Combining Active Sampling with Parameter Estimation and Prediction in Single Networks

Joseph J. Pfeiffer III JPFEIFFER@PURDUE.EDU
Jennifer Neville NEVILLE@CS.PURDUE.EDU
Department of Computer Science, Purdue University, West Lafayette, IN

Paul N. Bennett PAUL.N.BENNETT@MICROSOFT.COM
Microsoft Research, Redmond, WA

Abstract

A typical assumption in network classification methods is that the *full* network is available to both learn the model and apply the model for prediction. Often this assumption is appropriate (publicly visible friendship links in social networks), however in other domains, while the underlying relational structure exists, there may be a cost associated with acquiring the edges. In this preliminary work we explore the problem domain of *active sampling*—where our goal is to maximize the number of positive (e.g., fraudulent) nodes identified, while simultaneously querying for network structure that is likely to improve estimates. We outline the problem domain formally and discuss five subdomains that are likely to be observed in real world scenarios. For our key finding, we show when the parameter estimates are learned from the distribution of labeled samples they are biased with respect to the parameters for the distribution of unlabeled samples, which negatively impacts the number of positive instances recalled. Additionally, we demonstrate that the estimation of the generative distribution from the labeled samples is also biased.

1 Introduction

With the emergence of social networks, large amounts of literature has focused on classification models tailored to the *relational* domain. These models make the underlying assumption that the full network is available to the learner. In many domains this is reasonable; for example, Facebook and Twitter users frequently have public friend listings.

However, in some domains the assumption of network availability is unreasonable. For example, consider a domain where a learner aims to identify positively labeled individuals in a network, such as finding fraudulent users in a phone network. Here, the majority of users are unlikely to be involved in fraudulent activity, meaning access to their phone records (and the underlying network) will be limited for privacy reasons. In another setting, a web crawler looking for new pages cannot know the entire network; rather, it must estimate which of the known pages that have not yet been crawled is most likely to have relevant information for possible users or contain links to new relevant pages. In both of these examples, we observe edges as we observe node labels. Further, there is *cost* associated with investigating nodes in these examples; investigating a potential fraudulent user can take a considerable amount of time, while we have a finite amount of computational resources available to crawl webpages.

In this domain of *active sampling* (Pfeiffer III et al., 2012), we are given the goal of maximizing the number of positive (e.g., fraudulent) nodes identified, while simultaneously querying for network structure. One benefit when we acquire a node during active sampling is that we observe connections between the node and its unlabeled neighbors. For example, when the web crawler examines an unlabeled page, it may give us new hyperlinks. Similarly, when we investigate a user for fraudulent behavior that investigation will reveal the user's phone records. This gives us relationships to nodes in the network that are still unlabeled; we term these nodes the *border*. The amount of information available to the learner, and when it becomes available, can change depending on the application.

Proceedings of the ICML Structured Learning Workshop, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

We identify five subdomains of interest, and discuss possible applications of each. For some of these subdomains, intrinsic attributes of unlabeled nodes may be unavailable, meaning relational information is key to prediction.

While the acquisition of relationships may help us to identify more positive instances, acquiring structure (and/or features) in this incremental fashion results in a *partial* observation of the data sampled from the underlying complete dataset. Specifically, considering that the positive acquisitions will (hopefully) appear at a rate greater than random, we can fall into a situation where there is a bias between the observed graph versus the underlying graph, as well as between the observed graph and remaining samples we are trying to predict. Should this occur, it makes the decisions of the sampler unreliable, as samplers assume accurate estimates when deciding which instance to acquire. We discuss this problem at length, and demonstrate that bias occurs when positives are sampled at a greater rate than occur in the network.

This work is divided as follows: in section 2 we discuss related work, while in section 3 we outline the five subdomains. In section 4 we formally define the domain while discussing the problems with estimation, while in section 5 we demonstrate the estimation bias empirically. Lastly, we conclude in section 6.

