Network Quantification Despite Biased Labels

Lei Tang
Computer Science and Engineering
Arizona State University
Tempe, AZ, USA
L.Tang@asu.edu

Huiji Gao
Computer Science and Engineering
Arizona State University
Tempe, AZ, USA
Huiji.Gao@asu.edu

Huan Liu
Computer Science and Engineering
Arizona State University
Tempe, AZ, USA
HuanLiu@asu.edu

ABSTRACT

The increasing availability of participatory web and social media presents enormous opportunities to study human relations and collective behaviors. Many applications involving decision making want to obtain certain generalized properties about the population in a network, such as the proportion of actors given a category, instead of the category of individuals. While data mining and machine learning researchers have developed many methods for link-based classification or relational learning, most are optimized to classify individual nodes in a network. In order to accurately estimate the prevalence of one class in a network, some *quantification* method has to be used. In this work, two kinds of approaches are presented: quantification based on classification or quantification based on link analysis. Extensive experiments are conducted on several representative network data, with interesting findings reported concerning efficacy and robustness of different quantification methods, providing insights to further quantify the ebb and flow of online collective behaviors at macro-level.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data Mining*; H.4.2 [**Information Systems Applications**]: Decision Support

General Terms

Algorithm, Experimentation

Keywords

Network Quantification, Prevalence Estimation, Classification-Based Quantification, Link-Based Quantification

1. INTRODUCTION

In many applications and domains, it is the aggregated mass, or in other words, the prevalence of one class, that plays a key role in decision making. Here are some examples:

- In the US presidential election, the candidate who wins a plurality of individual votes in a state wins the state vote.
- A senator might need to collect the mass opinion and prioritize requests based on the urgency and the number of requests [18].
- An accurate estimation of the prevalence of influenza incidents in certain area can help the authority to allocate attentions, money and resources accordingly [12].
- Companies may want to estimate the proportion of positive or negative responses from customers to take corresponding strategic actions toward a new-generation product.

All the example above share one common characteristic: it is the aggregated mass of certain properties that matters.

The problem of accurately estimating the prevalence of one class in samples is referred as *quantification* [5]. It has been studied in various domains. For example, in the medical statistics field, Zhou et al. estimate the prevalence of a disease in a test population with an imperfect binary diagnostic test with known sensitivity and specificity [24]; Sociologists perform content analysis of blog posts to examine the support with respect to two different candidates [10]; Forman compares various strategies of quantification, aiming at quantifying the counts and costs of different classes in HP customer calls [5] so that HP can allocate human resources accordingly. Quantification is also applied to help semi-supervised learning when labeled samples and unlabeled samples follow different class distributions [23].

Quantification calls for attention because 1) The basic assumption of training and test sharing the same distribution as the foundation in many machine learning techniques are not necessarily true in reality; 2) Many classifiers are optimized for individual predictions. It might induce biases in quantification, especially when the class are imbalanced and insufficient; 3) The accuracy of classifiers can be highly dependent on the availability of training samples, which often involves tremendous human efforts. However, the prevalence might be estimated precisely even with few labeled samples if a robust quantification method is exploited.

All the aforementioned studies focus on quantification when data are presented in conventional attribute format. But in many situations, the data are presented in a relational network, such as the citations among papers, the transportation network and financial transactions between different entities. Recently, with the expanded use of web and social

MLG'10, July 24-25, 2010, Washington, DC, USA

media, oceans of user interaction data are produced in network format. This flood of network data provides valuable opportunities to study collective behavior such as the political views of users, the happiness of people, the opinions and sentiments of online users, the likelihood of product sales online, etc. Essentially, networks offers a new type of information source.

With abundant information provided online, we aim to quantify the collective behavior, i.e., the number of users that are involved in certain type of activities, preferences, or behaviors. Generally, both attribute data (e.g., user tweets, status updates, blog posts, tags, shared content) and network data (e.g., user interaction, friendship network, following/follower network) are available. Ultimately, we hope to exploit both kinds of information collected from social media to quantify the prevalence of users of certain classes.

