence, our algorithm *can be used to quantify the uncertainty in a wide range of applications involving node feature prediction and associated decision making*. Furthermore, with extensive experiments on real life networks, we show that in many scenarios uncertainty in node feature prediction on networks and subnetworks can be reliably assessed by only observing a small fraction of the network data.

Our contributions can be summarized as follows:

- GraphBoot is the first scalable and computationally efficient method to *quantify uncertainty* and its dynamics in *node feature* analysis of large social networks. GraphBoot offers a speed-up of up to two orders of magnitude compared to existing methods.
- We derive theoretical properties of GraphBoot and illustrate its utility for node feature analysis on a wide range of synthetic and real world large networks of up to 4 million nodes. Our results indicate that node features can be predicted from as few as 5% of all nodes in most of the cases.
- We report results of our bootstrapped approach on synthetic and real world networks, and show that the new bootstrap approach provides the most competitive performance in quantifying uncertainty in node feature mining, yielding the most calibrated and sharp confidence intervals.
- We propose a new information saturation criterion that allows quantifying how much information is needed in social network analysis. With this new criterion, we show that the decision to increase or limit the queried data can be made efficiently and objectively.
- We apply GraphBoot in mental health study on Twitter. The case study results suggest a high utility of our approach to assess the uncertainty in estimates of users' features associated with depression.

The paper is structured as follows. Relevant work is discussed in Section 2. Section 3 defines the problem and provides background on graphs and validation metrics for uncertainty quantification. Section 4 states the algorithms of GraphBoot. Section 5 shows GraphBoot's performance in synthetic and real life networks. The paper is concluded with discussion in Section 6.

2 RELATED WORK

In the context of this project, we outline two primary relevant research fields, namely, i) network analysis with uncertainty quantification on graphs, and ii) mental health studies on online social networks.

Network analysis and uncertainty quantification. Network analysis has been widely studied in statistics, machine learning, and social sciences. The first results go back to the 1960s in a context of social network studies (see, for instance, [16, 17]). For recent overviews on modeling, analysis and mining of complex networks see [2, 32, 33].

Despite these early works, almost nothing is known on how to evaluate estimation uncertainties in network mining, that is, how to obtain reliable sampling standard errors and confidence intervals, *without relying on extensive, costly and practically infeasible simple random sampling (SRS)*. Conventional routes for statistical inference rely on SRS of nodes and either on subsequent application of the central limit theorem (CLT), which results in normal-based confidence intervals, or on resampling of nodes in SRS and the associated Efron confidence intervals [12, 28, 38]. However, SRS of *many independently sampled* nodes is simply impossible in many applications (consider groups of drug addicts, high HIV risk groups, or supporters of ISIS on Facebook).

In their seminal paper, Snijders and Borgatti [35] aim to address these challenges by quantifying uncertainty in graph mining with a data-driven bootstrap. However, their algorithm allows to assess uncertainty only in estimation of densities of small networks and under the assumptions that the whole network is available upfront and the observed network data are error-free, which make the algorithm prohibitive in terms of computational resources, data storage, and data access. Recently, [15, 36] proposed a nonparametric bootstrap method that allows to reliably quantify uncertainty in estimation of a network degree distribution and its functions, while observing only a part of the network. However, the algorithms of [15, 36] do not scale up when the number of nodes increases beyond 10,000 nodes. Finally, [5] proposed a subsampling algorithm for subgraph sampling that is applicable only to exchangeable graphs of high density. Hence, to our knowledge, there exists no algorithm to reliably quantify uncertainty in estimation of node features that is computationally efficient and feasible for large social network analytics. The proposed algorithm GraphBoot is the first attempt in this direction.

Mental health. In our research, we use mental health as an example of application domains. In this field, data from dedicated subgroups of social networks, such as r/SuicideWatch data from Reddit.com in [23], have been mined to analyze user behavior. In addition, crowd-sourcing [10] and manual selections [19] have been used to locate mental health related users. In GraphBoot, we use a combination of manual labels and machine learning to first locate and then expand the set of mental health related accounts. Compared to earlier works [10] that poll users to create labeled data,

GraphBoot starts with a limited number of seeds and reaches more users with fewer labels.

On the Twitter network, studies by De Choudhury et al. [9, 11] have identified prominent features of depressed users' accounts. In the GraphBoot estimates, we use five of these features (e.g., usage of words related to depression treatment) in Section 5.3. In these studies, users are primarily identified with their mental health related word usages. Hwang et al. [20] mine the usage of 14 stigmatizing words, such as "crazy" and "insane", in terms of the senses they are used in. In a similar approach, Harman et al. [19] mine usage of word groups (e.g., anger and swear words) in tweets, and present box plots for frequency of these groups for a variety of mental illnesses.

In these works, the user features that are deemed useful are manually chosen by domain experts. This approach cannot scale up to large networks. Instead, GraphBoot can automatically identify which features should be employed. Furthermore, GraphBoot can give uncertainty estimates for each feature.

3 PROBLEM STATEMENT AND PRELIMINARIES

Suppose that we observe a subset of a large online social network and aim to answer questions of social importance, such as: What types of drugs are primarily used by teenagers? By sampling relevant node features (e.g., age, gender, type and frequency of drugs used) from a network, we can make an inference, and test hypotheses about the whole population based only on the available subnetwork data.

Our analytical solutions are motivated by the following questions: How confident are we in any of our conclusions about the population of teenagers? How high is the sampling error and how biased is the estimate? How large the sampled subnetwork should be to derive a reliable conclusion?

Here we define key concepts that appear throughout the paper.

Networks. We consider a graph $\mathcal{G} = (N, E, F)$, with a set of nodes N, a set of undirected and non-weighted edges E, and a map of feature values $F = \{f_i\}_{i=1}^{|N|}$. That is, each node $n \in N$ is equipped with a feature $f_n \in F$ and $f_1, \ldots, f_{|N|}$ are independent random variables that can be either discrete (e.g., type of drugs) or continuous (e.g., frequency of drug use). Immediate neighbors of a node n are denoted with Γ_n. Let G_n be an induced subgraph of $\mathcal{G}$ that results from sampling on $\mathcal{G}$, and let F_n be the associated map of feature values.