2 Related Work

Our work borrows elements from Active Learning, but with an orthogonal goal. In the typical active learning scenario labels are expensive to acquire, meaning the active learner needs to acquire nodes which learn the true model as quickly as possible (Bilgic & Getoor, 2008; Kuwadekar & Neville, 2011). Further, in relational active learning the network is assumed observed. In contrast, in active sampling we focus on acquiring instances whose label is likely positive and avoid instances which are likely negative, as well as acquire network structure throughout the process.

As part of this work, we address parts of the problem domains defined in (Namata et al., 2012) and (Garnett et al., 2012). In (Namata et al., 2012), the authors have a fixed set of instances they would like to estimate, but cannot query them directly. The authors search for neighboring nodes likely to provide information on the desired set. In the problem of *active search* defined by (Garnett et al., 2012), the authors define a similar problem structure where a classifier looks for positive instances to label, but with a fully observed network. A key component of both of these works is the estimation of the probability that a node is positive, given the other nodes in the network. However, in both of these works it is possible the samples are biased towards a particular label, which could lead to a distribution of labeled samples which differs from the unlabeled samples. Where the authors largely sidestep this possible bias, we focus on it, determining if this problem exists in the data. Lastly, (Pfeiffer III et al., 2012) was preliminary work which introduced active sampling. We define the problem in a formal manner and demonstrate their assumption of a stationary border distribution was incorrect.

3 Problem Overview and Scenarios

When we perform active sampling, we have the explicit goal of acquiring a large number of positive instances while minimizing acquisitions of negative instances. That is, we have a sampler which iteratively chooses instances to be labeled by an expert, or outside oracle. Additionally, as sample labels are acquired we receive *partial* information about the remaining unlabeled instances. That is, we receive information in addition to the labels of acquired samples, information about the unlabeled instances such as features, or information about edges to labeled samples. We use that additional information, along with the previously acquired labels, to model the distribution of the remaining samples, and allow the sampler to utilize that distribution when choosing from the remaining instances. However, as our sampler will acquire predominately positive samples, the core question in this work is whether the distribution of *labeled*, or sampled, instances will bias our parameter estimates with respect to the distribution of *unlabeled* instances.

This problem domain can be split into various individual cases, or subdomains, depending on what information becomes available to the learner during the process. Problem scenarios may differ depending on the subdomain: we outline five reasonable cases and give examples of an application that fits each case.

Case 1. In the "standard" within-network relational learning scenario, the data available at to us the beginning of the process contains all relationships and all attributes for all nodes. Thus, all that we iteratively acquire is the labels, which we use to estimate the likelihood of the remaining instances being positive. This is the scenario presented in (Garnett et al., 2012).

A real-world application which falls into this subdomain is companies who maintain a social network for their users (such as Facebook or Twitter), and are interested in investigating and removing fraudulent or offensive users. In this scenario, the company has all friendship links and individuals' attributes given to them by their user base, but must take time to investigate individuals' content by hand to determine if the account is being used in a fraudulent or illegal manner.

Case 2. In this scenario we begin with the network obscured, with only a handful of instances to use when we begin our search. As we iteratively choose which instances to label, we also acquire the labeled instances' relationships, which gives us partial relational information about their neighbors. In this case, we acquire the intrinsic attributes of both the acquired node and its neighboring nodes.

One application in this subdomain is searching for fraudulent trading behavior, where the investigation of one individual reveals their phone calls and emails to other individuals. In this case we are able to gather feature information of the unlabeled neighbors from documents found when investigating the labeled node.

Case 3. This case is similar to case 2; however, we only acquire attributes for labeled instances. Thus, we do not have any intrinsic features for the unlabeled instances and must estimate their distribution using only the relationships provided by their labeled neighbors. This instance is possibly the most difficult, as attributes of an instance are unavailable until the investigator takes the time to investigate the node. The nature of this problem requires us to use relational learning, as all intrinsic features are unavailable for unlabeled nodes. These features are ones that may not be known to friends, like financial records.

Case 4. This is a generalization of cases 2 and 3, where *part* of our intrinsic attributes become available when our neighbors are acquired, but other features are only acquired when an instance is labeled.

