Should We Be Confident in Peer Effects Estimated From Partial Crawls of Social Networks?

Jiasen Yang[*] and **Bruno Ribeiro**[†] and **Jennifer Neville**[*†]
Departments of Statistics[*] and Computer Science[†]
Purdue University, West Lafayette, IN
{jiaseny,ribeirob,neville}@purdue.edu

Abstract

Research in social network analysis and statistical relational learning has produced a number of methods for learning relational models from large-scale network data. Unfortunately, these methods have been developed under the unrealistic assumption of full data access. In practice, however, the data are often collected by crawling the network, due to proprietary access, limited resources, and privacy concerns. While prior studies have examined the impact of network crawling on the structural characteristics of the resulting samples, this work presents the first empirical study designed to assess the impact of widely used network crawlers on the estimation of peer effects. Our experiments demonstrate that the estimates obtained from network samples collected by existing crawlers can be quite inaccurate, unless a significant portion of the network is crawled. Meanwhile, motivated by recent advances in partial network crawling, we develop crawl-aware relational methods that provide accurate estimates of peer effects with statistical guarantees from partial crawls.

Introduction

The recent explosion of large-scale network datasets has fueled a great deal of interest in learning *relational* models to identify *peer effects*. For example, political views are often correlated among friends in social networks. While much work has been done in the relational learning community to develop models and algorithms for estimation and inference in networks, a primary assumption underlying these works is that a *full* network is available for learning. However, the network datasets used to study peer effects are typically *samples* of a larger network. In particular, it is often the case that researchers do not have *random access* to the full network structure and that sampling is only possible via repeated crawling from a node to one of its neighbors.

In this work, we provide, to the best of our knowledge, the first empirical study to assess the accuracy and reliability of peer effect estimates from crawled network data, comparing five different sampling methods across five network datasets. We learn relational models from the crawled samples and assess the quality of their parameter estimates, model predictions, and confidence intervals for the estimated parameters. We show that data collected by existing network crawlers, when used for estimation in relational methods, often produce inaccurate estimates of peer effects unless a large portion of the network is crawled. Most importantly, we find that two researchers, using different partial crawls, could reach widely divergent estimates, which affects the trustworthiness and reproducibility of the results. For instance, in a large social network dataset, we show that most methods would incorrectly assess the correlation between a user's age and their friend's zodiac sign (*cf.* Figure 2a).

Meanwhile, we also demonstrate that accurate estimation of peer effects is possible through the design of *crawl-aware* relational methods. Recently, (Avrachenkov, Ribeiro, and Sreedharan 2016) proposed a general crawling method that yield estimates with statistical guarantees of edge-based functions in graphs. Based on their sampling method, we derive an improved crawl-aware parameter estimation algorithm, and provide a nonparametric approach to computing confidence intervals for the estimated peer effects.

Summary of Contributions

- We conduct a set of experiments to investigate the impact of network sampling upon the estimation of relational learning models under an access-restricted scenario.
- We show that estimates based on popular network crawling methods are often inaccurate and may lead to incorrect conclusions regarding peer effects.
- We introduce crawl-aware relational methods that can accurately estimate network-wide peer effects and construct well-calibrated confidence intervals from crawling only a small portion of a large attributed social network.

Problem Definition

The goal of this work is to study the effects of sampling on the estimation of relational models in large social networks under an access-restricted scenario. More specifically, we are interested in accurately estimating model parameters in order to effectively assess *peer effects*—*i.e.*, the importance of relational features involving the neighbors of a node.[1]

We assume that random access to the full network structure is not available, and that the network can only be accessed via crawling. Specifically, we assume (*i*) the avail-

[1] In this work, the peer effects are represented by the parameters in a relational model, and we shall use the terms interchangeably.

ability of a seed node in the network, (*ii*) the ability to query for the attributes of a sampled node, and (*iii*) the ability to transition to neighbors of a sampled node.

Given such an access pattern, and assuming that the full network cannot be crawled, the task is to accurately estimate peer effects by learning a relational model over the sampled network. If we refer to the estimates that a learning algorithm would obtain from the full network as *global* estimates and those from the sampled network as *sample* estimates, then the ideal method should produce (*i*) unbiased sample estimates (*w.r.t.* the global estimates), and (*ii*) accurate assessments of the uncertainty associated with the sample estimates (*e.g.*, confidence intervals).