Formulae for uncertainty quantification on node features. Let $\alpha \in (0, 1)$ be a given significance level and $\theta(F)$ be a statistical parameter on a network (e.g., quantiles of features F), and $\hat{\theta}(F_n)$ be its empirical observed counterpart based on F_n. Then, the problem statement can be mathematically formalized as follows. What do we know on $\Pr\{|\theta(F) - \hat{\theta}(F_n)| \geqslant \epsilon\}$, for a given $\epsilon > 0$ and sample size n? Can we construct a reliable $(1-\alpha)100\%$-confidence interval (CI) for a parameter of network features $\theta(F)$? That is, can we find lower L_n and upper U_n bounds such that $\Pr\{\theta(F) \in [L_n, U_n]\} = 1 - \alpha$?

Seeds and waves. Let n_s be a node drawn with simple random sampling from N. This node n_s is called a *seed* as it acts as a starting point for a sample from the network. A *wave* $w_l(s)$ around the seed node n_s is the union set of all nodes and edges that can be reached from n_s by a path p of length $|p| \leqslant l$, where $l \in \mathbb{N}^0$. Thus, wave $w_0(s)$ contains only the seed itself.

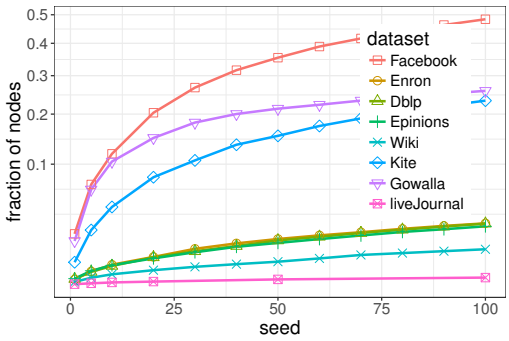


Figure 1: Fraction of nodes visited during sampling.

Embeddedness [18]. We use node embeddedness to quantify how a node's neighborhood overlaps with those of its neighbors. The embeddedness of a node n is $Emb(n) = 1/|\Gamma_n| \sum_{m \in \Gamma_n} |\Gamma_n \cap \Gamma_m|/|\Gamma_n \cup \Gamma_m|$. Embeddedness of a node with degree $d = 0$ is zero.

Confidence intervals and their validation metrics. Let $A = [L_n^A, U_n^A]$ and $B = [L_n^B, U_n^B]$ be two competing $(1-\alpha)100\%$-confidence intervals for the network feature parameter $\theta(F)$. Suppose that over a set of Monte Carlo experiments, $\Pr\{\theta(F) \in [L_n^A, U_n^A]\} = 1 - \alpha_A$ and $\Pr\{\theta(F) \in [L_n^B, U_n^B]\} = 1 - \alpha_B$. Then $1 - \alpha_A$ and $1 - \alpha_B$ are called empirical coverages, and we prefer the *calibrated* confidence interval i.e., with the coverage closest to the nominal level of $1 - \alpha$. Among two intervals, where the first confidence interval under-covers $\theta(F)$ (i.e., $1-\alpha_A < 1-\alpha$), and the second confidence interval over-covers $\theta(F)$ (i.e., $1 - \alpha_B > 1 - \alpha$), we prefer the *conservative*, or over-covering confidence interval B. Furthermore, among A and B with similar coverages, we prefer a confidence interval with a shorter length. Such a preferred confidence interval is called *sharp*. Hence, to compare A and B, we introduce the relative sharpness (RS) criterion

$$RS = \theta(F)^{-1}(\{U_n^B - L_n^B\} - \{U_n^A - L_n^A\})100\%, \quad (1)$$

which yields a relative gain or loss of using confidence interval B over A. Positive values of RS implies that B is sharper (shorter), whereas negative values mean the opposite.

4 NODE FEATURE BOOTSTRAP ALGORITHMS

Sampling-resampling. We employ snowball-like sampling of the network features in parallel around multiple seeds simultaneously. We sample nodes around seeds by the Algorithm 1. The distinguishing characteristic of Alg. 1 compared with the Labeled Snowball with Multiple Inclusions (LSMI, [15, 36]) is that SFINKS collects feature information from the nodes. Similar to LSMI, SFINKS does not reuse edges: each edge can be used in the sampling process only once (used edges are removed, see line 17 of SFINKS).

Algorithmic complexity. In the graph $\mathcal{G} = (N, E, F)$, SFINKS (Alg. 1) chooses m seeds, and for each seed moves on the graph

input : Network $\mathcal{G} = (N, E, F)$; # seeds m, $m \ll |N|$; # waves d.
output: Approximation of sampling probabilities for bootstrap, $\pi^{(\leqslant d),*}$, and two feature lists: of seeds, $\mathcal{L}_s$, & waves, $\mathcal{L}_d$.

$S : Set \leftarrow$ Sample m seeds randomly without replacement from N;
$\mathcal{L}_s : List \leftarrow$ feature values of S;
$\mathcal{L}_d : List \leftarrow \{\}$;
$\pi^{(\leqslant d),*} \leftarrow \{\}$;
$S_0 : Set \leftarrow S$;
$w \leftarrow 1$;
while $w \leqslant d$ **do**
  $N' : Set \leftarrow \{\}$;
  $E' : Set \leftarrow \{\}$;
  foreach *node* $n \in S_{w-1}$ **do**
    foreach *edge* $e \in E | e = \langle n, n^* \rangle$ **do**
      $\mathcal{L}_d \leftarrow \mathcal{L}_d \cup f_{n^*}$;
      $N' \leftarrow N' \cup n^*$;
      $E' \leftarrow E' \cup e$;
  $E \leftarrow E \setminus E'$;
  $S_w \leftarrow N'$;
  $\pi^{(\leqslant d),*} \leftarrow \pi^{(\leqslant d),*} \cup |L_d|/|N|$;
  $w++$;
return $\pi^{(\leqslant d),*}$, $\mathcal{L}_s$, and $\mathcal{L}_d$

Algorithm 1: SFINKS: Sampling Features In NetworKS.

in a breadth first fashion, while deleting already used nodes, until it reaches all nodes within d waves. Hence, SFINKS compensates for discovering the same nodes multiple times, and as a result SFINKS both minimizes sample bias and speeds up the discovery process. For example, with 2 waves and 10–100 seeds, SFINKS has a speedup of 13–175 times over [15, 36] and standard snowball designs. Indeed, for $k = |E|/|N|$ average neighbors for each seed, SFINKS is $O(m \times k^w) = O(k^w) = O(|N|^w)$, when $m \ll |N|$ and $k_{max} = (|N| - 1)/2$. As we show in Section 5, in the real world graphs, $k \ll |N|$ and w values as low as 2 suffice. Hence the complexity of SFINKS is reduced to a very scalable subquadratic form of $o(|N|^2)$.