For example, when considering the fraud scenario, information from documents (such as emails) may give us attributes of an investigated node's neighbors, but other features such as financial records of those neighbors remain hidden until those nodes are investigated.

Case 5. Lastly, we have a generalization of cases 2-4, where the existence of all nodes is known initially. Further, we may have a handful of attributes for all nodes which is also known. This could exist in the case of brokers, where we may have access to general information about all registered brokers.

4 Formalized Domain

In this section we discuss notation to be used throughout the paper, as well as develop the problem domain more formally than in (Pfeiffer III et al., 2012).

Let capitalized letters indicate a variable and a bolded capitalized letter indicate a vector of variables. Let $\mathbf{X} \in \mathcal{X}$ indicate an *attribute* vector from a space of attribute vectors, while $Y \in \mathcal{Y}$ indicates a *label* from a space of labels (we focus on the label space $\mathcal{Y} = \{0, 1\}$).

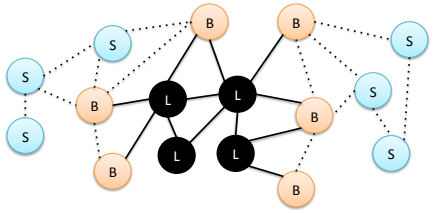


Figure 1. An example iteration of a network for active sampling. Prior labeled instances would have been chosen according to their probability of being positive.

Let $\mathbf{V}$ indicate a set of vertices, with $V_i \in \mathbf{V}$ indicating a specific vertex in the set of vertices. Each vertex V_i has a corresponding label Y_i and set of attributes $\mathbf{X}_i$. Define $\mathbf{E} = \mathbf{V} \times \mathbf{V}$ as the set of *possible* relationships between vertices, where $E_{ij} \in \{0, 1\}$, a '1' indicating a relationship between v_i, v_j and '0' indicating no relationship between v_i, v_j. Define a *graph* to be $G = \langle \mathbf{V}, \mathbf{E} \rangle$. For simplicity, G implies the entire graph, whether or not all of the edges are observed (more discussion later). Lastly, let $N(v_i)$ denote the Markov blanket, or *neighbors*, of a vertex v_i between it and other vertices for which the relationship exists, $N(v_i) = \{v_j | v_j \in \mathbf{V} \text{ and } E_{ij} = 1 \in \mathbf{E}\}$.

The emphasis on 'observed' versus 'unobserved' edges becomes important as we progress through this work; namely, these terms refer to whether the random variable E_{ij} has been *revealed to us*. Thus, observing the state of the relationship between v_i, v_j does not imply $E_{ij} = 0$ (this is the dominant observation). The set $\mathbf{E}$ is the ground truth of whether a relationship exists between every two instances in the network.

4.1 Active Sampling Domain

In active sampling, we have a number of various observed and unobserved sets of items, which can change from case to case. First, the label $Y_i \in \mathbf{Y}$ associated with the vertex $V_i \in \mathbf{V}$ is unobserved until the instance is acquired. In the majority of cases (2-5), the relationships $E_{ij} \in \mathbf{E}$ between vertices $V_i, V_j \in \mathbf{V}$ are also initially unobserved. The active sampling problem is one of acquiring labels and edges, thus variables $Y_i \in \mathbf{Y}$ and $E_{ij} \in \mathbf{E}$ can become observed. The variables that can either be unobserved or observed, depending on the case, is the vertex $V_i \in \mathbf{V}$ and the attributes $\mathbf{X}_i \in \mathbf{X}$.

We refer to the observed labels in $\mathbf{Y}$ as *labeled*, with the labeled set being $\mathbf{Y}^l$, meaning each $Y_i \in \mathbf{Y}^l$, the value $y_i = Y_i$ is known. As stated before, our edges are acquired as the active sampling process progresses, meaning our set of observed edges increases based on the variables $\mathbf{Y}^l$.

