This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2925355, IEEE Transactions on Knowledge and Data Engineering

GraphBoot: Quantifying Uncertainty in Node Feature Learning on Large Networks

Cuneyt Gurcan Akcora, Yulia R. Gel, Murat Kantarcioglu, *Senior Member, IEEE,* Vyacheslav Lyubchich, Bhavani Thuraisingham, *Fellow, IEEE*

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. Although many techniques have been developed for estimating various node features, quantifying uncertainty in such estimations has received little attention. Furthermore, most existing methods study networks parametrically, which limits insights about necessary quantity of queried data, reliable feature estimation, and estimator uncertainty.

Uncertainty quantification is critical for answering key questions, 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 learned reliably? More importantly, how can we evaluate uncertainty of our estimators? Uncertainty quantification is not equivalent to network sampling but constitutes a key complementary concept to sampling and the associated reliability analysis.

To our knowledge, this paper is the first work that sheds light on uncertainty quantification and uncertainty propagation in social network feature mining. We propose a novel non-parametric bootstrap method for uncertainty analysis of node features in social network mining, derive its asymptotic properties, and demonstrate its effectiveness with extensive experiments. Furthermore, we develop a new metric based on dispersion of estimations, enabling analysts to assess how much more information is needed for increasing prediction reliability based on the estimated uncertainty. We demonstrate the effectiveness of our new uncertainty quantification methodology with extensive experiments on real life social networks, and a case study of mental health on Twitter.

Index Terms—Uncertainty quantification, social network, mental health, bootstrap, graph analytics.

1 Introduction

Although social network analysis is a vast multi-disciplinary research area at the intersection of computer science, statistics, and social studies (see [1], [2], [3] for recent reviews), and despite the fact that uncertainty quantification (UQ) is rapidly gaining attention in various facets of data analytics, understanding of uncertainty and its dynamics in social network analysis is very limited.

There exist multiple sources of uncertainty in analysis of complex networks, including but not limited to uncertainty due to node and/or edge sampling, uncertainty due to only partially observed network information, e.g., for the cases of hard-to-reach populations and confidentiality restrictions, uncertainty due to estimation of node, edge, and other network attributes, and uncertainty due to approximation of the true underlying network data generating process by a certain model. In this paper we primarily focus on addressing uncertainty in estimation of node features, which includes uncertainty due to sampling and uncertainty due to the type of the estimation method.

The majority of recent studies in graph mining focuses on efficiency and scalability of sampling strategies in estimating network statistics, while uncertainty analysis of node feature estimates remains virtually unexplored. Furthermore, while uncertainty quantification and sampling are tightly interrelated, it is important to emphasize that *uncertainty quantification shall not be mistaken with sampling*. That is, sampling allows us to estimate network statistics but does not provide an insight on reliability of the obtained estimates. In turn, uncertainty quantification is the way to systematically address *credibility* of the obtained estimates.

UQ becomes even more important for many applications, ranging from finding radical groups and their supporters on Facebook to peer-driven drug abuse prevention campaigns on Twitter, where we have access to limited data. These social network applications involve first constructing subnetworks (e.g., drug abusers) and then using the data within the subnetwork for population estimates (e.g., average age of drug abusers), inference and feature mining (e.g., whether a given Twitter user is a potential drug abuser).

Such node feature estimation with subnetworks raises a number of important questions that are tightly linked to the reliability of any conclusions and decisions based on incomplete or noisy data. That is, are the queried subnetwork data representative of the target population? Do we have to enlarge the subnetwork by collecting more data? Which node features can be estimated confidently and which cannot be, with the given samples? Finally, considering all these challenges, what do we know about the uncertainty of our estimates and how reliable are they?

The pioneering attempt for uncertainty quantification on social networks is due to Snijders et al. [4]. However, uncertainty quantification in conjunction with social network mining still remains scarcely explored. In particular, to our knowledge, only steps toward uncertainty quantification on

- C. G. Akcora, Y. R. Gel, M. Kantarcioglu, and B. Thuraisingham are with the University of Texas at Dallas, Richardson, TX 75080, USA. E-mail: {cuneyt.akcora, ygl, muratk, bhavani.thuraisingham}@utdallas.edu.
- V. Lyubchich is with the University of Maryland Center for Environmental Science, Cambridge, MD 21613, USA. E-mail: lyubchich@umces.edu.

Manuscript received Month dd, 20YY; revised Month dd, 20YY.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2925355, IEEE Transactions on Knowledge and Data Engineering

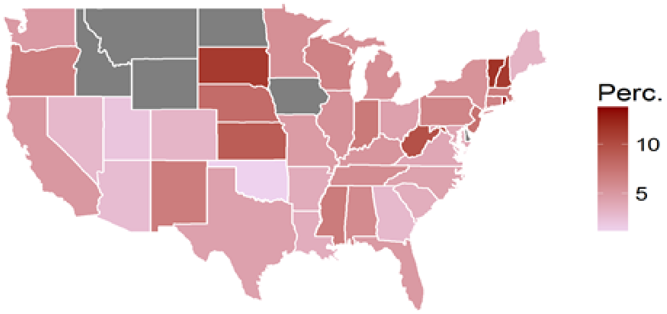


Fig. 1: Ratio of mental health related Twitter accounts in USA. Our non-parametric graph learner framework can learn node features and quantify the uncertainty in its results (gray implies states with inadequate data). This case study is detailed in Sec. 5.3.

social networks have been taken by [5], [6] in a Bayesian framework; [7], [8] study uncertainty quantification for centrality indices in animal and physiological networks, while [9], [10] propose bootstrap for estimators of degree distribution. However, these algorithms are limited to only very small networks, rely on (semi)parametric assumptions that are hard to validate in most real life scenarios, 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 estimation with GraphBoot. This new scalable bootstrap based approach provides 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 uncertainty in a wide range of applications involving node feature learning and associated decision making*. Furthermore, with extensive experiments on real life networks, we show that in many scenarios uncertainty in node feature learning on networks and subnetworks can be reliably assessed by only observing a small fraction of the network data.

The importance of the proposed research can be summarized as follows:

- GraphBoot is the first scalable and computationally efficient *non-parametric 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 bootstrap based approach on synthetic and real world networks, and show that the new bootstrap based approach provides the most competitive performance in quantifying uncertainty in node feature mining, yielding the most calibrated and sharp confidence intervals, comparing to the baseline approaches.
- 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. With very little human supervision, we were able to discover a group of interest that is many times bigger in size than what state of the art mental health research works [11], [12] have been analyzing. The case study results suggest a high utility of our approach in assessing 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 UQ validation metrics. 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

We outline two primary relevant research fields, namely, i) network analysis with Uncertainty Quantification, and ii) mental health studies on online social networks.

Uncertainty quantification in network analysis. 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, [13], [14]). For recent overviews on modeling, analysis and mining of complex networks see [1], [2], [15].

Despite these early works, almost nothing is known on how to evaluate estimation uncertainties in network mining, that is, how to obtain reliable estimates of the sampling errors and confidence intervals for the parameters of interest, without relying on extensive, costly and practically infeasible simple random sampling (SRS).