Fig. 2 shows a toy network ($|N| = 23$) with a structured network sample called *patch* (shaded area) of $m = 2$ seeds and $d = 3$ waves.

Seeds $S = \{1, 2\}$ are selected with simple random sampling from nodes. Node features in this case are node degree, gender, state, type or number of drugs used. Considering the number of drugs used, $\mathcal{L}_s = \{1, 2\}$. By following edges emanating from S, neighborhoods of higher orders are located and information on the nodes' features is recorded. The nodes discovered at each step of growing the patch in Fig. 2 are as follows: $S_1 = \{3, 4, 5, 6\}$, $S_2 = \{7, 7, 8, 9, 10\}$, and $S_3 = \{7, 8, 9, 10, 11\}$. The resulting $\mathcal{L}_3 = \{2, 1, 0, 3, 1, 1, 1, 1, 1, 1, 1, 1, 0\}$.

The GraphBoot algorithm (Alg. 2) creates point estimates and bootstrap confidence intervals of the feature values, using the samples obtained with node sampling. Since probabilities of a node to be included into a patch are different for $\mathcal{L}_s$ and $\mathcal{L}_d$, separate bootstrapping schemes are employed:

- Inclusion probabilities of elements in $\mathcal{L}_s$ are all $|S|/|N|$; use resampling with replacement.
- Inclusion probabilities of elements in $\mathcal{L}_d$ are proportional to their node degrees; account for that using the inverse of the degrees in resampling with replacement.

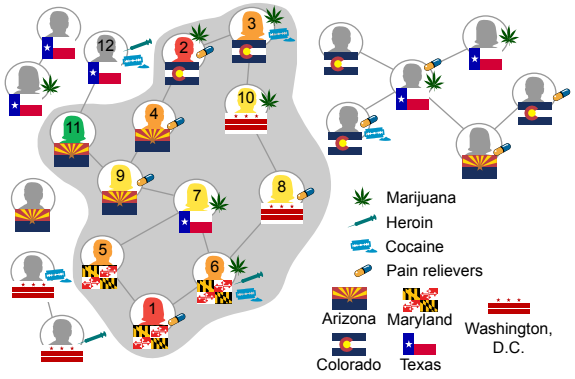


Figure 2: A patch (shaded area) around persons 1 and 2 in a network of $|N| = 23$ people. Starting from the seeds 1 and 2 simultaneously, nodes 3, 4, 5, and 6 are discovered in the first wave. Node 11 can only be discovered at wave 4.

After all b bootstrapped values are obtained, we calculate empirical quantiles from the bootstrap distribution stored in the vector I. For example, by bootstrapping the number of drugs used by a sampled person in Fig. 2 $b = 1000$ times, we obtain a 95% BCI of (0.8, 1.4) for the average value of this feature.

input : SFINKS objects: estimated sampling probabilities $\pi^{(\leqslant d),*}$, $\mathcal{L}_s$, and $\mathcal{L}_d$; # bootstrap samples b, confidence level $1 - \alpha$.
output: Bootstrap confidence interval BCI.

$I : List \leftarrow []$;
for 1 *to* b **do**
  initialize $\mathcal{M} : Map$ to 0 for feature values from $\mathcal{L}_s$;
  for 1 *to* $|\mathcal{L}_s|$ **do**
    Sample one f_n^* from $\mathcal{L}_s$ randomly with replacement;
    $\mathcal{M}_x++$;
  initialize $\mathcal{N} : Map$ to 0 for feature values from $\mathcal{L}_d$;
  for 1 *to* $|\mathcal{L}_d|$ **do**
    Sample f_n^* from $\mathcal{L}_d$ with replacement, with weight proportional to $\pi^{(\leqslant w),*}$, where $w \in \{1, \ldots, d\}$ or approximations thereof by reciprocal degree;
    $\mathcal{N}_x++$;
  $t^*_{\bar{F}} = \dfrac{\sum_{f_n^* \in \mathcal{M}} f_n^*/\pi_n^{(0),*} + \sum_{f_n^* \in \mathcal{N}} f_n^*/\pi_n^{(\leqslant d),*}}{\sum_{f_n^* \in \mathcal{M}} 1/\pi_n^{(0),*} + \sum_{f_n^* \in \mathcal{N}} 1/\pi_n^{(\leqslant d),*}}$;
  $I \leftarrow I \cup t^*_{\bar{F}}$;
$I \leftarrow sort(I)$;
$BCI = \left(I_{\lfloor b\alpha/2 \rfloor},\ I_{\lfloor b(1-\alpha/2) \rfloor}\right)$;
return BCI

Algorithm 2: GraphBoot: Bootstrap for graph features

Algorithmic complexity. Algorithm 2 is linear in the size of seed list $\mathcal{L}_s$, and wave list $\mathcal{L}_w$ for b bootstraps from each list, where $|\mathcal{L}_s| + |\mathcal{L}_w| = n$, hence $O(b(|\mathcal{L}_s| + |\mathcal{L}_w|)) = O(bn)$. In practice, we use $b \ll n$, hence the complexity of the GraphBoot becomes $O(n)$.

Consistency of the Estimator. In this section, we will prove the consistency of Algorithm 2 in quantifying the estimation uncertainties. Note that our ultimate goal is *not to estimate* unknown node features F, which is a longstanding problem [2, 32], but to *reliably quantify uncertainties in estimation* of F.

Let S be the set of seeds, S_1 be the set of immediate neighbors of S (i.e., the first wave), and S_i be the set of immediate neighbors of S_{i-1} (i.e., the i-th wave), where $i \in \mathbb{N}^+$. Let $S^{(\leqslant d)} = S_1 \cup S_2 \cup \ldots \cup S_d$. Let $Q(N)$ be a sampling design on N such that $\pi_n^{(0)} = \Pr\{n \in S\} > 0$, $\forall n \in N$, $\pi_n^{(\leqslant d)} = Pr\{n \in S^d, n \notin S\}$, and $\pi_{nw}^{(\leqslant d)} = Pr\{n \in S^{(\leqslant d)}, w \in S^{(\leqslant d)}\}$.