However, network data are different from conventional attribute data that has been well studied in traditional data mining. When data instances are connected in a network, they are not independently identically distributed, hence collective inference is commonly used for prediction [15]. It is not clear how the collective inference process would affect the final quantification performance, which, in part, motivates us to develop this piece of work. For another, it is not surprising to have a network with millions of entities. Existing quantification methods rely on classification thus collective inference, which can be too computational expensive for practical use. It is demanding to develop efficient quantification methods for large-scale networks.

As an initial attempt, we concentrate on prevalence estimation with network data in this work. We first formally define the network quantification problem and emphasize its difference from conventional classification. Higher classification accuracy does not necessarily lead to better prevalence estimation. In section 3, we review existing quantification methods based on classification and discuss how it can migrate to handle network data. Since collective inference can be time consuming, we propose a simple link-based quantification method which does not require individual predictions, thus eliminating the collective inference procedure. We conduct comprehensive comparison and report empirical results on several benchmark network data sets in section 5, with concluding remarks and future work in Section 6.

2. NETWORK QUANTIFICATION

The expanded availability of web and social media has lead to the flourish of network data. The kind of interaction can provide valuable information to study collective behavior. Suppose the behavior can be captured by certain classes such as whether a user supports a certain political view, whether one likes one product, whether he would like to vote for a presidential candidate, etc. The categories can also be generalized to properties such as locations, preferences or sentiments of online users, and other attributes of entities in a network. In the simplest case, we consider binary classes (i.e., $\{+,-\}$) in this work, as most problems of multiple classes or multiple labels can often be converted into multiple binary classification problems.

Network Quantification is to estimate the prevalence of one class in a network. Without loss of generality, we always aim to compute the prevalence of the positive class. In particular, the problem is defined as follows:

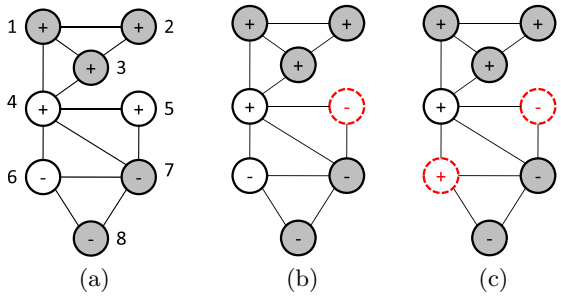


Figure 1: Shaded nodes are the ones with known labels while others are unknown. The nodes in red dashed outline denote the misclassified ones. (a) the ground truth of classes associated with each node in the network, $p(+) = 5/8$**; (b)** $accuracy = 2/3$**,** $\tilde{p}(+) = 4/8$**,** $|\tilde{p}(+) - p(+)| = 1/8$**; (c)** $accuracy = 1/3$**,** $\tilde{p}(+) = 5/8$**,** $|\tilde{p}(+) - p(+)| = 0$**.**

> **Given:** a network $N(V, E)$ where V is the vertex set, E is the edge set; known class values $\mathbf{y}^\ell \in \{+,-\}^\ell$ of certain nodes V^ℓ in the network;
> **Output:** the proportion $\tilde{p}(+)$ of nodes belonging to the positive class in the network.

One natural straightforward solution is *sampling*. That is, we count the number of positive and negative instances in $\mathbf{y}^\ell$ and compute the proportion directly. However, this relies on the assumption that the labeled nodes are representative of the whole population. In this work, we do not hinge on a carefully-designed sampling procedure, but allow the labeled nodes following a *different* distribution from the whole population. That is, the available labels are *biased*. Because 1) the labeling process might introduce biases in the available samples. Depending on the labeling procedure, different biases might be inserted into the labeled samples. For example, suppose the classes are about the sentiment of actors in a network and they are self-reported. For many users, the class information might be missing. Happy user may be more likely to share their mood online. 2) Networks often grow and evolve. For instance in social media, each day we have new users joining a network and new connections occurring between existing ones. It is impractical or impossible to always keep the labeled samples mirroring the whole network. Hence, we remove the strict assumption that labeled samples follow the same distribution as the whole network. In our setup, labeled samples are treated as obtained from a black box. They are biased and not guaranteed to follow the same class distribution as the whole population.