Conventional statistical inference relies on learning from node data and relies either on applying the central limit theorem, which results in normal distribution based confidence intervals, or on resampling of nodes in simple random sampling and the associated Efron confidence intervals [16], [17], [18]. However, randomly choosing many nodes and learning from their data is not feasible in real life settings. For example, locating and learning from groups of drug addicts, HIV risk groups, or terror supporters on an online social network such as Facebook would require knowing who belong to these groups.

In their seminal paper, Snijders and Borgatti [4] 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, [9], [10] proposed a nonparametric bootstrap method that allows to reliably

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2925355, IEEE Transactions on Knowledge and Data Engineering

quantify uncertainty in estimation of a network degree distribution and its functions, while observing only a part of the network. However, the algorithms of [9], [10] do not scale up when the number of nodes increases beyond 10,000 nodes. In turn, a subsampling algorithm of [19] focuses on subgraph sampling that is applicable only to exchangeable graphs of high density. Finally, [5], [6], [7], [8] emphasized a critical role of uncertainty quantification in analysis of psychological and animal social networks. Among these studies, [5], [6] employed a Bayesian inference for uncertainty quantification of animal social interactions which requires parametric assumptions on a prior distribution. In turn, [7] and [8] utilized a randomization technique, where certain node characteristics, e.g., node centrality measures, are re-sampled or permuted without accounting for network dependence structure among nodes and then are compared against some simulated random networks. In all these cases, the order of the considered networks is less than 1,000 nodes and the whole network is assumed to be fully observed upfront.

Hence, to our knowledge, there exists no algorithm to 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 [20], have been mined to analyze user behavior. In addition, crowdsourcing [21] and manual selections [22] have been used to locate mental health related users. In GraphBoot, we use a combination of manual labels and machine learning algorithms to first locate and then expand the set of mental health related accounts. Compared to earlier works, such as [21], 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. [12], [23] have identified prominent features of depressed users' accounts. In the GraphBoot estimates in Section 5.3, we use five of these features (e.g., usage of words related to depression treatment). In these studies, users are primarily identified with their mental health related word usages. Hwang et al. [11] 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. [22] 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 can be mined. Furthermore, GraphBoot can give uncertainty estimates for each feature.

3 Problem Statement and Preliminaries

Assume that we discover a subnetwork of a large online social network and aim to answer questions of social importance, such as, what types of drugs are primarily used by users in depression? By extracting relevant node features (e.g., age, gender, type and frequency of drugs used) from a network, we can make an inference, and test our hypotheses about the whole population based only on the available subnetwork data.

Objectives. Our analytical solutions are motivated by the following questions: How confident are we in any of our conclusions about the population of users? How big is the learning error and how biased is the estimate? How large the discovered subnetwork should be to derive a reliable conclusion?

Here we define key network concepts and validation metrics that appear throughout the paper.

Network. We consider a network $\mathcal{G} = (N, E, F)$, with a set of nodes N, a set of undirected and unweighted edges E, and a map of feature values $F = \{f_i\}_{i=1}^{|N|}$. That is, each node $n \in N$ is assigned a feature $f_n \in F$ and $f_1, \dots, f_{|N|}$ are 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_m be an induced subnetwork of $\mathcal{G}$ that is discovered on $\mathcal{G}$, and let F_m 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 $\hat{\theta}(F_m)$ be its empirical observed counterpart based on F_m. Then, the problem statement can be mathematically formalized as follows:

What do we know on $\Pr\{|\theta(F) - \hat{\theta}(F_m)| \geqslant \epsilon\}$*, for a given* $\epsilon > 0$ *and subnetwork size* m*? 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_m *and upper* U_m *bounds such that* $\Pr\{\theta(F) \in [L_m, U_m]\} = 1 - \alpha$*?*

Simple random sampling. A simple random sample is a subset of nodes taken from a statistical population in which each node of the subset has an equal probability of being chosen. Each node is chosen randomly and entirely by chance. A simple random sample is meant to be an unbiased representation of a population [24]. If sample size is substantially lower than population size, then simple random sampling without replacement is essentially equivalent to simple random sampling with replacement.

Seeds and waves. Let n_s be a node selected with simple random sampling from N. This node n_s is called a *seed* as it acts as a starting point for discovering a subnetwork of 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, the zeroth wave $w_0(s)$ contains only the seed itself.

Embeddedness [25]. We use node embeddedness to quantify how a node's neighborhood overlaps with those of its neighbors. The embeddedness of a node n is

$$Emb(n) = (1/|\Gamma_n|) \times \sum_{l \in \Gamma_n} |\Gamma_n \cap \Gamma_l| / |\Gamma_n \cup \Gamma_l|,$$

where Γ_n are the neighbors of the node n, and Γ_l is the neighbors of its neighbor l. If all neighbors of node n are neighbors with each other, $Emb(n) = 1$.

Confidence intervals and their validation metrics. Let $A = [L_m^A, U_m^A]$ and $B = [L_m^B, U_m^B]$ be two competing $(1-\alpha)100\%$-confidence intervals for the network feature parameter $\theta(F)$. Suppose that over a set of Monte Carlo experiments, $\Pr\{\theta(F) \in [L_m^A, U_m^A]\} = 1 - \alpha_A$ and $\Pr\{\theta(F) \in [L_m^B, U_m^B]\} = 1 - \alpha_B$. 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 $1 - \alpha$. From two alternative 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, between 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

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

which represents a relative gain or loss of using confidence interval B over A. Positive relative sharpness implies that B is sharper (shorter), whereas negative relative sharpness means the opposite.

4 Node Feature Bootstrap Algorithms

Our approach is based on two selection-re-selection algorithms. In the selection stage (Algorithm 1), we randomly sample seeds and then select neighbors around those seeds – that is, we use the idea of snowball sampling but in contrast to snowball design, we remove all edges that have been already used to locate a node and we allow for multiple inclusions of the same node. As a result, the new sampling design approach allows us to reduce the estimation bias. In the re-selection stage (Algorithm 2), we now deal with the already sampled nodes and apply resampling with replacement, or bootstrap to these nodes.

Selection-re-selection. In the sampling design, we employ snowball-like discovery of the network in parallel around multiple seeds simultaneously. We select nodes around seeds by Algorithm 1, which we call the SFINKS algorithm. The distinguishing characteristic of SFINKS compared with the Labeled Snowball with Multiple Inclusions (LSMI, [9], [10]) is that SFINKS collects feature information from the nodes. Similar to LSMI, one of the key characteristics of SFINKS is that SFINKS does not reuse edges: each edge can be used in the sampling process only once (used edges are removed), and that SFINKS accounts for node multiplicity: nodes with higher degrees can be accounted multiple times.