Given a set of sampled nodes and their features, we propose a modified Hájek estimator for functions of F. For instance, to estimate the mean level $\bar{F}$ of features on N, we propose

$$t_{\bar{F}} = \frac{\sum_{n\in S} f_n/\pi_n^{(0)} + \sum_{S^{(\leqslant d)}} f_n/[\pi^{(\leqslant d)}\hat{\mu}_S^{-1}]}{\sum_{n\in S} 1/\pi_n^{(0)} + \sum_{S^{(\leqslant d)}} 1/[\pi^{(\leqslant d)}\hat{\mu}_S^{-1}]}, \quad (2)$$

where $\hat{\mu}_S = \sum_{n\in S} k_n/|S|$, i.e., an unbiased mean value estimator based on S, and k_n is a degree of a node n. The first term in numerator of (2) corresponds to an estimator of the feature total on seeds and the first term in denominator of (2) is an estimator of unknown $|N|$. The second terms in numerator and denominator of (2) are the corresponding (rescaled) estimators of feature totals and $|N|$ respectively, based on neighbors.

Following the bootstrap algorithm 2, we now construct a bootstrap estimator of a mean feature level on N that could be used to quantify sampling uncertainties in a model-free manner:

$$t^*_{\bar{F}} = \frac{\sum_{n\in S} f^*_n/\pi_n^{(0),*} + \sum_{S^{(\leqslant d)}} f^*_n/\pi^{(\leqslant d),*}}{\sum_{n\in S} 1/\pi_n^{(0),*} + \sum_{S^{(\leqslant d)}} 1/\pi^{(\leqslant d),*}}. \quad (3)$$

Theorem 4.1. **Consistency of node feature bootstrap.** *Let $S \cup S^{(\leqslant d)}$ be a set of nodes sampled from N, l be the cardinality of $S \cup S^{(\leqslant d)}$, and $f_1, \ldots, f_l$ be node features observed on a set $S \cup S^{(\leqslant d)}$. Then, the limiting distributions of $t_{\bar{F}}$ and $t^*_{\bar{F}}$ are identical. That is, as $|N| \to \infty$, $l \to \infty$ and $l/|N| \to \infty$*

$$\sup_f |\sqrt{|N|}(t_{\bar{F}} - \bar{F}) - \sqrt{|N|}(t^*_{\bar{F}} - E^* t^*_{\bar{F}})| \to \quad 0 \quad (4)$$

in probability.

Proof. Let $R(N)$ be a rejective (also called maximum entropy or Poisson) sampling design on N. Simple random sampling without replacement is a particular case of rejective sampling with equal drawing probabilities [29]. The Hájek and Horvitz–Thompson estimators obtained by rejective sampling designs are known to be asymptotically normally distributed (see [4, 6]). Now let us first consider a distribution of $t_{\bar{F}}$. Note that seed nodes in S are obtained with simple random sampling without replacement, that is, via a rejective or maximum entropy sampling. Hence, the first terms in numerator and denominator of (2), namely, $\sum_{n\in S} f_n/\pi_n^{(0)}$ and $\sum_{n\in S_0} 1/\pi_n^{(0)}$ are both asymptotically normally distributed.

Now let us turn to the second terms in numerator and denominator of (2), which involves sampling probabilities $\pi^{(\leqslant d)}$ on waves $S_1, \ldots, S_d$. In general, probabilities $\pi^{(\leqslant d)}$ for $d > 1$ are unknown, unless G is a tree [34]. However, following [21], if we assume that all neighbors of a node n are included in the sample up to wave $d-1$, then sampling probability $\pi^{(\leqslant d)}(n)$ for a node $n \in N$ can be approximated by a function of its degree k_n. That is,

$$\begin{aligned} \pi^{(\leqslant d)} \approx \pi^{(\leqslant d)}(k_n) &= 1 - \left(1 - \frac{|S^{(\leqslant d)}|}{|N|}\right)^{k_n} \\ &\approx k_n \frac{|S^{(\leqslant d)}|}{|N|}, \end{aligned} \quad (5)$$

where the last term is due to the Taylor approximation of a convergent power series within an open unit circle. Note that even when the assumption that all neighbors of a node n are already included into previous waves does not hold, the bias due to this simplification in the numerator of (2) is corrected by the respective bias in the denominator of (2) (see Section 4.2 of [21] for the detailed discussion).

Given (5), if k_n is concentrated around mean degree of G, drawing probabilities $\pi^{(\leqslant d)}\hat{\mu}_{S_0}^{-1}$ of neighbors in the design $Q(N)$ satisfies an approximation $\left(k_n\hat{\mu}_{S_0}^{-1}|S^{(\leq d)}|\right)/(|N|) \approx |S^{(\leq d)}|/|N|$. Hence, the divergence of a design $Q(N)$ from a rejective design $R(N)$ (see [4]) $D(Q(N)||R(N)) = \sum_n Q(n)\log\left[Q(n)/R(n)\right] \to 0$. That is, the design $Q(N)$ can be approximated by a high entropy design. As a result, Theorem 4.2 of [6] implies that $\sum_{S^{(\leqslant d)}} f_n/[\pi^{(\leqslant d)}\hat{\mu}_S^{-1}]$ and $\sum_{S^{(\leqslant d)}} 1/[\pi^{(\leqslant d)}\hat{\mu}_S^{-1}]$ are asymptotically normally distributed. Hence, limiting distributions of all four summarands in (2) is normal, and invoking a delta-method implies that $t_{\bar{F}}$ is asymptotically normally distributed [38].

Now let us turn to a limiting conditional bootstrap distribution of $t^*_{\bar{F}}$. Derivations of limiting conditional bootstrap distributions of the first terms in the numerator and denominator of (3) mirrors the case of (2). In turn, neighbors in the bootstrap Algorithm 2 are resampled with probabilities proportional to their reciprocal degree $1/k_n$. Hence, (5) implies that $\pi^{(\leqslant d),*} \approx |S^{(\leqslant d)}|/|N|$. The remaining derivations for (3) mirrors the case of (2).