Background and Related Work

Denote a graph by $G = (V, E)$, where V is the set of vertices and $E \subseteq V \times V$ is the set of edges. For a node $v \in V$, denote its neighbors by $\mathcal{N}_v = \{u \in V : (u, v) \in E\}$ and its degree by $d_v = |\mathcal{N}_v|$. Finally, let $\mathbb{1}\{\cdot\}$ denote the indicator function.

Network Sampling Algorithms We note that under a crawl-based scenario, any technique involving random node/edge selection will be infeasible.

Snowball sampling (BFS) traverses the network via a breadth-first search. *Forest fire (FF)* (Leskovec and Faloutsos 2006) samples ("burns") a random fraction of a node's neighbors, and repeats this process recursively for each "burned" neighbor. *Random walk sampling (RW)* performs a random walk on the network by transitioning from the current node to a randomly selected neighbor at every step. *Metropolis-Hastings random walk (MH)* (Gjoka et al. 2010) sets the transition probability from node u to v as $\mathbf{P}^{\mathsf{MH}}_{u,v} = \min(1/d_u, 1/d_v)$ if $v \in \mathcal{N}_u$ and $1 - \sum_{w \neq u} \mathbf{P}^{\mathsf{MH}}_{u,w}$ if $v = u$, which yields a uniform stationary distribution over nodes.

Random walk tour sampling (TS) (Avrachenkov, Ribeiro, and Sreedharan 2016) is a recently proposed method that exploits the regenerative properties of random walks. Given an initial seed node, the algorithm first performs a short random walk to collect a set of seed nodes $\mathcal{S} \subseteq V$, and then proceeds to sample a sequence of random walk *tours*. Specifically, the k-th random walk tour starts from a sampled node $v_1^{(k)} \in \mathcal{S}$ and transitions through a sequence of nodes $v_2^{(k)}, \cdots, v_{\xi_k - 1}^{(k)} \in V \backslash \mathcal{S}$ until it returns to a node $v_{\xi_k}^{(k)} \in \mathcal{S}$. The algorithm repeats this process to sample m such tours, denoted $\mathcal{D}_m(\mathcal{S}) = \{(v_1^{(k)}, \ldots, v_{\xi_k}^{(k)})\}_{k=1}^m$. Since the successive returns to a seed node in $\mathcal{S}$ act as renewal epochs, the renewal reward theorem (Brémaud 1999) guarantees that sample statistics computed from each tour will be independent.

Relational Learning Models We utilize statistical relational learning models to estimate peer effects in networks.

Relational Bayes classifier (RBC) (Neville, Jensen, and Gallagher 2003; Macskassy and Provost 2007) is a widely used and interpretable relational model. RBC is similar to the conventional naive Bayes classifier except that the target class of a node is conditioned on the attributes and class label of its neighbors. The posterior probability of a node $v \in V$ with attributes $\mathbf{x}_v \in \mathbb{R}^k$ taking on label y_v is given by

$$\Pr(y_v | \mathbf{x}_v, \mathcal{N}_v) \propto \Pr(y_v) \prod_{i=1}^{k} \Pr(x_{v,i} | y_v) \prod_{u \in \mathcal{N}_v} \prod_{j=1}^{k+1} \Pr(\phi_{u,j} | y_v), \quad (1)$$

where $\boldsymbol{\phi}_u \triangleq (y_u, \mathbf{x}_u) \in \mathbb{R}^{k+1}$ contains both the class label and attribute values of node u. Parameters in the RBC model include the prior probability $\Pr(y_v)$ of node v having label y_v, conditional probabilities $\Pr(x_{v,i}|y_v)$ of node v having attribute value $x_{v,i}$ given label y_v, and conditional probabilities $\Pr(\phi_{u,j}|y_v)$ that a neighboring node u has attribute/label $\phi_{u,j}$ given that node v has label y_v. The conditional probabilities correspond to the peer effects we are interested in.

Comparison of Network Sampling Methods While prior studies have examined the impact of crawling on network analysis, our work differs from theirs in significant ways.

Leskovec and Faloutsos (2006) studied the impact of network sampling methods on the *structural* characteristics of the resulting sample, but they did not consider the impact of sampling on the estimation of relational models in *attributed* networks. On the other hand, Ahmed, Neville, and Kompella (2012) studied the impact of network sampling on the performance of the *weighted-vote relational neighbor (wvRN)* classifier of (Macskassy and Provost 2007). However, the wvRN does not contain any parameters—it simply predicts a node's label via a majority vote among its neighbors—and is therefore incapable of estimating peer effects.