Algorithmic complexity. In the network $\mathcal{G} = (N, E, F)$, SFINKS chooses m seeds, and for each seed moves on the network in a breadth first fashion, while deleting already used edges, 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 [9], [10] and standard snowball designs. Indeed, for $k = |E| / |N|$ average neighbors for each seed, SFINKS is $\mathcal{O}(m \times k^w) = \mathcal{O}(k^w) = \mathcal{O}(|N|^w)$, when $m \ll |N|$ and $k_{max} = (|N| - 1)/2$. Fig. 2 shows that number of seeds must be kept low, otherwise a big fraction of the network is visited in the learning process. 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** $o(|N|^2)$.

```
input : Network $\mathcal{G} = (N, E, F)$; number of seeds $m$,
        $m \ll |N|$; number of waves $d$.
output: Approximation of selection probabilities for
        bootstrap, $\pi^{(\leqslant d),*}$, and two feature lists: of
        seeds, $\mathcal{L}_s$, and waves, $\mathcal{L}_d$.
$S : Set \leftarrow$ Select $m$ 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
            $\mathcal{L}_d \leftarrow \mathcal{L}_d \cup f_{n^*}$;
            $N' \leftarrow N' \cup n^*$;
            $E' \leftarrow E' \cup e$;
    $E \leftarrow E \setminus E'$;
    $S_w \leftarrow N'$;
    $\pi^{(\leqslant d),*} \leftarrow \pi^{(\leqslant d),*} \cup |\mathcal{L}_d| / |N|$;
    $w = w + 1$;
return $\pi^{(\leqslant d),*}$, $\mathcal{L}_s$ and $\mathcal{L}_d$;
```

Algorithm 1: SFINKS: Sampling Features In NetworKS.

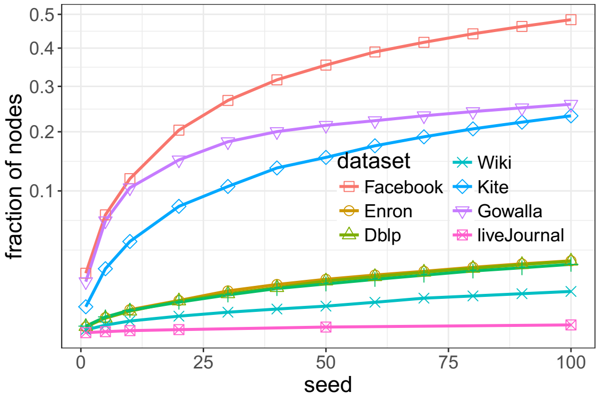


Fig. 2: Fraction of nodes visited for wave 2.

Example 1. *Fig. 3 shows a toy network ($|N| = 23$) with a structured network sample called patch (shaded area) of $m = 2$ seeds and $d = 3$ waves. Initial seeds $S = \{1, 2\}$ are selected with simple random sampling from nodes N. Node features in this case are defined as the number of drugs used. Seed node 1 uses Pain relievers and node 2 uses Pain relievers and Marijuana. Considering the number of drugs used, for these seed nodes $\mathcal{L}_s = \{1, 2\}$. 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. 3 are as*

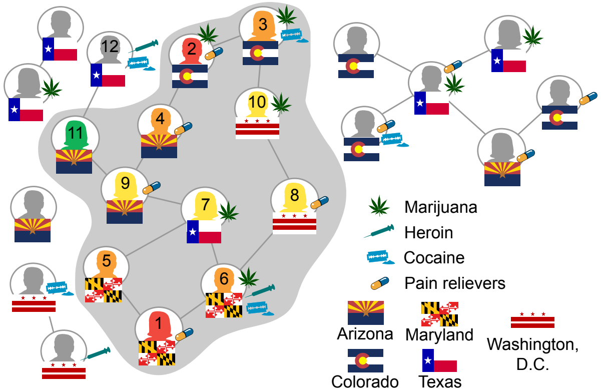


Fig. 3: A patch (shaded area) around persons 1 and 2 in a network of $|N| = 23$ people. Starting from the seeds 1 and 2 simultaneously, nodes 3, 4, 5, and 6 are discovered in the first wave. Node 12 could have only been discovered at wave 4.

follows: $S_1 = \{3, 4, 5, 6\}$, $S_2 = \{7, 7, 8, 9, 10\}$, *and* $S_3 = \{7, 8, 9, 10, 11\}$. *The resulting* $\mathcal{L}_3 = \{2, 1, 0, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0\}$. *Note that the nine values in* $\mathcal{L}_3$ *come from the nine nodes discovered in three waves, i.e.,* $S_1 \cup S_2 \cup S_3$. *The two zero values come from nodes 5 and 11.*

Once the discovery process is concluded, the GraphBoot algorithm (given in Algorithm 2) creates point estimates and bootstrap confidence intervals of the feature values mined from the discovered nodes. Since probabilities of a node to be included into a patch are different for $\mathcal{L}_s$ and $\mathcal{L}_d$, we employ separate bootstrapping schemes to reduce bias:

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

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 selected person in Fig. 3 $b = 1000$ times, we obtain a 95% BCI of (0.8, 1.4) for the average value of this feature.

Algorithmic complexity. Alg. 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 $|\mathcal{L}_s| + |\mathcal{L}_w|$, hence $\mathcal{O}(b(|\mathcal{L}_s| + |\mathcal{L}_w|))$. In practice, we use $b \ll |\mathcal{L}_s| + |\mathcal{L}_w|$, hence the complexity of the GraphBoot becomes $\mathcal{O}(|\mathcal{L}_s| + |\mathcal{L}_w|)$.

Consistency of the estimator. In this section, we prove the consistency of Alg. 2 in quantifying the estimation uncertainties.

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 $\pi_n^{(0)} = \Pr\{n \in S\} > 0$, $\forall n \in N$, $\pi_n^{(\leqslant d)} = \Pr\{n \in S^d, n \notin S\}$.

Given a set of selected nodes and their features, we propose a modified Hájek estimator for functions of F. For

```
input  : SFINKS objects: estimated sampling
         probabilities $\pi^{(\leqslant d),*}$, feature values $\mathcal{L}_s$, and
         $\mathcal{L}_d$; number of bootstrap replications $b$,
         confidence level $1 - \alpha$.
output: Bootstrap confidence interval $BCI$ for the
         feature.
$I : List \leftarrow []$;
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
        Choose one $f_n^*$ from $\mathcal{L}_s$ randomly with
         replacement;
        $\mathcal{M}_{f_n^*} = \mathcal{M}_{f_n^*} + 1$;
    initialize $\mathcal{N} : Map$ to 0 for feature values from $\mathcal{L}_d$;
    for 1 to $|\mathcal{L}_d|$ do
        Choose $f_n^*$ 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;
        $\mathcal{N}_{f_n^*} = \mathcal{N}_{f_n^*} + 1$;
    $t_{\bar{F}}^* = \dfrac{\sum_{f_n^* \in \mathcal{M}} f_n^*/\pi_n^{(0),*} + \sum_{f_n^* \in \mathcal{N}} f_n^*/\pi_n^{(\leqslant d),*}}{\sum_{f_n^* \in \mathcal{M}} 1/\pi_n^{(0),*} + \sum_{f_n^* \in \mathcal{N}} 1/\pi_n^{(\leqslant d),*}}$;
    $I \leftarrow I \cup t_{\bar{F}}^*$;
$I \leftarrow sort(I)$;
$BCI = \left(I_{[b\alpha/2]},\ I_{[b(1-\alpha/2)]}\right)$;
return $BCI$;
```

Algorithm 2: GraphBoot: Bootstrap for graph features

instance, to estimate the mean level $\bar{F}$ of features on N, we propose

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

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 key intuitive idea behind (2) is to combine estimators based on seeds and neighbors into a plausible joint estimator. That is, the numerator of (2) estimates the feature total, where the first and second terms in the numerator of (2) correspond to estimators based on seeds and neighbors, respectively. Since during the sampling stage, probability of sampling a neighbor depends on the node degree of the associated seed, we rescale estimators based on neighbours with the unbiased mean degree estimator based on seeds S, $\hat{\mu}_S$. Similarly, the denominator in (2) estimates the unknown $|N|$, and the first and second terms of the denominator in (2) correspond to estimators of $|N|$ based on seeds and neighbors, respectively.