We define the set of *observed* edges ($\mathbf{E}^{obs}$) to be the neighbors of the vertices corresponding to the set $\mathbf{Y}^l$.

Graphically, this is represented in Figure 1, where the solid lines represent the observed edges. Formally:

$$E_{ij}^{obs} = \begin{cases} 1 & \text{if } E_{ij} = 1 \text{ and } \left(Y_i \in \mathbf{Y}^l \text{ or } Y_j \in \mathbf{Y}^l\right) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This results in the observed graph $G^{obs} = \langle \mathbf{V}, \mathbf{E}^{obs} \rangle$. Conversely, we wish to define the set of *unobserved* labels and edges. These can be described succinctly as $\mathbf{Y}^{un} = \mathbf{Y} \backslash \mathbf{Y}^l$ and $\mathbf{E}^{un} = \mathbf{E} \backslash \mathbf{E}^{obs}$.

The corresponding unobserved graph is $G^{un} = \langle \mathbf{V}, \mathbf{E}^{un} \rangle$. Additionally, we define the Markov blanket of a node in the observed graph representation: as $N^{obs}(v_i) = \left\{ v_j | v_j \in \mathbf{V} \text{ and } E_{ij}^{obs} = 1 \in \mathbf{E} \right\}$.

Lastly, we introduce disjoint subsets of unlabeled instances known as the *border* set and the *separate* set. The border set $\mathbf{Y}^b = \left\{ y_i | y_i \in \mathbf{Y}^{un} \text{ and } N^{obs}(y_i) \neq \emptyset \right\}$ is the set of instances which are unlabeled, but have at least one known edge due to a neighbor label having been observed. Similarly, separate set $\mathbf{Y}^s = \left\{ y_i | y_i \in \mathbf{Y}^{un} \text{ and } N^{obs}(y_i) = \emptyset \right\}$ is the unlabeled instances with no revealed neighbors.

In cases 2-4, the separate sets are naturally empty.

4.1.1 Subdomain Specific Notation

Aside from Case 1, we discuss how the attribute presence and vertex presence can change for each of the cases. For cases 2-4, note that the observed vertices $\mathbf{V}^{obs} = \left\{ v_i | N^{obs}(\mathbf{v}_i) \neq \emptyset \right\}$ are either the *neighbors* of labeled instances, or are labeled themselves.

Case 2. The attributes are acquired either when the node is acquired or when the node's neighbor is acquired. This results in our observed features being $\mathbf{X}^{obs} = \left\{ \mathbf{x}_i | N^{obs}(\mathbf{x}_i) \neq \emptyset \right\}$.

Case 3. The attributes are only acquired when the node is acquired, with the observed features being $\mathbf{X}^{obs} = \left\{ \mathbf{x}_i | y_i \in \mathbf{Y}^L \right\}$

Case 4. Now, the attributes have two different types: those which are acquired when their neighbor is acquired $\mathbf{X}^{obs_n} = \left\{ \mathbf{x}_i | N^{obs}(\mathbf{x}_i) \neq \emptyset \right\}$, and those which are acquired only on acquisition $\mathbf{X}^{obs_l} = \left\{ \mathbf{x}_i | y_i \in \mathbf{Y}^L \right\}$. The observed edges are then simply the union of these: $\mathbf{X}^{obs} = \mathbf{X}^{obs_n} \cup \mathbf{X}^{obs_l}$.

Case 5 Unlike the previous cases, this case guarantees knowledge of the node, which in turn may result in an additional given set of attributes $\mathbf{X}^{obs_g}$, in addition to the attributes defined in Case 4. Thus, $\mathbf{V}^{obs} = \mathbf{V}$ and $\mathbf{X}^{obs} = \mathbf{X}^{obs_g} \cup \mathbf{X}^{obs_n} \cup \mathbf{X}^{obs_l}$.

Naturally, the unobserved attributes are those which are not in the observed set $\mathbf{X}^{un} = \mathbf{X} \backslash X^{obs}$, while the unobserved vertices are also those not in the observed set $\mathbf{V}^{un} = \mathbf{V} \backslash \mathbf{V}^{obs}$.