An alternative approach is to exploit the label information to construct a classifier, predict the labels of those unlabeled ones and count the proportion (which is denoted as *Classify & Count* (CC) in [5]) . We emphasize that classification and quantification are two different, though highly correlated tasks. Existing within-network classification or relational learning techniques are optimized for individual predictions. The prediction performance is not necessarily correlated with network quantification accuracy. Figure 2 shows a toy example. Figure 1a shows the problem setup. We are given a network of 8 nodes. The shaded ones are those whose class values are known. Two instances of predictions are shown in Figures 1b and 1c, respectively, where the red nodes are misclassified. Evidently, the former one in Fig-

Given: network $N(V, E)$;
labels $\mathbf{y}^\ell$ of some nodes;
maximum number of hops k_{max};
Output: the prevalence of the positive class $\tilde{p}(+)$.

Build relational classifier based on N and $\mathbf{y}^\ell$;
Perform collective inference to obtain prediction score y^u;
CC: $\hat{p}(+) = \frac{1}{|V|}|\mathbf{y} > 0|$.
AC: perform cross validation on y^ℓ to estimate tpr and fpr;
estimate $\hat{p}(+)$ following Eq. (3).
for threshold $t = \min\{\mathbf{y}\}, \cdots, \max\{\mathbf{y}\}$
estimate $\hat{p}_t(+)$ following AC;
end
T50: $\hat{p}(+)$ is estimated as $\hat{p}_t(+)$ when $tpr - fpr = 50\%$.
MS: discard invalid $\hat{p}_t(+)$;
output $\hat{p}(+)$ the median of valid $\{\hat{p}_t(+)\}$.

Figure 2: Classification-based Quantification

ure 1b, though predicts more accurately, produces higher the quantification error (1/8 deviation from the ground truth). On the contrary, in the latter case, 2/3 of test data are misclassified but the quantification error is 0 as those misclassified ones happen to cancel out. Thus, in order to quantify the prevalence of classes accurately, we have to employ techniques that are designed for accurate quantification.

3. QUANTIFICATION BASED ON CLASSIFICATION

A common practice in quantification is to rely on an imperfect classifier to estimate the class prevalence. Essentially, the probability of a binary classifier to output positive predictions in the test data is:

$$p(pos) = p(pos|+)\tilde{p}(+) + p(pos|-)\tilde{p}(-) \quad (1)$$
$$= tpr \cdot \tilde{p}(+) + fpr \cdot (1 - \tilde{p}(+)) \quad (2)$$

where $p(pos|+)$ (and $p(pos|-)$) is the probability of predicting positive given the true class is positive (negative), which corresponds to the true (false) positive rate, and $\tilde{p}(+)$ reflects the ground truth prevalence of positives in the test set. Since tpr and fpr can be estimated from cross validation on training data, $p(pos)$ can be computed based on the individual predictions in the test data, the estimated true prevalence in the test data can be computed as

$$\tilde{p}(+) = \frac{p(pos) - fpr}{tpr - fpr}. \quad (3)$$

This corrected quantification method first appears in [14] and is denoted as *Adjusted Count* (AC) in [5].

As seen in Eq. (3), a default threshold value of a classifier might introduce a very small difference between tpr and fpr, leading to unreliable estimates. In the worse case, the solution does not exist when $tpr = fpr$. Hence, Forman proposed several heuristics to set classification thresholds resulting in different tpr and fpr values as to obtain more accurate quantification. One recommended heuristic (denoted as T50) is to set a threshold so that $tpr = 50\%$.

Another more reliable quantification method is *median sweep* (MS). As suggested by the name, MS exhausts all the possible classification thresholds resulting in distinctive tpr and fpr values. For each threshold, an estimate of the prevalence is obtained. The final prevalence quantity is the median of all the estimates.