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

$$t_{\bar{F}}^* = \frac{\sum_{n \in S} f_n^*/\pi_n^{(0),*} + \sum_{S^{(\leqslant d)}} f_n^*/\pi_n^{(\leqslant d),*}}{\sum_{n \in S} 1/\pi_n^{(0),*} + \sum_{S^{(\leqslant d)}} 1/\pi_n^{(\leqslant d),*}}. \tag{3}$$

Theorem 1 (Consistency of node feature bootstrap). *Let* $S \cup S^{(\leqslant d)}$ *be a set of nodes selected 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 $l/|N| \to 0$

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

in probability.

Proof. Let $R(N)$ be a rejective (also called maximum entropy or Poisson) selection design on N. Simple random selection without replacement is a particular case of rejective selection with equal drawing probabilities [26]. The Hájek and Horvitz–Thompson estimators obtained by rejective selection designs are known to be asymptotically normally distributed (see [27], [28]).

Now let us first consider a distribution of $t_{\bar{F}}$. Note that seed nodes in S are obtained with simple random selection without replacement, that is, via a rejective or maximum entropy selection. 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 selection 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 [29]. However, following [30], if we assume that all neighbors of a node n are included in the sample up to wave $d-1$, then selection probability $\pi^{(\leqslant d)}(n)$ for a node $n \in N$ can be approximated by a function of its degree k_n. That is,

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

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 [30] 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 [27]) $D(Q(N)||R(N)) = \sum_n Q(n) \log\left[Q(n)/R(n)\right] \to 0$. That is, the design $Q(N)$ can be approximated by a high entropy design. As a result, Theorem 4.2 of [28] 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 summands in (2) are normal, and invoking a delta-method implies that $t_{\bar{F}}$ is asymptotically normally distributed [17].

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 re-selected 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 network data to query 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 discovery 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 [31]. Formally, let $x_1, \ldots, x_n$ be features associated with n units. Then the Gini index (GI) is defined as $g = \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|/(2n \sum_{i=1}^n x_i)$. GI is widely used in economics to measure income inequality. In particular, g ranges from 0, when all individuals have equal income, to 1, when entire income is assigned to a single person. GI is also used in network studies to evaluate sparsity and centrality (for a review see [32]).

In a bootstrap context, we expect that the GI for BCI lengths will decrease as selection variability decreases. Hence, the optimal number of seeds and waves in the graph discovery framework can be determined by the minimum of GIs for BCI lengths, e.g., via the elbow plot (see Fig. 6). At the same time, if GI does not decrease, it suggests that selection 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 selection that necessarily needs to be low for reliable estimates of node features.

5 Experiments

We use GraphBoot to quantify uncertainty of node feature learning in simulated and real life networks where the ground truth is known (Sections 5.1 and 5.2) and unknown (Section 5.3). We then show the efficiency of our methods for varying sizes of the discovered subgraphs in Section 5.4.[1]

5.1 Simulated Networks

We use 10^4 simulated networks of order $|N| = 10^4$ which follow a power-law GutenbergRichter degree distribution with parameters 0.01 and 2 [33]. 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 competing baseline approaches, we use simple random selection of 50 seeds to construct normal and nonparametric bootstrap confidence intervals [16], [24]. Our choice

[1] The Scala/Spark implementation of the algorithms is available on https://github.com/cakcora/GraphBoot.

TABLE 1: Observed coverage (%) of 95% confidence intervals (width is in parentheses) on simulated data. GraphBoot is based on a sample of 20 seeds with 2 waves; 90, 200 and 350 nodes are minimum, average and maximum numbers of total nodes (i.e., unique seeds+neighbors) used by GraphBoot across considered synthetic networks.

		Features f_i ($i = 1, \ldots, \|N\|$)		
Interval	n	$k_i^2 + Pois(1)$	$k_i^2 + Pois(4)$	$k_i + N(0,1)$
GraphBoot	—	98.7 (4.88)	98.7 (4.90)	94.4 (0.58)
SRS normal	20	81.2 (13.63)	81.6 (13.74)	92.2 (1.81)
	50	87.4 (9.26)	87.6 (9.32)	93.7 (1.17)
	90	89.9 (7.17)	90.2 (7.21)	94.0 (0.87)
	200	92.2 (4.89)	92.4 (4.91)	94.4 (0.59)
	350	93.4 (3.71)	93.4 (3.72)	94.6 (0.44)
SRS bootstrap	20	81.7 (12.77)	82.1 (12.89)	91.2 (1.75)
	50	88.0 (8.97)	88.1 (9.03)	93.4 (1.15)
	90	90.2 (7.03)	90.4 (7.07)	93.8 (0.86)
	200	92.4 (4.86)	92.7 (4.88)	94.2 (0.58)
	350	93.7 (3.72)	93.7 (3.74)	94.6 (0.44)

in selecting the competing alternatives is dictated by the following rationale:

- First, to the best of our knowledge, there exist no other techniques for Uncertainty Quantification in network node feature mining (again here we would like to underline that Uncertainty Quantification is not to be confused with network sampling).
- Second, to illustrate the utility of the proposed network bootstrap, we allow the competing baseline methods to use up to 350 seeds vs. 20 seeds used by bootstrap. Note that in many real life scenarios, e.g., analysis of terrorist activity and HIV risk propagation, it is substantially more challenging to get information from new seeds than from the neighbors of the already identified seeds.

Table 1 shows the performance of GraphBoot in quantifying the uncertainty in feature averages. Following the anonymous reviewers' suggestion, Table 1 includes two extreme cases (a) same number of seeds in the baseline method of simple random sampling (SRS) as in the proposed bootstrap approach, and (b) minimum, average and maximum numbers of total sampled nodes (i.e., unique seeds + neighbours) in the baseline SRS method as engaged in the proposed bootstrap approach. As expected, we find that under the case (a), the baseline SRS method delivers liberal confidence intervals (CIs) which largely undercover, i.e., empirical coverage probability is 82–92% for the expected coverage level of 95%. Remarkably, even for 200 seeds (i.e., the case (b)), SRS tends to yield lower coverage than the expected 95% level, while for the same number of total nodes, the proposed bootstrap method delivers either calibrated or moderately conservative CIs. Only the case of SRS for 350 seeds yields relatively well calibrated CIs with coverage of 94–95%. Note that conservative CIs are preferred over liberal CIs.