Hence, both limiting distribution $t_{\bar{F}}$ and limiting conditional distribution of $t^*_{\bar{F}}$, given G, coincides, which concludes the proof of (4).□

Gini elbow criterion: Can bootstrap help to decide how much to sample objectively? A set of bootstrap confidence intervals (BCI) in Alg. 2 for different numbers of seeds and waves contains a wealth of information on structural properties of $\mathcal{G}$, and can be used to assess a level of sampling uncertainty on $\mathcal{G}$. Intuitively, if the node features are estimated sufficiently well, then the bootstrap distributions for similar numbers of seeds and waves shall not be too different from each other, that is, a certain level of information saturation is reached when increasing a number of seeds and waves yields incremental or no improvement. Hence, we can study distributional properties of BCIs, and we start from homogeneity analysis of BCI lengths with the Gini index.

The Gini index g is a measure of statistical heterogeneity [14]. Formally, let $x_1, \ldots, x_n$ be features associated with n units. Then the Gini index (GI) is defined as $g = \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|/(n\sum_{i=1}^n x_i)$. GI is widely used in economics to measure income inequality. In particular, g ranges from 0, when all individual have equal income, to 1, when all income is assigned to a single person. GI is also used in network studies to evaluate sparsity and centrality (for reviews see [30] and references therein).

Table 1: Observed coverage (%) of 95% confidence intervals (width is in parentheses) on simulated data

Interval	Features f_i $(i = 1, \ldots, \lvert N\rvert)$		
	$k_i^2 + Pois(1)$	$k_i^2 + Pois(4)$	$k_i + N(0,1)$
GraphBoot	98.7 (4.88)	98.7 (4.90)	94.4 (0.58)
SRS normal	87.4 (9.26)	87.6 (9.32)	93.7 (1.17)
SRS bootstrap	88.0 (8.97)	88.1 (9.03)	93.4 (1.15)

In a bootstrap context, we expect that the GI for BCI lengths will decrease as sampling variability decreases. Hence, the optimal number of seeds and waves in the graph sampling framework can be determined by the minimum of GIs for BCI lengths, e.g., via the elbow plot (see Fig. 5). At the same time, if GI does not decrease, it suggests that sampling variability is high, and the information extraction for this node feature is limited or even impossible. Note that with this measure we do not account for estimation bias but rather focus on intrinsic variability of sampling that necessarily needs to be low for reliable estimates of node features.

5 EXPERIMENTS

We use GraphBoot to quantify uncertainty of node feature estimations in simulated and real life networks where the ground truth is known (Sections 5.1 and 5.2) and unknown (Section 5.3).

The Scala/Spark implementation of the algorithms is available on Github[1].

5.1 Simulated Networks.

We use 10^4 simulated networks of order $|N| = 10^4$ and polylogarithmic degree distribution with parameters 0.01 and 2. We simulate features using linear and non-linear functions of the node degrees with added noise. To obtain CIs for the mean feature value, we apply SFINKS and GraphBoot, with reciprocal degree weights, based on 20 seeds, 2 waves, and $b = 500$. As a competing approach, we use simple random sampling of 50 nodes to construct normal and non-parametric bootstrap confidence intervals [12, 37]. Table 1 shows the performance of GraphBoot in quantifying the uncertainty in feature averages. Compared to the two alternatives, the new feature bootstrap approach delivers coverage closest to the nominal level 95% and yields the sharpest confidence intervals.

5.2 Real Life Networks

We used GraphBoot on eight datasets of varying order, sparsity and embeddedness (Table 2) to quantify uncertainty in node feature estimation. We start from average degrees as a target statistical parameter. The LiveJournal dataset is the biggest, with 4M nodes and 35M edges [24]. The Facebook [39] dataset contains 817K edges among 63K New Orleans network users. The DBLP [26] network is an undirected co-authorship network. Gowalla [8] and BrightKite [8] are undirected location based social networks.

In directed networks, we created edges when both nodes have an outgoing edge to each other. The LiveJournal edges are friendships, whereas in the Wikipedia network [25], an edge is created among nodes who have edited each other's talk page. In the Enron [22] network, we create an edge between two nodes who have shared emails. The Epinions [31] network is the only signed network in our dataset, where an edge between two users indicates their trust for each other.

[1] www.github.com/cakcora/GraphBoot

Table 2: Summary statistics for number of nodes, edges, mean degree, fraction of sampled nodes at 20 seeds and 2 waves (Cov@20) and average node embeddedness (Emb).

	$\lvert N\rvert$	$\lvert E\rvert$	μ	Cov@20	Emb
LiveJournal	4M	35M	13.40	8e-5	0.088
Dblp	317K	1M	6.62	0.0097	0.305
Gowalla	196K	950K	9.60	0.1481	0.073
Wiki	94K	361K	7.60	0.0015	0.005
FB	63K	817K	25.64	0.2031	0.071
Kite	58K	214K	7.30	0.0795	0.049
Enron	36K	183K	10.02	0.0054	0.192
Epinions	31K	103K	6.63	0.0048	0.046

Estimation. Figure 3 shows GraphBoot wave 2 results in degree estimation on undirected and directed networks[2] with $b = 1000$. All results are averages of 50 runs. As networks have different mean degrees, we present the GraphBoot degree estimate $\hat{x}$ in a relative form, $\hat{x}/\mu$, where μ is the true mean degree given in Table 2.

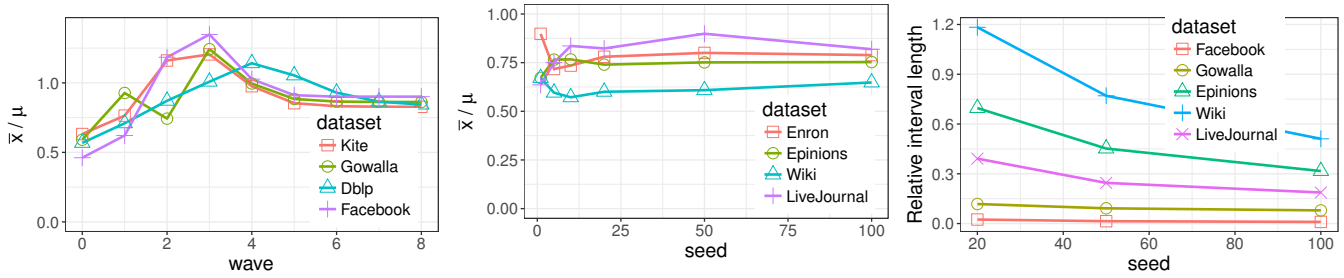
Figure 3 shows that undirected networks can be efficiently sampled by GraphBoot with 2 or 3 waves for average degrees. The proportion of used nodes depends on the distribution of degrees and embeddeddness in a network. Due to this low proportion, directed networks in Fig. 3b have lower estimates. Thus, sampling in these networks must start with more seeds or continue for more waves. Based on Fig. 3c, GraphBoot confidence intervals are much shorter for undirected networks than for the directed ones.