Given: network $N(V, E)$;
labels $\mathbf{y}^\ell$ of some nodes;
maximum number of hops k_{max};
Output: the prevalence of the positive class $\tilde{p}(+)$.

for $k = 1, 2, \cdots, k_{max}$
for each node i in V
estimate $p(\hat{i}^k|+)$ and $p(\hat{i}^k|-)$ based on N and $\mathbf{y}^\ell$;
estimate $p(\hat{i}^k)$ in N;
compute the prevalence $\tilde{p}_i^k(+)$ based on Eq. (6);
end
end
discard all invalid $\tilde{p}_i^k(+)$;
output $\tilde{p}(+)$ as the median of valid $\{\tilde{p}_i^k(+)\}$.

Figure 3: Link-based Quantification

Fortunately, all the quantification methods above can be migrated to a network setting. The nuance is that we have to use a classifier tailored for network classification. As long as it outputs a prediction score for each node in the network, the classification-based quantification methods can be applied. In later experiments, we will study and compare these different quantifiers handling network data.

At present, one common strategy for prediction in a network is collective inference [11, 15]. It assumes the labels of one node depend on the labels (or attributes) of its neighbors, which is consistent with the homophily effect observed in many social networks [17]. An iterative process is required to determine the class labels for the unlabeled data in turn, so that the inconsistency between neighboring nodes is minimized. When networks scale to millions of nodes, this computation can be very expensive. Since the goal of quantification diverges from that of typical classification, *can we avoid the classification process and address quantification directly*? In the next section, we provide a simple solution.

4. QUANTIFICATION BASED ON LINKS

In this part, we present one link-based quantification method that does not involve a classification component. Inspired by Eq. (1), we model the connections to one node as a mixture of distributions conditioned on classes. Let $p(\hat{i})$ denote the probability of another node connecting to node i, $p(\hat{i}|+)$ is the probability of a connection from a node of positive class to vertex i, and $p(\hat{i}|-)$ the probability of a node of negative class connecting to vertex i. We have the following equation:

$$p(\hat{i}) = p(\hat{i}|+)\tilde{p}(+) + p(\hat{i}|-)\tilde{p}(-). \quad (4)$$

Note that $p(\hat{i})$ can be estimated directly from a network, and both $p(\hat{i}|+)$ and $p(\hat{i}|-)$ can be estimated from the labeled samples. As $\tilde{p}(+) + \tilde{p}(-) = 1$, it follows that

$$\tilde{p}(+) = \frac{p(\hat{i}) - p(\hat{i}|-)}{p(\hat{i}|+) - p(\hat{i}|-)}, \quad (5)$$

if $p(\hat{i}|+) \neq p(\hat{i}|-)$. Take node 4 in Figure 1a as an example. We have

$$p(\hat{4}) = 5/8,$$

since node 4 is connecting to 5 other nodes among all the 8 nodes. Similarly, we have

$$p(\hat{4}|+) = 2/3; \quad p(\hat{4}|-) = 1/2.$$

According to Eq. (5), it follows that

$$\tilde{p}(+) = \frac{5/8 - 1/2}{2/3 - 1/2} = 75\%.$$

Note that the solution in Eq. (5) is not defined if $p(\hat{i}|+) = p(\hat{i}|-)$. It may also reside outside the valid prevalence range between 0 and 1. In that case, we simply discard it. As each node i in the network can lead to an estimate, we collect all the valid estimates and output the median as the prevalence, because median is well known for its robustness.

Another issue we want to emphasize is the sparsity issue. The number of neighboring nodes to labeled samples might be too few, leading to insufficient valid estimates. In this case, we can expand the neighboring nodes by considering all the nodes that are k-hops away. Correspondingly, we replace $p(\hat{i})$ by $p(\hat{i}^k)$, which is *the probability of a node in the network that is exactly k-hops away from node i*. Similarly, we can substitute $p(\hat{i}|+)$ and $p(\hat{i}|-)$ by $p(\hat{i}^k|+)$ and $p(\hat{i}^k|-)$, respectively. Then, we have the following general form for prevalence estimation when considering nodes that are exactly k hops away from node i:

$$\tilde{p}_i^k(+) = \frac{p(\hat{i}^k) - p(\hat{i}^k|-)}{p(\hat{i}^k|+) - p(\hat{i}^k|-)}. \quad (6)$$

Eq. (5) is a special case when $k = 1$. As a small-world effect is observed in many networks, $k \leq 3$ often can generate sufficient valid estimates. The detailed algorithm of this link-based quantification is summarized in Figure 4.