5.2 Real Life Networks - Quantifying Uncertainty with Ground Truth

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 [34]. The Facebook [35] dataset contains 817K edges among 63K New Orleans network users. The DBLP [36] network is an undirected co-authorship network. Gowalla [37] and BrightKite [37] are undirected location based social networks.

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

Network	$\|N\|$	$\|E\|$	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

In directed network, the LiveJournal edges are friendships, whereas in the Wikipedia network [38], an edge is created among nodes who have edited each other's talk page. In the Enron [39] network, we create an edge between two nodes who have shared emails. The Epinions [40] network is the only signed network in our dataset, where an edge between two users indicates their trust for each other. In directed networks, we created edges among two nodes when the nodes have outgoing edges to each other, i.e., $n_1, n_2 \in \mathcal{G}.N$ and directed edges $\langle n_1 \rightarrow n_2 \rangle$, $\langle n_2 \rightarrow n_1 \rangle$ exist in the dataset. In this case an edge is recorded between n_1 and n_2, i.e., $\langle n_1, n_2 \rangle \in E$.

Estimation. Fig. 4 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.

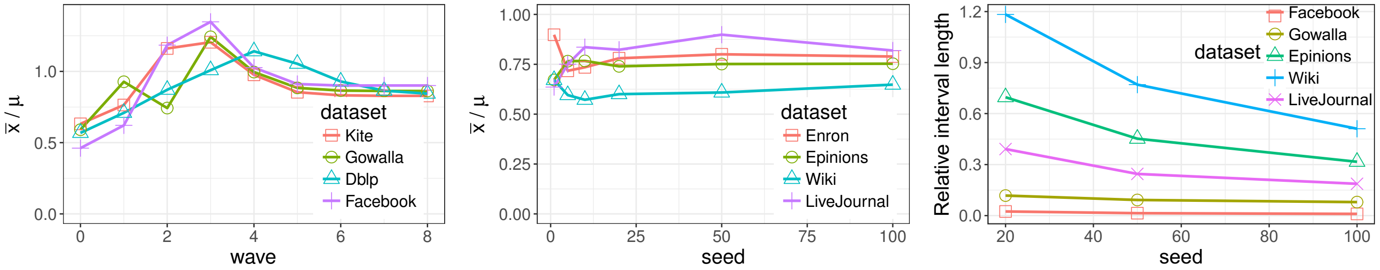
Fig. 4 shows that undirected networks can be efficiently queried by GraphBoot with 2 or 3 waves for average degrees. The proportion of used nodes depends on the distribution of degrees and embeddedness in a network. Due to this low proportion, directed networks in Fig. 4b have lower estimates. Thus, sampling in these networks must start with more seeds or continue for more waves.

As we noted in Section 4, in addition to the estimated value, GraphBoot also reports a confidence interval BCI for the estimated statistic. This confidence interval shows dispersion of estimated values in 1,000 bootstraps. Fig. 4c shows that GraphBoot confidence intervals are much shorter for undirected networks than for the directed ones.

How would baseline UQ methods perform? Fig. 5 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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2925355, IEEE Transactions on Knowledge and Data Engineering



(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 seeds.

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

Fig. 4: GraphBoot relative degree estimates on networks, and confidence interval lengths for the estimates.

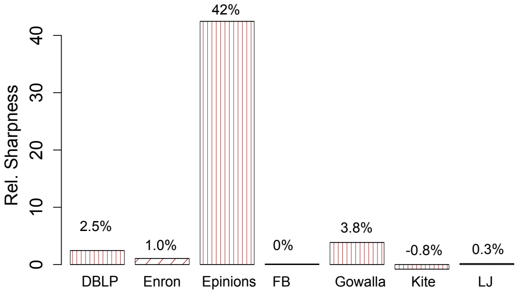


Fig. 5: Relative Sharpness (RS) comparison of GraphBoot bootstrap and normal-based confidence intervals. Positive values of RS implies a gain of GraphBoot bootstrap CI vs. normal-based CI (in %), and reverse. In all cases, except of Enron, all CIs contain the true population parameter.

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. 6 shows the GI of BCI lengths for varying waves and seeds. As expected, increasing number of seeds leads to lower selection variability and, as a result, to lower Gini Index values. Fig. 6a and 6b can be used as an elbow plot for selecting optimal number of seeds in a selection design for a particular node feature, so that adding extra seeds does not reduce dispersion. For instance, in Fig. 6b, all networks reach low GI around 25 seeds, and hence we can conclude that 25 seeds is a reasonable selection size. Undirected networks tend to deliver low GI even with few seeds, which can be attributed to degree assortativity or bigger diameters.

Fig. 6c shows a counterpart study of GI against varying numbers of waves. While all GIs 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 selection. 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. 6a and 6c, with a very slow decay of selection variability.

Furthermore, Fig. 6c 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.

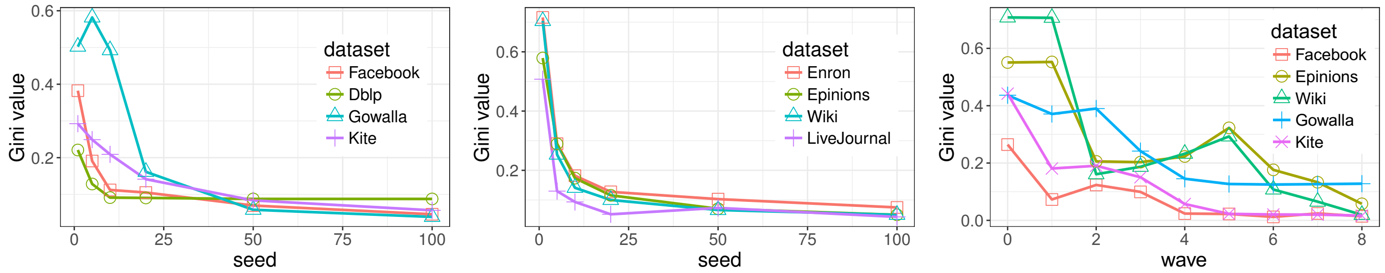
Gini Index values of GraphBoot degree estimations on real life networks show that depending on the network coverage and network type, GraphBoot can provide degree estimates with as few as 25 seeds and 2 waves. Furthermore, by observing how the inequality of estimated confidence interval lengths change,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

5.3 Twitter Sub-networks - 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 [20], and ii) there are well known mental health related features [21] to query with GraphBoot.

Querying a mental health subnetwork. Our node querying 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



(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.

Fig. 6: GIs of confidence interval lengths from GraphBoot degree estimations on real life networks.

health issues themselves. These initial accounts are found by manually searching mental health related keywords, such as depression and mental health, on Twitter. *Note that for this subnetwork querying, 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 the Twitter API 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.