How would a conventional normal based inference perform? Fig. 4 presents a comparison of normal-based and GraphBoot bootstrap confidence intervals in terms of relative sharpness (RS). Positive RS implies a gain of GraphBoot vs. normal-based CI (in %), reverse is true for negative RS. In most of the cases, RS is positive, i.e., GraphBoot bootstrap CIs outperform normal-based CIs. The improvement of GraphBoot is the most overwhelming in Epinions (42.4%). The only case where GraphBoot delivers a less sharp CI than a normal-based approach is for Kite, however, the loss is minor (0.8%).

These results show that GraphBoot delivers more accurate (sharper) confidence intervals, and hence, is a more reliable and preferred method for uncertainty quantification in network mining. Another aspect is that a normality-based inference would require querying the feature values of too many nodes, whereas GraphBoot requires a very limited number of seeds neighborhoods around them.

How does dispersion in estimations indicate an uncertainty? Fig. 5 shows the Gini index (GI) of BCI lengths for varying waves and seeds. As expected, increasing number of seeds leads to lower sampling variability and, as a result, to lower GI. Figures 5b and 5a can be used as an elbow plot for selecting optimal number of seeds

[2] In the rest of this paper, we will omit some datasets in figures to have visually discernible results.



(a) Average degree estimates on undirected networks for 20 seeds and varying waves.

(b) Average degree estimates on directed networks for 2 waves and varying number of seeds.

(c) Relative confidence interval length, $|BCI|/\hat{x}$, for GraphBoot wave 2 results.

Figure 3: GraphBoot relative degree estimates on networks, and confidence interval lengths for the estimates.

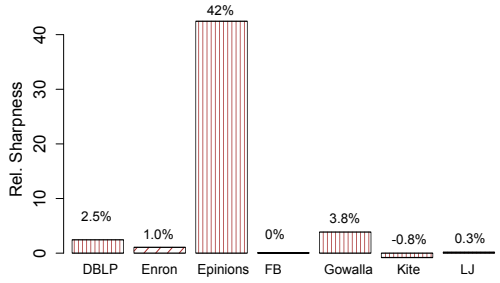


Figure 4: Relative Sharpness (RS) comparison of GraphBoot bootstrap and normal-based confidence intervals. Positive values of RS implies a gain of GraphBoot bootstrap CI vs. normal-based CI (in %), and reverse. In all cases, except of Enron, all CIs contain the true population parameter.

in a sampling design for a particular node feature, so that adding extra seeds does not reduce sampling dispersion. For instance, in Fig. 5b, all networks reach low GI around 25 seeds, and hence we can conclude that 25 seeds is a reasonable sampling size. Undirected networks tend to deliver low GI even with few seeds, which can be attributed to degree assortativity or bigger diameters.

Fig. 5c shows a counterpart study of GI against varying numbers of waves. While all GI also tend to decrease as the number of waves increases, the rate of decrease is noticeably different than for a case of GI as a function of seeds. For instance, some networks, such as DBLP and Facebook, exhibit a rapid decrease of GI already at wave 1 – thereby, implying that already wave 1 contains a large portion of information that can be extracted by sampling. In contrast, other undirected networks, such Kite and Gowalla, show a noticeably slower rate of decrease in GI, suggesting that reliable estimation requires a higher number of waves. Generally, the Gowalla network appears to be an outlier in both Figs. 5a and 5c, with a very slow decay of sampling variability.

Furthermore, Fig. 5c indicates a clear distinction between directed and undirected networks, that is, the GIs of directed networks tend to be higher especially for smaller waves. Intuitively, this implies that neighbors of a node are not sufficiently similar to the node and to achieve higher accuracy, we need more data.

GIs of GraphBoot degree estimations on real life networks show that depending on the network coverage and network type, GraphBoot can provide degree estimates with as few as 25 seeds and 2 waves. Furthermore, by observing how the inequality of estimated confidence interval lengths change, GraphBoot can continue to sample more data until a predefined level is reached for a given GI.

5.3 Quantifying Uncertainty without Ground Truth

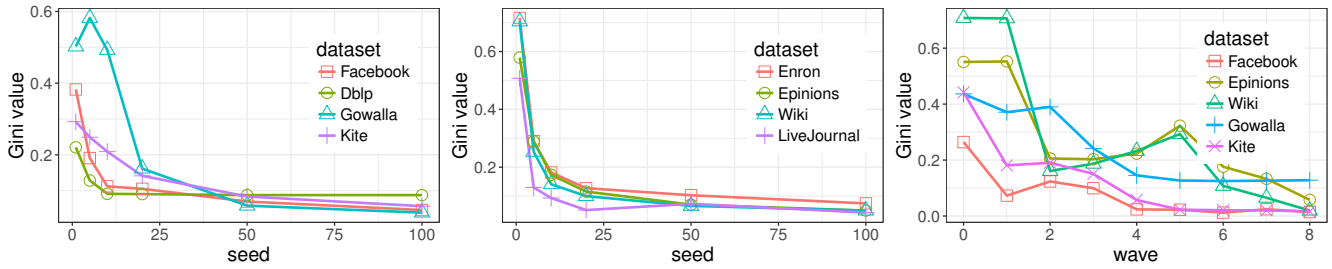
In many real life scenarios, the network (e.g., Twitter) consists of different types of nodes, and the estimated feature may not exist for all the nodes. Furthermore, the estimated feature may be relevant to a subset of nodes only. Consider the case where we estimate the age of first time car-buyers in the USA. We first need to classify users into nationalities, and sample on the US nationals only. This approach involves using a subset of nodes on the network rather than the full network. In this section, we will describe how GraphBoot can quantify the quality of a estimations in such a scenario.

As an illustrative example, we chose to carry out a case study of mental health research on the Twitter network, because i) mental health is widely discussed on Twitter by many users [23], and ii) there are well known mental health related features [10] to sample with GraphBoot.