5. EXPERIMENT SETUP

Below, we describe evaluation measures, benchmark data sets and quantification methods for comparison.

5.1 Evaluation Measure

Let $\hat{p}$ and p denote the estimation and the ground truth of prevalence, respectively. Following [5], we adopt three different measures for comparison: *bias*, *absolute error* and *KL divergence*.

- Bias: $\hat{p} - p$. Bias is to study whether a quantifier shows any bias toward overestimation or underestimation. But a positive bias and a negative bias in another run might lead to zero average bias, thus another natural solution is to use the *absolute error*.
- Absolute Error: $|\hat{p} - p|$. By averaging the error over benchmark data sets, we can have an overall sense of how a quantification method works. However, it seems reasonable to look at relative errors since an estimate of 1% when the ground truth is 5% should be worse than an estimate of 41% when the ground truth is 45%. So *KL divergence* can be exploited.
- KL divergence (a.k.a. *normalized cross entropy*):

$$D_{KL}(\hat{p}||p) = \hat{p}\log\frac{\hat{p}}{p} + (1-\hat{p})\log\frac{1-\hat{p}}{1-p}. \quad (7)$$

The divergence becomes zero if $\hat{p} = p$ and goes to infinity when p approaches to 0 or 1.

5.2 Data Sets

For comparison purpose, we use three benchmark network data sets[1]: CoRA, IMDb and Industry. CoRA comprises computer science research papers. The network is constructed based on the citations among them, with classes being the topics of each paper. IMDb is obtained from Internet Movie Database[2]. It contains movies released in the United States between 1996 and 2001, with class labels identifying whether the opening weekend box-office receipts will exceed $2 million. Two movies are connected if they share a production company. Industry contains companies that are linked via co-occurrence in text documents. The companies are classified into 12 industry sectors.

Table 1: Statistics of Different Network Data

Data	CoRA	IMDb	Industry
# nodes	4240	1169	2189
# edges	22516	51481	13062
density	0.0025	0.050	0.054
# classes	7	2	4
prevalence	$6.3 - 32.2\%$	42.7%	$12.2\% - 27.8\%$
diameter	16	7	8

Note that some data sets have more than one classes. Hence, we convert them into multiple binary classification problems by treating one class as positive and all the remaining ones as negative. Some classes have too few instances, thus they are not included as a task for quantification evaluation, but their connections are still considered. In total, there are 12 classes for prevalence estimation. Some other statistics concerning each network data are presented in Table 5.2.

To evaluate the effectiveness of different quantifiers, we fix the number of negative instances to be 100, and change the number of positive instances ranging from 10 to 100. In particular, we want to examine cases when the class distribution of labeled samples is highly skewed. In another case, we fix the number of negative instances to 500, and change the number of positive instances from 10 to 100 again. Then the number of positive instances in labeled samples is as low as around $2\% - 20\%$. Each setup is repeated for 10 runs, and the average results are reported.

5.3 Quantification Methods

In the experiments, both classification-based and link-based quantification are included for comparison. Classification-based quantification methods are:

- Classify and Count (CC);
- Adjusted Count (AC) based on Eq. (3);
- Probabilistic Classify and Count (pCC);
- Probabilistic Adjusted Count (pAC);
- T50 assigning a threshold such that $tpr = 50\%$ in Eq. (3);
- Median Sweep (MS). We exhaust threshold values of $0.01, 0.02, \cdots, 0.99$. Each threshold produces a prevalence estimate and the final output is the median among all the valid ones.