To train the model also with MentalHealth- vectors, we downloaded the Twitter data set of Cheng et al. [41] and created feature vectors of users in a similar way. The user profiles in the Cheng dataset were assumed to be not related to mental health, i.e., these users' vectors were labeled as MentalHealth-. As 36% of Twitter users scored positive for depression in [21], we decided to combine 13 MentalHealth+ vectors with 26 randomly selected MentalHealth- vectors.

Out of 1,895 wave 1 nodes that were found by using the 39 user profiles as seeds, 1,784 of them had public tweets. Remaining nodes either have not tweeted yet, or chose to hide their tweets by making their profiles private. 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 [42]) of 0.46, with values: 8 false positives, 33 true positives, 22 false negatives and 91 true negatives. Phase 0 accuracy of the random forest model is thus found to be 80.4%.

Phase 1. In phase 1, we added the manually labeled vectors to the model input, and trained a new model on the expanded dataset. The new random forest model was then used to classify neighbors of 258 wave 1 nodes who were labeled as MentalHealth+ in phase 0 by the random forest classifier. 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.

TABLE 3: Classification numbers in phases. In phase=1, 194 out of 197 accounts have been manually assigned a label. Three accounts were left as ”no decision”.

Ph.	Verified 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

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. Furthermore, Fig. 1 shows density of mental health related accounts on a map.

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.

Mining 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 [21]. 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 [21], Day and Night, are used to estimate daily posting habits of Twitter users (See the code repository for this dataset). 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 Gini Index values for these groups are given in Fig. 7a. Day and Night estimates of GraphBoot are 15.8 and 3.92, respectively. GI of these time related statistics are given in Fig. 7b. Gini Index changes of both word groups and time aspects show that GraphBoot wave 1 computations are better than wave 0 and wave 2 estimates. Overall, these statistics show low Gini Index values in GraphBoot estimates, and hence can be viewed as estimable statistics for mental health studies.

Control Experiments. As a control experiment, we devised two statistics, username length and registration 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.

In this experiment we sample the username and year features from the subnetwork and hypothesize that adding more seeds or employing bigger wave values will not reduce the uncertainty in the estimates of these features. This is due to our assumption that the two features do not depend on the mental health related nature of Twitter accounts. If our hypothesis wrong, we expect to see reduced uncertainty when more data is sampled from the network.

GraphBoot estimates for these statistics are 2013.6 for year and 11.9 for username. GraphBoot results in Fig. 7c show that the Gini Index values for these statistics are high, and even increase at waves 1 and 2.

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

5.4 Epinions Network - Discovery with Size Effects

So far, we have looked into the uncertainty of learning features from the same set of nodes on a network. In this section, we will use the Epinions Ratings dataset [34] and show uncertainty for estimating multiple features from the same network where node sets that have these features have different sizes. This experimental setting allows us to see how features of varying node sets from the same network can be compared in terms of their estimation uncertainties.

The Epinions dataset consists of user given ratings (i.e., -1 or +1 for trusted and distrusted content, respectively) to articles written by other users. Each article has a topic, and a user might have multiple articles on the same or different topics. This allows us to create a network for specific topic; we define each feature value f_{tn} as the number of articles with topic t written by user n. Nodes are users who have written at least one article on topic f_t. The edges are trust/distrust ratings of users; two users are connected if any of them rate an article written by the other user.

We chose six topics from the Epinions network; three most popular topics (i.e., H1, H2 and H3, with each having at least 3000 users) and three moderately common topics (i.e., L1, L2 and L3 with each having 1000–3000 users). We refer to these topics as H and L topics, respectively.

In Figure 8a, we show relative values of estimations for H and L topics on the Epinions network. Both H and L topics reach stable estimates with 10 seeds. However L topics are shown to have higher Gini values in Figure 8b. Despite this, these values are very low (i.e., 0.04) compared to Gini values from Figure 6. Similarly, as shown in Figure 8c, the confidence intervals are much narrower in topic networks.

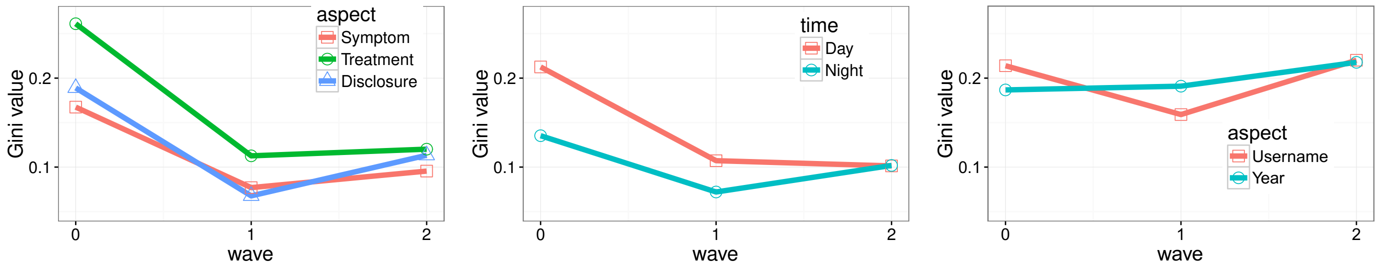
This is due to shared topic interests on the network which results in a better discovery process; Epinions users rate each other for their content and this results in a network where nodes that are close to each other on the network write about and rate similar topics [43]. As a result, the discovery process from initial seeds results in lower Gini index because waves are more likely to discover nodes that are similar to the initial sees. Resulting low Gini index reflects this fact and indicates an efficient network discovery process.

5.5 Feature estimation and Homophily

An aspect in feature estimation is that homophily may lead to biased estimates. To diminish the estimation bias due to homophily, we can increase the number of sampled seeds and evaluate stability of the resulting confidence intervals along with the sufficient number of seeds to deliver stable performance, using, for example, an elbow plot. Furthermore, we argue that getting trapped in homophilous groups is very difficult when we start from randomly selected seeds and seed number is above 20. Considering high network orders (billions of nodes on Facebook), probability of randomly picking multiple seeds from the same homophilous community is low. In Figure 4a and 4b we show that estimates are stable for more than 20 seeds. Although these experiments are degree based, we believe that similar results will hold for other features. Although we have graph datasets from real life networks, privacy policies of homophilous networks such as Facebook do not allow feature mining on nodes. As such, currently, we are not able to complete an experiment to test this hypothesis, and leave this issue as a future work. Finally, in many cases what we are interested in are features in homophilous groups. For example, when monitoring mental health among HIV patients, it is useful to follow and locate homophilous relations, so that more members of this small subnetwork can be discovered, and subsequently their features are mined. In our Twitter experiments, we address this aspect and our experiments are targeted for feature mining from hard-to-reach subnetworks.