Sampling a mental health subnetwork. Our node sampling starts with randomly selecting 13 accounts that tweet about mental health; 6 of these are organizational accounts that are used to create awareness on mental health issues, whereas the remaining 7 are users that experience mental health issues themselves. *Note that for this subnetwork sampling, we can employ any community detection method. In this experiment, the quality of the subnetwork is not our primary concern.*

Our overall goal is to locate the subnetwork with as few manually given labels as possible. We employed a two phased approach to classify Twitter users as mental health related (i.e., MentalHealth+) or not mental health related (i.e., MentalHealth-). Below, we will outline our approach.

Phase 0. We queried Twitter APIs for the last 20 tweets of the 13 seeds, and used these tweets to create seed feature vectors. Each feature in the vector is a word, and feature value is the number of times the word appears in the tweets of the seed. We labeled these seed vectors as mental health related.



(a) GraphBoot wave 2 GIs of undirected networks for varying seeds.

(b) GraphBoot wave 2 GIs of directed networks for varying seeds.

(c) GraphBoot 20 seed GIs while varying waves for select datasets.

Figure 5: GIs of confidence interval lengths from GraphBoot degree estimations on real life networks.

Table 3: Classification numbers in phases.

Ph.	Labels	Found	Classi.	MHealth+
0	MHealth+: 13 MHealth-: 26	1.8K	1.7K	258 (14%)
1	MHealth+: 51 MHealth-: 143	40K	30K	1.1K (3.8%)

Table 4: Top 7 U.S. states with the highest numbers of classified accounts in phase 1.

State	CA	TX	NY	FL	GE	IL	PE
MHealth+	72	42	47	28	11	18	20
MHealth-	1.3K	911	827	543	372	314	301

In order to train the model also with MentalHealth- vectors, we downloaded the Twitter data set of Cheng et al. [7] and created feature vectors of users in a similar way. These vectors were labeled as non-mental health related. As 36% of Twitter users scored positive for depression in [10], we decided to combine 13 MentalHealth+ vectors with 26 randomly selected MentalHealth+ vectors.

Out of 1,895 wave 1 nodes that were found, 1,784 of them had public tweets. MentalHealth vectors were used in a 10 tree Random Forest Classifier to classify these 1,784 Twitter users. In total, the phase 0 model classified 258 accounts as MentalHealth+.

Four Ph.D. students independently classified 197 Twitter users from phase 0 into MentalHealth+, MentalHealth- and "no decision" classes manually. Validation results show an inter-annotator agreement (Fleiss' Kappa [13]) of 0.46, with values: 8 false positives, 33 true positives, 22 false negatives and 91 true negatives. Phase 0 precision of the random forest model is thus 80.4%.

Phase 1. In phase 1, we added the manually labeled vectors to the model input, and trained a new model. The new random forest model was then used to classify neighbors of 258 wave 1 nodes who were labeled as MentalHealth+ in phase 0. Overall, this reduced the number of wave 2 nodes to be classified from 198K to 40K. Table 3 shows the statistics of our model learning. In Phases 0 and 1, 14% and 3.8% of users were classified as mental health related, respectively.

We also used the bio location of Twitter users with Google Location API to assign users to US states. The results are given in Table 4.

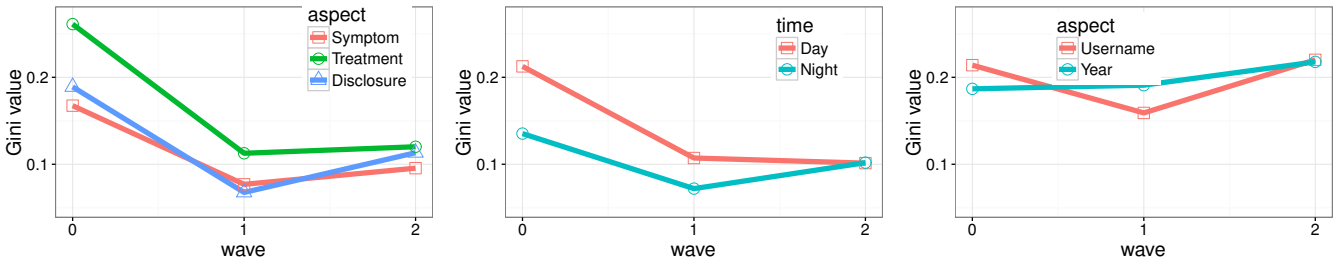
At the end of phase 1, we create a network from MentalHealth+ users where edges among these nodes are learned by querying Twitter again. Including seeds, phase 0 and phase 1 users, our network consists of 1,466 nodes and 4,581 edges.

Sampling for various features. On the created mental health network, we used GraphBoot to estimate five node features from mental health research. Three of these features, Symptom, Treatment and Disclosure are related to the specific groups of words used by users [10]. De Choudhury et al. report that "these words appear with high frequency in the posts from the depression class" of Twitter users. GraphBoot estimates the number of appearances from these groups. Two other statistics from [10], Day and Night, are used to estimate daily posting habits of Twitter users. We formally define these statistics as follows:

Symptom: Number of times Symptom related words, such as anxiety, appear in the tweets of user. **Treatment:** Number of times Treatment related words, such as medication, appear in the tweets. **Disclosure:** Number of times Disclosure related words, such as fun, appear in the tweets. **Day:** From the last 20 tweets of a user, the number of tweets posted before 6AM and after 9PM. **Night:** From the last 20 tweets of a user, the number of tweets posted after 6Am and before 9PM.

GraphBoot wave 2 estimates for word groups are 4.45 for Disclosure, 1.71 for treatment and 2.27 for Symptom. Change of GIs for these groups are given in Fig. 6a. Day and Night estimates of GraphBoot are 15.8 and 3.92, respectively. GI of these time related statistics are given in Fig. 6b. GI changes of both word groups and time aspects show that GraphBoot wave 1 computations have better estimates. Overall, these statistics show low GIs in GraphBoot estimates, and hence can be viewed as estimable statistics for mental health studies.

As a control experiment, we devised two statistics, Username and Year, that have no basis in mental health analysis. In Username, we estimate the average length of a user's screen name on the mental health network. E.g., www.twitter.com/POTUS has a username of length 5. The Year statistic gives the Twitter registration year of a user, such as 2017 for the @POTUS user. GraphBoot estimates for these statistics are 2013.6 for Year and 11.9 for Username. GraphBoot results in Figure 6c show that the GIs for these statistics are high, and even increase at waves 1 and 2.



(a) GIs for Symptom, Treatment and Disclosure aspects. **(b) GIs for time aspects from mental health accounts.** **(c) GIs for Username and Registration Year aspects.**

Figure 6: GIs for Twitter mental health experiments.