4.2 Relationship Between the Observed Distribution and Generative Distribution

In active sampling, we have a number of conditionals which we now formalize. The first is that we still expect the *entire* network G to have been drawn according to $P_{\theta^*}(\mathbf{Y}|\mathbf{X}, G)$. As part of the active sampling process, we are acquiring positive instances in the network while trying to avoid sampling instances which do not match our desired labeling. In active sampling the edges are also revealed iteratively, as only edges which have at least one labeled endpoint are observed. Thus, although the data was drawn from the generating distribution, we must estimate according to: $P_{\theta^{obs}}\left(\mathbf{Y}|\mathbf{X}, \mathbf{E}^{obs}\right)$.

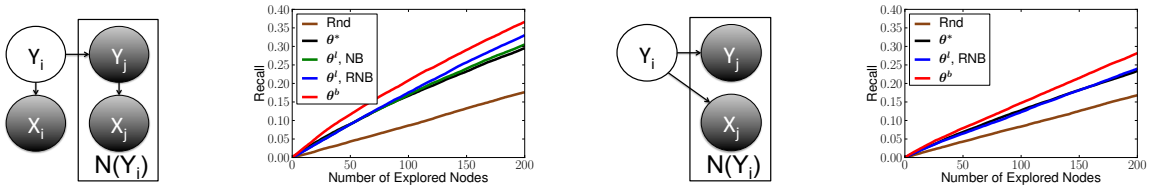
Taking the partially observed sampling scenario with the most information, where the attributes and vertices are known but edges and labels are unobserved (case 2), let $\mathcal{P}(\mathbf{E}^{un})$ be the power set of the combinations of possible values of the unobserved edges, with $\mathbf{E}_P^{un} \in \mathcal{P}(\mathbf{E}^{un})$ being a particular combination of possible edge values. Similarly, let $\mathcal{P}(\mathbf{Y}^{un})$ be the power set of the combinations of possible values of the unobserved labels, with $\mathbf{Y}_P^{un} \in \mathcal{P}(\mathbf{Y}^{un})$ being a particular combination of possible labels. Then, ideally, we can estimate:

$$P(\theta) = \int_{\mathbf{Y}_P^{un} \mathbf{E}_P^{un}} \Big[P\left(\theta | \mathbf{Y}_P^{un}, \mathbf{E}_P^{un}, \mathbf{Y}^l, \mathbf{X}, \mathbf{E}^{obs}\right) P\left(\mathbf{Y}_P^{un}, \mathbf{E}_P^{un} | \mathbf{Y}^l, \mathbf{X}, \mathbf{E}^{obs}\right) d\mathbf{Y}_P^{un} \mathbf{E}_P^{un} \Big]$$

Once a distribution for the parameters θ has been defined, a learner can use either maximum likelihood estimation (MLE) or maximum pseudolikelihood estimation (MPLE) to maximize and recover an estimate for θ^*(Xiang & Neville, 2008).

There are two problems with this approach, both involving the conditional distribution of the unobserved edges. First, the number of unobserved edges is $O(|\mathbf{V}|^2)$, meaning that computation in even modestly-sized networks is intractable. Secondly, there is no intuitive way to model $P\left(\mathbf{E}_P^{un} | G^{obs}\right)$ – the field of link prediction largely centers on heuristics based on hop distances, with edges presumed missing at random, or (in a dynamic sense) occurring at a subsequent point in time. Furthermore, in edge prediction most edges are presumed present, with only a handful of edges which are unobserved (Liben-Nowell & Kleinberg, 2007). This is not the case in active sampling, when the far majority of edges are unobserved.

The situation becomes more complicated when we remove the assumptions that the attributes and vertices are observed. In such a situation we must also marginalize over the possible number of nodes, as well



(a) Model (b) Recall (c) Model (d) Recall

Figure 2. Comparison of recalls for different parameters for Case 2 (a-b) and Case 3 (c-d)

as the assignments of attributes to those nodes. This further limits the feasibility of computing the generative distribution directly, and makes using it for producing estimates on the known vertices difficult.