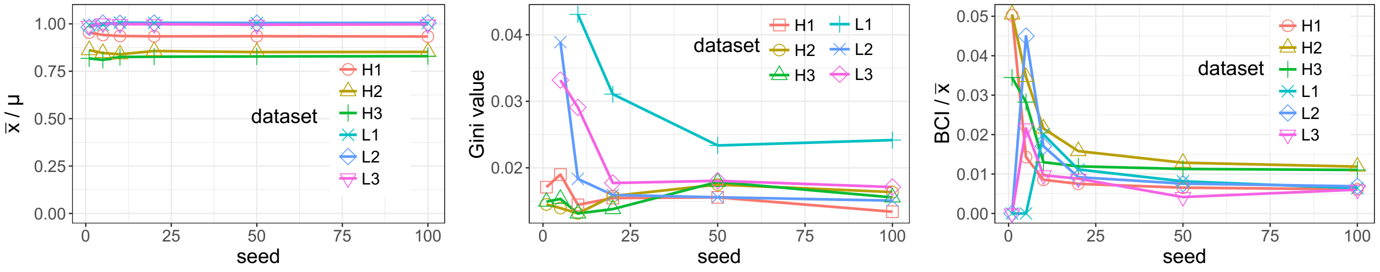
6 Conclusion

In this paper, we quantify estimation uncertainties of node features in a variety of large social networks. Our experiments include a case study of learning mental health related features on the Twitter network efficiently. We show that the match between the estimated feature and the network type



(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.

Fig. 7: GIs for Twitter mental health experiments.



(a) GraphBoot wave 2 relative results for varying seed counts.
(b) Gini values for estimation uncertainties for varying seed counts.
(c) Relative confidence interval length, $|BCI|/\bar{x}$, for GraphBoot wave 2 results.

Fig. 8: GraphBoot performance on Epinions topic networks. H topics are the three most common topics in the dataset.

and size greatly affects the estimation quality. Undirected networks where edges are created with mutual consent provided better feature estimates, whereas directed networks have inequalities within samples. Topical networks, such as Epinions, provide lower uncertainty in feature learning because edges among nodes are a good proxy of shared interests. In all cases, GraphBoot provides adequate quality indications for the learning process.

We plan to expand GraphBoot to quantifying uncertainty in motif estimation and feature-based anomaly detection.

Acknowledgment

Gel has been partially supported by NSF IIS 1633331, NSF ECCS 1824716 and NSF DMS 1736368, Kantarcioglu has been partially supported by NSF IIS 1633331, NSF CICI 1547324, NSF CNS 1837627, NSF OAC 1828467, NIH award 1R01HG006844, and ARO award W911NF-17-1-0356, and Lyubchich has been partially supported by NSF IIS 1633355.

References

[1] N. K. Ahmed, N. Duffield, J. Neville, and R. Kompella, "Graph sample and hold: A framework for big-graph analytics," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014, pp. 1446–1455. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623757

[2] O. Simpson, C. Seshadhri, and A. McGregor, "Catching the head, tail, and everything in between: A streaming algorithm for the degree distribution," in *Proceedings of the 2015 IEEE International Conference on Data Mining (ICDM)*, ser. ICDM '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 979–984. [Online]. Available: http://dx.doi.org/10.1109/ICDM.2015.47

[3] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer, "Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks," *Ann. Appl. Stat.*, vol. 9, no. 1, pp. 166–199, 03 2015. [Online]. Available: https://doi.org/10.1214/14-AOAS800

[4] T. A. B. Snijders and S. P. Borgatti, "Non-parametric standard errors and tests for network statistics," *Connections*, vol. 22, no. 2, pp. 61–70, 1999.

[5] C. T. Butts, "Network inference, error, and informant (in)accuracy: a bayesian approach," *Social Networks*, vol. 25, no. 2, pp. 103 – 140, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378873302000382

[6] D. R. Farine and A. Strandburg-Peshkin, "Estimating uncertainty and reliability of social network data using Bayesian inference," *Royal Society Open Science*, vol. 2, no. 9, p. 150367, 2015.

[7] D. Lusseau, H. Whitehead, and S. Gero, "Incorporating uncertainty into the study of animal social networks," *Animal Behavior*, no. 75, pp. 1809–1815, 2008.

[8] S. Epskamp, D. Borsboom, and E. I. Fried, "Estimating psychological networks and their accuracy: A tutorial paper," *Behavior Research Methods*, vol. 50, no. 1, pp. 195–212, 2018. [Online]. Available: https://doi.org/10.3758/s13428-017-0862-1

[9] M. E. Thompson, L. L. Ramirez Ramirez, V. Lyubchich, and Y. R. Gel, "Using the bootstrap for statistical inference on random graphs," *Canadian Journal of Statistics*, vol. 44, no. 1, pp. 3–24, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cjs.11271

[10] Y. R. Gel, V. Lyubchich, and L. L. Ramirez Ramirez, "Bootstrap quantification of estimation uncertainties in network degree distributions," *Scientific Reports*, vol. 7, no. 1, Dec 2017. [Online]. Available: http://par.nsf.gov/biblio/10039995

[11] J. D. Hwang and K. Hollingshead, "Crazy mad nutters: The language of mental health," in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 52–62. [Online]. Available: https://www.aclweb.org/anthology/W16-0306

[12] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar, "Discovering shifts to suicidal ideation from mental health content in social media," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ser. CHI

'16. New York, NY, USA: ACM, 2016, pp. 2098–2110. [Online]. Available: http://doi.acm.org/10.1145/2858036.2858207

[13] L. A. Goodman, "Snowball sampling," *The Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 148–170, 1961. [Online]. Available: http://www.jstor.org/stable/2237615

[14] M. Granovetter, "Network sampling: Some first steps," *American Journal of Sociology*, vol. 81, no. 6, pp. 1287–1303, 1976. [Online]. Available: http://www.jstor.org/stable/2777005

[15] J. P. Scott and P. J. Carrington, *The SAGE Handbook of Social Network Analysis*. Sage Publications Ltd., 2011.

[16] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, ser. Monographs on Statistics and Applied Probability. Boca Raton, Florida, USA: Chapman & Hall/CRC, 1993, no. 57.

[17] A. W. v. d. Vaart, *Asymptotic Statistics*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[18] D. Pfefferman and C. R. Rao, *Sample Surveys: Inference and Analysis*. Amsterdam: Elsevier/North-Holland, 2009.

[19] S. Bhattacharyya and P. J. Bickel, "Subsampling bootstrap of count features of networks," *Ann. Statist.*, vol. 43, no. 6, pp. 2384–2411, 12 2015. [Online]. Available: https://doi.org/10.1214/15-AOS1338

[20] M. Kumar, M. Dredze, G. Coppersmith, and M. De Choudhury, "Detecting changes in suicide content manifested in social media following celebrity suicides," in *Proceedings of the 26th ACM Conference on Hypertext &#38; Social Media*, ser. HT '15. New York, NY, USA: ACM, 2015, pp. 85–94. [Online]. Available: http://doi.acm.org/10.1145/2700171.2791026

[21] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Seventh international AAAI conference on weblogs and social media*. AAAI, July 2013. [Online]. Available: https://www.microsoft.com/en-us/research/publication/predicting-depression-via-social-media/

[22] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 51–60. [Online]. Available: https://www.aclweb.org/anthology/W14-3207

[23] M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference*, ser. WebSci '13. New York, NY, USA: ACM, 2013, pp. 47–56. [Online]. Available: http://doi.acm.org/10.1145/2464464.2464480

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

[25] M. Granovetter, "Economic action and social structure: The problem of embeddedness," *American Journal of Sociology*, vol. 91, no. 3, pp. 481–510, 1985. [Online]. Available: https://doi.org/10.1086/228311

[26] D. Pfeffermann and C. R. Rao, *Handbook of Statistics_29A: Sample Surveys: Design, Methods and Applications*. Amsterdam: Elsevier, 2009, vol. 29.