Such not diminishing and even deteriorating GIs imply that the used statistics are not reliable; the sampler can stop sampling node features for these statistics.

In the next section, we offer a holistic view on GraphBoot's performance, and outline further directions in feature sampling.

6 CONCLUSION

In this paper, we quantify estimation uncertainties of node features in a variety of large social networks. Our results in degree experiments show that the match between the estimated feature and the network type greatly affects the estimation quality. Undirected networks where edges are created with mutual consent provided better degree estimates, whereas directed networks have inequalities within samples. In both cases, GraphBoot provides adequate quality indications for the sampling process.

In the future we plan to expand GraphBoot to quantifying uncertainty in motif estimation and feature-based anomaly detection.

REFERENCES

[1] Nesreen K Ahmed, Nick Duffield, Jennifer Neville, and Ramana Kompella. Graph sample and hold: A framework for big-graph analytics. In *ACM SIGKDD 2014*. 1446–1455.

[2] N. K Ahmed, N. Duffield, J. Neville, and R. Kompella. Graph sample and hold: A framework for big-graph analytics. In *ACM SIGKDD 2014*. 1446–1455.

[3] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. Four degrees of separation. In *Websci'2012*. 33–42.

[4] Yves G Berger. 1998. Rate of convergence for asymptotic variance of the Horvitz–Thompson estimator. *J. of Stat. Plan. & Infer.* 74, 1 (1998), 149–168.

[5] S. Bhattacharyya and P. J. Bickel. 2015. Subsampling Bootstrap of Count Features of Networks. *Ann. of Stat.* 43, 6 (2015), 2384–2411.

[6] Helene Boistard, Hendrik P Lopuhaä, Anne Ruiz-Gazen, and others. 2012. Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Elect. J. of Stat.* 6 (2012), 1967–1983.

[7] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *ACM CIKM 2010*.

[8] Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *ACM SIGKDD 2011*. 1082–1090.

[9] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Websci'2013*. 47–56.

[10] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media.. In *ICWSM*. 2.

[11] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *ACM CHI 2016*. 2098–2110.

[12] B. Efron and R. J. Tibshirani. 1993. *An introduction to the bootstrap*. Chapman & Hall, Boca Raton.

[13] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psych. Bulletin* 76, 5 (1971), 378.

[14] J. L. Gastwirth. 1972. The estimation of the Lorenz curve and Gini index. *REStat* 54, 3 (1972), 306–316.

[15] Y. R. Gel, V. Lyubchich, and L. L. Ramirez Ramirez. 2017. Bootstrap quantification of estimation uncertainties in network degree distributions. *Sci. Rep.* 7 (2017), 5807.

[16] Leo A Goodman. 1961. Snowball sampling. *Ann. Math. Stat.* (1961), 148–170.

[17] Mark Granovetter. 1976. Network sampling: Some first steps. *Amer. J. of Sociol.* 81, 6 (1976), 1287–1303.

[18] Mark Granovetter. 1985. Economic action and social structure: The problem of embeddedness. *Am. J. of Sociology* 91, 3 (1985), 481–510.

[19] Glen Harman, Mark Coppersmith, and Craig Dredze. 2014. Quantifying mental health signals in Twitter. *ACL 2014* 51 (2014).

[20] Jena D. Hwang and Kristy Hollingshead. Crazy Mad Nutters: The Language of Mental Health. In *CLPsych@NAACL-HLT 2016*. 52–62.

[21] Johannes Illenberger and Gunnar Flötteröd. 2012. Estimating network properties from snowball sampled data. *Social Networks* 34, 4 (2012), 701–711.

[22] Bryan Klimt and Yiming Yang. 2004. The enron corpus: A new dataset for email classification research. In *European Conference on ML*. Springer, 217–226.

[23] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. Detecting changes in suicide content manifested in social media following celebrity suicides. In *ACM HT 2015*. 85–94.

[24] Jérôme Kunegis. Konect: the koblenz network collection. In *ACM WWW 2013*. 1343–1350.

[25] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *ACM SIGCHI 2010*. 1361–1370.

[26] Michael Ley. 2009. DBLP: Some Lessons Learned. *Proc. VLDB Endow.* 2, 2 (Aug. 2009), 1493–1500.

[27] Ahmad Mahmoody, Charalampos E Tsourakakis, and Eli Upfal. 2016. Scalable Betweenness Centrality Maximization via Sampling. *arXiv preprint arXiv:1609.00790* (2016).

[28] D. Pfeffermann and C.R. Rao. 2009. *Sample Surveys: Inference and Analysis*. Elsevier, Amsterdam.

[29] Danny Pfeffermann and Calyampudi Radhakrishna Rao. 2009. *Handbook of Statistics_29A: Sample Surveys: Design, Methods and Applications*. Vol. 29. Elsevier.

[30] Xiangju Qin, Pádraig Cunningham, and Michael Salter-Townshend. 2015. The influence of network structures of Wikipedia discussion pages on the efficiency of WikiProjects. *Social Networks* 43 (2015), 1–15.

[31] Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. 2003. Trust management for the semantic web. In *Int. Semantic Web Conf.* Springer, 351–368.

[32] John Scott and Peter J Carrington. 2011. *The SAGE handbook of social network analysis*. SAGE.

[33] Olivia Simpson, C Seshadhri, and Andrew McGregor. In *ICDM2015*. 979–984.

[34] Tom A B Snijders. 1992. Estimation on the basis of snowball samples: how to weight? *BMS* 36, 1 (1992), 59–70.

[35] T. A. B. Snijders and S. P. Borgatti. 1999. Non-parametric standard errors and tests for network statistics. *Connections* 22, 2 (1999), 61–70.

[36] Mary E Thompson, Lilia Leticia Ramirez Ramirez, Vyacheslav Lyubchich, and Yulia R. Gel. 2016. Using the bootstrap for statistical inference on random graphs. *CJS* 44, 1 (2016), 3–24.

[37] S. K. Thompson. 2012. *Sampling*. Wiley, Hoboken.

[38] A. W. van der Vaart. 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[39] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P. Gummadi. On the Evolution of User Interaction in Facebook. In *ACM SIGCOMM WOSN'09*.

[40] Y. Zhang, E.D. Kolaczyk, and B.D. Spencer. 2015. Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks. *AoAS* 9, 1 (2015), 166–199.