For classification with network data, we adopt wvRN implemented in NetKit-SRL [15] as it is recommended based

[1]http://netkit-srl.sourceforge.net/data.html

[2]http://www.imdb.com/

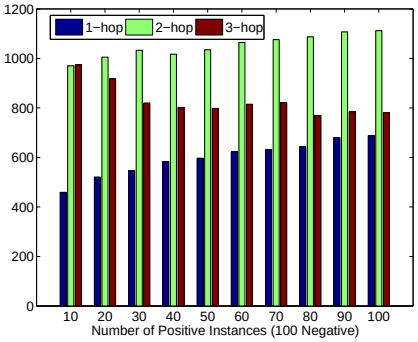


Figure 4: Valid Estimates by Connections of k Hops

on extensive comparison across different data. Since all classification-based methods need tpr and fpr to apply Eq. (3), we apply 10-fold cross validation to estimate tpr and fpr as suggested by [5] after we obtain prediction scores for all the nodes.

The link-based quantification (LBQ) is also compared. As for LBQ, it involves a parameter k_{max} which is defined to be the maximum number of hops to consider for computation. If we only consider 1-hop connections, essentially only those nodes that are 1-hop away from the labeled nodes can help quantification. The majority of the network structure is not considered. When we expand LBQ to consider connections of nodes that are 2-hop away, a large-portion of the network structure is taken into account. Hence more estimates based on Eq. (6) can be obtained. However, considering 3-hop connections might introduce too much noise, let alone the computational cost. Figure 5.3 shows the average number of valid estimates contributed from connections of exactly k hops on Industry data. A similar trend is observed on other data. Note that 2-hop produces the maximum number of valid estimates. 3-hop enforces LBQ to look at more connections, but most of them cannot produce a valid estimate based on Eq.(6). Hence, we only include the comparison when k_{max} is set to 2 or 3 (denoted as LBQ2 and LBQ3, respectively).

6. EXPERIMENTS

In this section, we compare different quantification methods based on several network data sets. In particular, we study the following questions:

- Is quantification necessary for network data?
- Which quantification method is more accurate?
- How efficient are different methods?

Due to space limit, we only report the performance of quantification methods averaged over different data sets and classes. Figures 5.3 and 5.3 show the quantification performance when we fix the number of negative instances to 100 and 500, respectively.

6.1 Is quantification necessary?

First, CC is highly unreliable. CC is biased depending on the class distribution in the training data as shown in Figure 5a With few positive instances, CC yields a negative bias; When positive instances increases, CC instead overestimates the class prevalence. When the negative instances increases to 500, almost all the methods shift toward negative bias. In this case, CC demonstrate a strong negative bias, implying that CC is quite sensitive to the bias as in the labeled samples.

The strong bias of CC is also verified by the high absolute error and KL-divergence. Most of the time, CC performs the worst. If we simply count the classification predictions, the quantification error can be huge. Hence, some correction must be adopted to reduce the error. If we use the probabilistic classify and count (pCC), the error is reduced, indicating a soft version of classification does help for quantification. However, when biases are presented in the labels, using a probabilistic prediction only provides limited refinement.

CC is even worse when we have a highly imbalanced training data as shown in Figure 5.3. As in this setting, the negative instances are increased to 500, whereas the number of positive instances ranges from 10 to 100, resulting in a highly skewed class distribution. This strong imbalance in training data results in a severe underestimate for CC as shown in Figure 6a. Both absolute error and KL-divergence are inflated. In general, both classification and link-based quantification methods outperform CC consistently as shown in Figures 5.3 and 5.3, confirming the necessity of quantification.

6.2 Which quantifier is more accurate?

Comparing different quantification methods, Median Sweep (MS) is no doubt the winner. It consistently outperforms other quantification methods in terms of absolute error and KL-divergence. This result is consistent with the empirical results on text data as reported in [5].