As modeling the distribution of unobserved edges is an intractable proposition, we must work with the observed edges to model $P_{\theta^{obs}}\left(\mathbf{Y}|\mathbf{X},\mathbf{E}^{obs}\right)$, where θ^{obs} are the parameters which maximize this distribution. We focus on estimating two distributions involving the border and labeled instances: $P_{\theta^l}\left(\mathbf{Y}_L|\mathbf{X},\mathbf{E}^{obs}\right)$, which is the joint distribution of the labeled border instances, and $P_{\theta^b}\left(\mathbf{Y}_B|\mathbf{Y}_L,\mathbf{X},\mathbf{E}^{obs}\right)$, which is the joint distribution of the border instances *given* the labeled instances. As the observed graph is not drawn from the generating distribution, and is instead drawn from an alternative distribution, we intend to examine whether the parameters match between the different distributions. Namely, do the border parameters equal the generative parameters $\theta^b = \theta^*$, and do the border parameters equal the labeled parameters $\theta^b = \theta^l$? The second case is of particular interest to us as θ^l can be estimated directly from the labeled dataset. If these parameters are close, then we can use the labeled sample to estimate our border.

5 Experiments

5.1 Datasets

We test the parameter bias on two datasets: one synthetic and one real-world. The real-world dataset is a collection of instances from the Purdue Facebook network, containing approximately 6000 nodes. We chose the feature "Conservative or Not Conservative" as our label, with the attributes being "Christian or Not Christian" and "Male or Female". For the synthetic network we began with the the Rovira email network (Guimera et al., 2003) and synthetically generated labels and 2 features on the structure, resulting in a network with label correlation 0.18.

5.2 Models

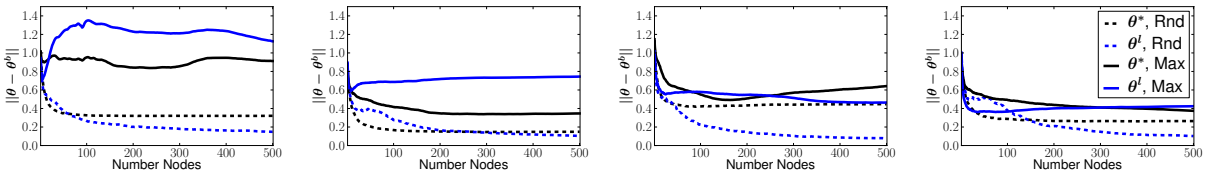
The conditionals for each method are in the form of Relational Naive Bayes (RNB) (Neville & Jensen, 2007). We compare parameters estimated from the labeled set (θ^l) against the true border parameters θ^b. Additionally, we compare the true generative parameters θ^* against the true border parameters θ^b, to determine how close the distribution of border nodes is to the generative distribution. Both θ^b and θ^* use unlabeled data for estimation (as well as labeled data), which allows us to determine how well estimation of θ^l from just the labeled set matches the desired parameters θ^b, as well as the generative parameters θ^*.

For estimation, we maximize the pseudolikelihood by treating the local conditional distributions of the nodes as independent samples (Xiang & Neville, 2008), and examine the performance of subdomains 2 and 3. Note that as the other cases are generalizations of 2 and 3, should bias exist in both of these cases it should exist for all cases. For cases 2 and 3, we use the graphical models shown in Figures 2.a and 2.c as our local conditionals. For case 2, a node's intrinsic features are available to the classifier, meaning we have the unlabeled nodes condition on their intrinsic features and their neighbors' labels. For case 3 the intrinsic features are unavailable to us, meaning we must condition the unlabeled instances solely on their relational information, which is their neighbors' features and labels.