Proposed Methodology

Given an unobserved network $G = (V, E)$, the tasks are (*i*) to estimate the parameters $\boldsymbol{\theta}$ in a relational model by crawling the network G from an initial set of seed nodes $\mathcal{S} \subseteq V$, and (*ii*) to assess the uncertainty associated with the estimates $\widehat{\boldsymbol{\theta}}$. Thus, the full procedure for estimating peer effects from a large social network should consist of three phases:

Crawling Crawl the network using a sampling method.

Estimation Estimate peer effects from the crawled network.

Calibration Compute confidence intervals for the estimates.

For the crawling phase, we shall employ the random walk tour sampling algorithm. Next, we discuss the details of our proposed estimation and calibration methodology.

Relational Model Estimation In general, we do not have any guarantees on the quality of peer effects estimated from a crawled network. However, if the sample were collected using the tour sampling algorithm, we propose applying Theorem 1 to accurately estimate the parameters in an RBC.

Theorem 1 *Given the sampled tours* $\{(v_1^{(k)}, \ldots, v_{\xi_k}^{(k)})\}_{k=1}^m$, *the following estimates for the prior, joint, and conditional probabilities in Eq.* (1) *are asymptotically unbiased:*

$$\widehat{\Pr}(c) \propto \frac{d_{\mathcal{S}}}{m} \sum_{k=1}^{m} \sum_{t=2}^{\xi_k - 1} \frac{1}{d_{v_t}} \mathbb{1}\{y_{v_t^{(k)}} = c\} + \sum_{v \in \mathcal{S}} \mathbb{1}\{y_v = c\} \quad (2)$$

$$\widehat{\Pr}(a, b) \propto \frac{d_{\mathcal{S}}}{m} \sum_{k=1}^{m} \sum_{t=3}^{\xi_k - 1} \mathbb{1}\{y_{v_{t-1}^{(k)}} = a\} \cdot \mathbb{1}\{y_{v_t^{(k)}} = b\} + \sum_{\substack{(u,v) \in E \\ u \in \mathcal{S} \text{ or } v \in \mathcal{S}}} \mathbb{1}\{y_u = a\} \cdot \mathbb{1}\{y_v = b\} \quad (3)$$

$$\widehat{\Pr}(a|b) = \widehat{\Pr}(a,b) \,/\, \sum_{c=1}^{H} \widehat{\Pr}(c,b)\,, \tag{4}$$

where $d_{\mathcal{S}} = |((\mathcal{S}\times V)\cap E)\backslash(\mathcal{S}\times\mathcal{S})|$ denotes the total number of outgoing edges from the seed nodes $\mathcal{S}$, and a, b, c take on arbitrary class values.

Proof The proof of asymptotic unbiasedness follows from an application of the renewal reward theorem (Brémaud 1999). The details are omitted due to space constraints. ■

Calibration of Estimated Parameters To construct well-calibrated confidence intervals for the estimated parameters, we propose to utilize bootstrap resampling (Efron 1979). In fact, this step can be performed as sampling progresses—by monitoring the confidence intervals, the practitioner can determine adaptively if more samples need to be collected.

For tours sampling, since the estimates computed from each tour are independent, and by Theorem 1 they are also unbiased, we can perform bootstrapping by treating the tours individually, sample with replacement, compute an estimate over the bootstrap sample, and repeat this process. We can then compute empirical confidence intervals over the bootstrap estimates. The theory of the bootstrap guarantees that the obtained confidence intervals will be well-calibrated. Among all the crawling methods under examination, tours sampling is the only approach capable of producing well-calibrated confidence intervals via resampling. This is due to the fact that BFS, FF, RW, and MH do not provide a list of node/edge samples that yield *i.i.d.* estimates of $\boldsymbol{\theta}$.

Experimental Evaluation

Dataset Description We perform experiments on five different attributed network datasets. As a preprocessing step, we take the giant component of all networks. Table 1 shows the summary statistics for each network after processing.

Facebook is a snapshot of the Purdue University Facebook network consisting of users who have listed their political views (whether or not they declare to be conservative).

Friendster-Large (Fri.-L.) and *Friendster-Small (Fri.-S.)* are processed from the entire Friendster social network crawl (Mouli et al. 2017). For Fri.-L., we take the subgraph containing all users with *age*, *gender*, and *marital status* listed in their profiles. For Fri.-S., we also include *zodiac*. We discretized the *age* attribute into four interval classes.

The observations we make are not restricted to social networks. We also experiment on two citation networks, *Communications (Comm.)* and *Computers*, both constructed from the NBER patent citations dataset (Hall, Jaffe, and Trajtenberg 2001).[2] The label of each patent indicates whether it was filed in a category related to comm. (computers).