[27] Y. G. Berger, "Rate of convergence for asymptotic variance of the horvitzthompson estimator," *Journal of Statistical Planning and Inference*, vol. 74, no. 1, pp. 149 – 168, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378375898001074

[28] H. Boistard, H. P. Lopuhaä, and A. Ruiz-Gazen, "Approximation of rejective sampling inclusion probabilities and application to high order correlations," *Electron. J. Statist.*, vol. 6, pp. 1967–1983, 2012.

[29] T. A. B. Snijders, "Estimation on the basis of snowball samples: How to weight?" *Bulletin of Sociological Methodology/Bulletin de Mthodologie Sociologique*, vol. 36, no. 1, pp. 59–70, 1992. [Online]. Available: https://doi.org/10.1177/075910639203600104

[30] J. Illenberger and G. Flötteröd, "Estimating network properties from snowball sampled data," *Social Networks*, vol. 34, no. 4, pp. 701–711, 2012.

[31] J. L. Gastwirth, "The estimation of the lorenz curve and gini index," *The Review of Economics and Statistics*, vol. 54, no. 3, pp. 306–16, 1972. [Online]. Available: https://EconPapers.repec.org/RePEc:tpr:restat:v:54:y:1972:i:3:p:306-16

[32] X. Qin, P. Cunningham, and M. Salter-Townshend, "The influence of network structures of Wikipedia discussion pages on the efficiency of WikiProjects," *Social Networks*, vol. 43, pp. 1–15, 2015.

[33] V. Latora, V. Nicosia, and G. Russo, *Complex networks: principles, methods and applications*. Cambridge University Press, 2017.

[34] J. Kunegis, "Konect: The koblenz network collection," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13 Companion. New York, NY, USA: ACM, 2013, pp. 1343–1350. [Online]. Available: http://doi.acm.org/10.1145/2487788.2488173

[35] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, ser. WOSN '09. New York, NY, USA: ACM, 2009, pp. 37–42. [Online]. Available: http://doi.acm.org/10.1145/1592665.1592675

[36] M. Ley, "Dblp - some lessons learned." *PVLDB*, vol. 2, no. 2, pp. 1493–1500, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/pvldb/pvldb2.html#Ley09

[37] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. New York, NY, USA: ACM, 2011, pp. 1082–1090. [Online]. Available: http://doi.acm.org/10.1145/2020408.2020579

[38] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1361–1370. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753532

[39] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Proceedings of the 15th European Conference on Machine Learning*, ser. ECML'04. Berlin, Heidelberg: Springer-Verlag, 2004, pp. 217–226. [Online]. Available: https://doi.org/10.1007/978-3-540-30115-8_22

[40] M. Richardson, R. Agrawal, and P. Domingos, "Trust management for the semantic web," in *Proceedings of the Second International Conference on Semantic Web Conference*, ser. LNCS-ISWC'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 351–368. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-39718-2_23

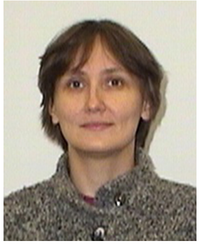
[41] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: A content-based approach to geo-locating twitter users," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10. New York, NY, USA: ACM, 2010, pp. 759–768. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871535

[42] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psych. Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[43] C. G. Akcora, B. Carminati, and E. Ferrari, "User similarities on social networks," *Social Network Analysis and Mining*, vol. 3, no. 3, pp. 475–495, 2013.

Cuneyt Gurcan Akcora is a Postdoctoral Fellow in the Departments of Statistics and Computer Science at the University of Texas at Dallas. He received his Ph.D. from University of Insubria, Italy and his M.S. from SUNY Buffalo, USA. His research interests are Data Science on complex networks and large scale graph analysis, with applications in social, biological, IoT and Blockchain networks. He is a Fulbright Scholarship recipient, and his research works have been published in leading conferences and journals including VLDB, ICDM and ICDE.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2019.2925355, IEEE Transactions on Knowledge and Data Engineering



Yulia R. Gel is Professor in the Department of Mathematical Science at the University of Texas at Dallas. Her research interests include statistical foundation of Data Science, inference for random graphs and complex networks, time series analysis and predictive analytics. She holds a Ph.D in Mathematics, followed by a postdoctoral position in Statistics at the University of Washington. Prior to joining UT Dallas, she was a tenured faculty member at the University of Waterloo, Canada. She held visiting positions at Johns Hopkins University, University of California, Berkeley, and the Isaac Newton Institute for Mathematical Sciences, Cambridge University, UK. She served as a Vice President of the International Society on Business and Industrial Statistics, and is a Fellow of the American Statistical Association.

Murat Kantarcioglu is a Professor in the Computer Science Department and Director of the UTD Data Security and Privacy Lab at the University of Texas at Dallas and a visiting scholar at Harvard University Data Privacy Lab. He is a recipient of NSF CAREER award, and Purdue CERIAS Diamond Award for Academic excellence. His research focuses on creating technologies that can efficiently extract useful information from any data without sacrificing privacy or security. His research has been supported by grants from NSF, AFOSR, ONR, NSA, and NIH. He has published over 160 peer reviewed papers related to data security, privacy and privacy-preserving data mining. His research work has been covered by the media outlets, such as Boston Globe, ABC News, and has received three best paper awards.

Vyacheslav Lyubchich received a PhD degree in Statistics from the Orenburg State University, Russia in 2011. In the same year, he was awarded the Government of Canada postdoctoral fellowship to continue his research in time series methodology at the Department of Statistics and Actuarial Science of the University of Waterloo, Canada. Since 2015, V. Lyubchich is a research assistant professor and a founding member of the Environmental Statistical Collaborative at the University of Maryland Center for Environmental Science, USA.

Bhavani Thuraisingham is the Louis A. Beecherl, Jr. Distinguished Professor of Computer Science and the Executive Director of the Cyber Security Research and Education Institute at the University of Texas at Dallas. She is also a visiting Senior Research Fellow at Kings College, University of London and a Fellow of the ACM, IEEE, and the AAAS. She has received several awards including the IEEE CS 1997 Technical Achievement Award, ACM SIGSAC 2010 Outstanding Contributions Award, and the ACM SACMAT 10 Year Test of Time Award. She has worked in industry (Honeywell), federal laboratory (MITRE), US government (NSF) and her work has resulted in 120 journal articles, 250 conference papers, 130 keynote and featured addresses, six US patents, fifteen books as well as technology transfer. She received her PhD from the University of Wales, Swansea, UK, and earned higher doctorate (D. Eng) from the University of Bristol, UK.