5.3 Methodology

Our main goal with evaluation is to determine whether our estimates are *biased* when instances are sampled to maximize the number of positive instances. To gauge the parameter distance, we use the mean squared error between the parameters learned from the labeled set (θ^l) and the true border set parameters θ^b. One concern is ensuring that the methods are given the same set of labeled (and border) instances, meaning they have the same target distribution. To this end, we choose the most probable instance according to the θ^b parameters. All methods are given this sample at each step, meaning they have the same set for evaluation.

We then compare the distances of the parameters θ^l against θ^b, evaluating the distance when nodes are sampled using the most probable selection versus a random sampling. This allows us to determine whether the observed parameter differences are due to the preference for positively selected instances, or simply due to variance from the sample.



(a) Case 2 - Rovira (b) Case 2 - Facebook (c) Case 3 - Rovira (d) Case 3 - Facebook

Figure 3. Parameter biases for cases 2 and 3.

To randomize the experiments, each trial begins with all methods being given 5 (paired) randomly selected positive instances. We conduct 100 trials for each dataset, and test cases 2 and 3 to determine whether the estimation bias is problematic for different cases.

5.4 Results

To begin, we test the recall of different samplers provided with parameters estimated from each distribution (θ^l, θ^b and θ^*), to ensure we are accumulating positive samples at a proportion greater than random for cases 2 (Figure 2.b) and 3 (Figure 2.d). We note that all methods outperform the random sampler; further, we also see (a) when the sampler is provided with the distribution based on the border parameters θ^b it outperforms a sampler using the θ^* distribution, and (b) that the sampler using the labeled set parameters θ^l is performing fairly well, although, as expected, not as well as the sampler using the distribution based on the border parameters θ^b. Additionally, for case 2 we see that the relational classifier outperforms the IID, meaning there is considerable information to be found in the relationships. Naturally, case 3 cannot have an IID classifier, as intrinsic features are unavailable.

Next, in Figure 3a,b, we compare case 2's parameter bias for both random selection and most probable positive greedy selection. For the random selection, the difference between the labeled set parameters and true border parameters converges as we acquire more nodes, due to the larger sample size resulting in smaller variance. In contrast, when greedily sampling we do not converge: the labeled set estimates are biased with respect to the true parameters.

Turning to case 3 (Figure 3c,d), we see similar results as in case 2, with considerable bias when we actively acquire positive instances. As cases 4 and 5 are generalizations of cases 2 and 3, this implies the bias exists for *all* problem cases.

6 Conclusions and Future Work

In this preliminary work we have introduced the problem of active sampling, where nodes are investigated and labeled without full access to the entire network. We began with a thorough description of the problem formulations, defining five problems that lie in this domain. We then demonstrated that this domain exhibits bias due to the oversampling of positive instances relative to negative.

This domain has multiple theoretical questions remaining to determine the cause of this bias and correct it. Such solutions pave the way for additional questions, such as the introduction of a utility function which accounts for a node's possible influence on other border nodes, or graph models which can estimate the conditional distributions of the unobserved edges given the observed graph.

Acknowledgements

This research is supported by NSF under contract numbers IIS-1017898, IIS-1149789, and IIS-1219015. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of NSF or the U.S. Government.

References

Bilgic, M. and Getoor, L. Effective label acquisition for collective classification. In *SIGKDD*, 2008.

Garnett, R., Krishnamurthy, Y., Xiong, X., Schneider, J., and Mann, R. Bayesian optimal active search and surveying. In *ICML*, 2012.

Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., and Arenas, A. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68:065103, Dec 2003.

Kuwadekar, A. and Neville, J. Relational active learning for joint collective classification models. In *ICML*, 2011.

Liben-Nowell, D. and Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58(7), May 2007.

Namata, G. M., London, B., Getoor, L., and Huang, B. Query-driven active surveying for collective classification. In *MLG*, 2012.

Neville, J. and Jensen, D. Relational dependency networks. *JMLR*, 8:2007, 2007.

Pfeiffer III, J. J., Neville, J., and Bennett, P. N. Active sampling of networks. In *MLG*, 2012.

Xiang, R. and Neville, J. Pseudolikelihood em for within-network relational learning. In *ICDM*, 2008.