Experiment Setup In each run of the simulation, we randomly select 50% nodes in the network to have observed labels, and the task is to infer the labels of the remaining nodes. Next, a labeled node is randomly selected as the seed node to initiate crawling for all the sampling methods.[3] In practice, querying a node will be associated with a certain cost, and we strictly control for the number of unique node-queries. For each method, we keep track of the parameter estimates and bootstrap confidence intervals as crawling progresses. We perform 10 runs of the simulation, and report the average performance and standard errors for all methods.

Table 1: Summary of Network Statistics

Dataset	$\lvert V\rvert$	$\lvert E\rvert$	Attributes
Facebook	14,643	336,034	*Political view*
Fri.-L.	3,146,011	47,660,702	*Age*, *gender*, *status*
Fri.-S.	1,120,930	19,342,990	*Age*, *gender*, *status*, *zodiac*
Comm.	855,172	5,269,278	*Communications-related*
Computers	855,172	5,269,278	*Cmputers-related*

Evaluation Criteria We measure the performance of the various network crawling methods in terms of:

- The quality of the RBC parameter estimates learned from a network sample crawled using that method. Specifically, we measure (*i*) the mean-absolute-error (MAE) between the *sample* estimate computed from the crawled sample and the *global* estimate computed from the entire graph, and (*ii*) the root-mean-square-error (RMSE) of the predicted class probabilities for the unlabeled nodes[4] using an RBC model equipped with the *sample* estimates.
- The quality of the confidence intervals obtained from the crawled sample, as measured by the coverage probability.

Evaluation of Estimation Performance Figure 1 shows the quality of the estimated parameters vs. the proportion of queried nodes in the network as crawling progresses. We observe that across all datasets, tour sampling (TS) consistently achieves smaller MAE in the estimated peer effects as well as lower RMSE in the predicted class probabilities. Also note that MH and RW usually outperfoms FF and BFS.

Evaluation of Calibration Performance Figure 2 shows the estimated bootstrap sampling distributions for two model parameters. [7] We observe that TS is the only method consistently capturing the global parameters. Figure 2 also shows examples of estimated peer effects in which the practitioner could be misled to draw the wrong conclusion regarding the

[2] While the edges in the citation networks are directed, we treat them as undirected edges in the experiments for simplicity.

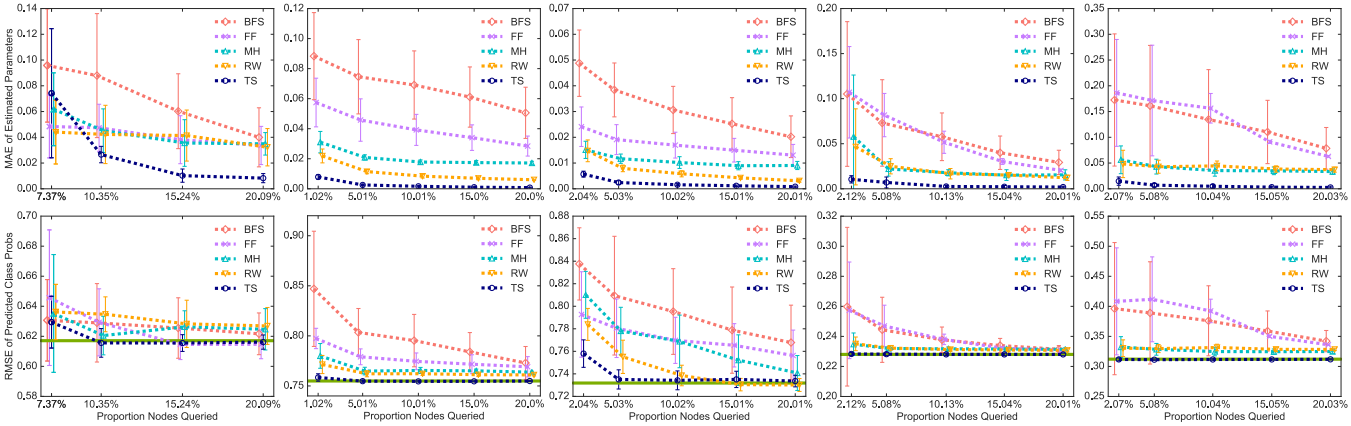
[3] In practice, one could always avoid querying unlabeled nodes; thus, we set all methods to crawl directly on the labeled subgraph.

[4] When predicting the class label for an unlabeled node, in addition to the attributes and class label of its neighbors, the attributes (but not the class label) of the unlabeled node are also available.

[5] For the Friendster results, the parenthesized attribute denotes the class label used for the prediction task, while all other attributes are used as features. The solid line in the RMSE plots correspond to the prediction error obtained using the *global* estimates. The plots are jittered horizontally to prevent the error bars from overlapping.

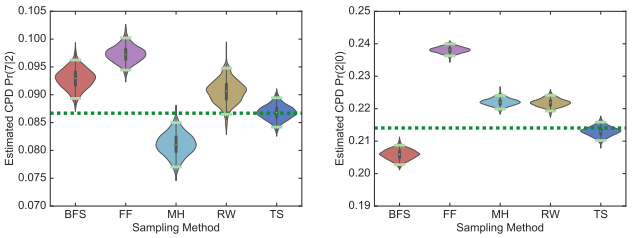
[6] The dashed line in each panel marks the values of the global parameter estimates, while the small horizontal bars on the violin plots indicate the estimated 95% bootstrap confidence interval.

[7] For BFS, FF, RW, and MH, we perform bootstrapping directly on the sampled nodes by treating each node as an *i.i.d.* instance.



(a) Facebook (b) Fri.-L. (Age) (c) Fri.-S. (Gender) (d) Comm. (e) Computers

Figure 1: MAE of estimated parameters (top row) and RMSE of predicted class probabilities (bottom row).[5]



(a) Probability of one having a friend of zodiac sign Taurus if one is 28 to 31 years old.

(b) Probability of one having a friend who is married given that one is female.

Figure 2: Estimated peer effects in Friendster-Small.[6]

existence of specific peer effects if their network data were crawled using conventional methods (BFS, FF, RW, MH).

To assess the calibration performance of each method, we compute 95% confidence intervals for the RBC parameters across 200 repeated trials, and calculate their empirical coverage probability (*i.e.*, the proportion of trials in which the estimated confidence interval contains the global estimate). Table 2 shows the results when 15% of each network have been crawled. We observe that the coverage probability for TS is higher than every other method across all datasets.

Table 2: Coverage probability of confidence intervals.

Dataset	BFS	FF	MH	RW	TS
Facebook	0.3333	0.4647	0.3570	0.3217	**0.9823**
Fri.-L. (Age)	0.0455	0.1136	0.1838	0.5789	**0.9424**
Fri.-L. (Status)	0.1136	0.0682	0.1131	0.5747	**0.9864**
Fri.-S. (Gender)	0.5146	0.3325	0.4345	0.5990	**0.9396**
Comm.	0.3333	0.3333	0.0000	0.0000	**1.0000**
Computers	0.0000	0.0000	0.0000	0.0000	**1.0000**

Conclusion

In this work, we have conducted the first empirical study to examine the impact of crawling on the estimation of peer effects in large-scale networks. Our experiments have shown that naively applying models to data collected by existing crawlers could lead to inaccurate parameter estimates and unreliable assessments of peer effects. To address this issue, we have developed crawl-aware relational estimation methods that produce accurate parameter estimates and well-calibrated confidence intervals with statistical guarantees.

Acknowledgments

This research is supported by NSF under contract numbers IIS-1149789, IIS-1546488, and IIS-1618690.

References

Ahmed, N.; Neville, J.; and Kompella, R. 2012. Network sampling designs for relational classification. In *ICWSM*.

Avrachenkov, K.; Ribeiro, B.; and Sreedharan, J. K. 2016. Inference in OSNs via lightweight partial crawls. In *SIGMETRICS*.

Brémaud, P. 1999. *Markov chains: Gibbs fields, Monte Carlo simulation and queues*. Texts in Applied Mathematics. Springer.

Efron, B. 1979. Bootstrap methods: Another look at the Jackknife. *Ann. Statist.* 7(1):1–26.

Gjoka, M.; Kurant, M.; Butts, C. T.; and Markopoulou, A. 2010. A walk in Facebook: Uniform sampling of users in online social networks. In *IEEE INFOCOM*.

Hall, B. H.; Jaffe, A.; and Trajtenberg, M. 2001. The NBER patent citation data file: Lessons, insights and methodological tools.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *SIGKDD*.

Macskassy, S. A., and Provost, F. J. 2007. Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.* 8:935–983.

Mouli, S. C.; Naik, A.; Ribeiro, B.; and Neville, J. 2017. Identifying user survival types via clustering of censored social network data. *arXiv:1703.03401*.

Neville, J.; Jensen, D.; and Gallagher, B. 2003. Simple estimators for relational Bayesian classifers. In *ICDM*, 609–612.