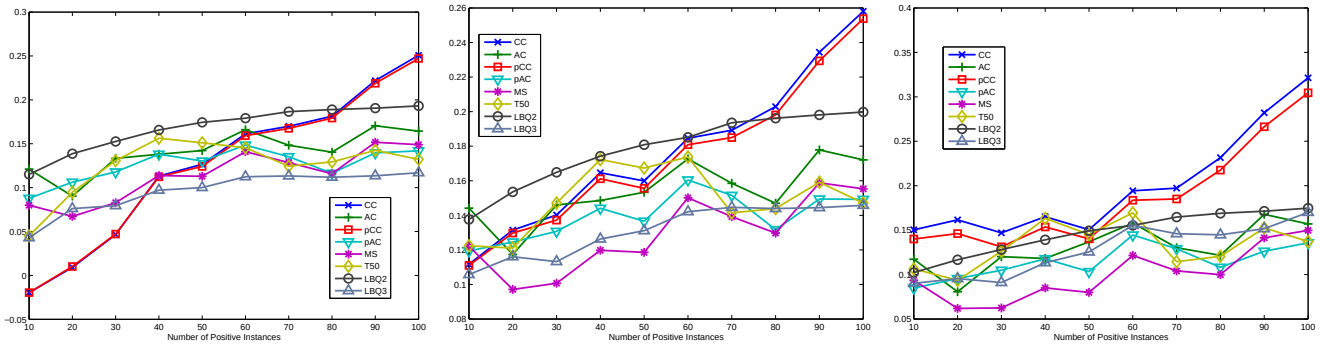
Besides MS, Probabilistic Adjusted Count (AC) seems to work fine. We notice that in [5], CC or AC based on probabilistic prediction is not included for comparison. However, based on our empirical result, it seems that soft classification often yields better quantification performance. However, AC or PAC does not always return a valid solution (i.e., a prevalence estimate outside [0, 1]), limiting its success[3].

LBQ, as shown in the figures, tends to yield lower absolute error or KL-divergence than simple CC and pCC. But it does not model the prevalence as accurate as classification-based quantification. On the other hand, its performance is sensitive to the number of hops for considering the links. However, LBQ is much more efficient than other classification based approaches, as we show in the next subsection,

6.3 How efficient is each method?

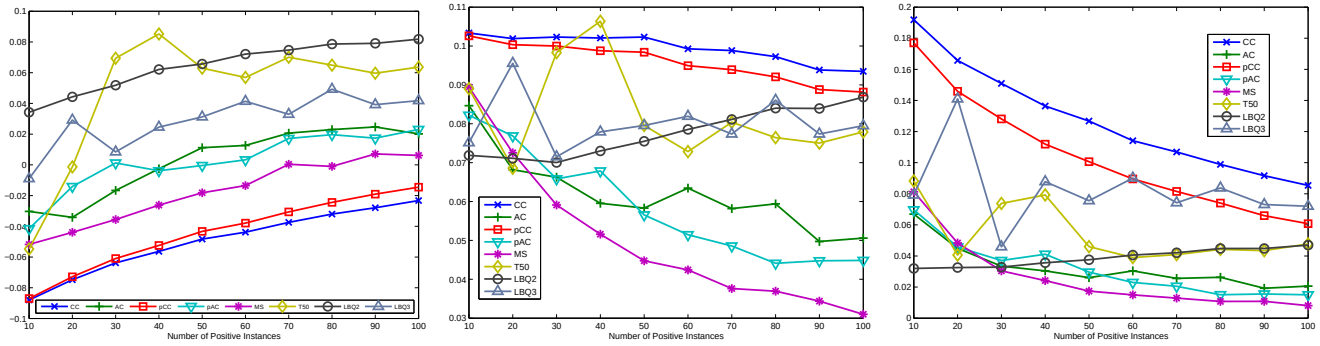
One advantage of LBQ is its efficiency. Since AC, T50 and MS all rely on individual prediction scores for quantification, collective inference has to be conducted, which often requires multiple scans of a given network thus time-consuming. Moreover, these methods need to estimate tpr and fpr in Eq. (3), involving tedious cross-validation. On the contrary, LBQ does not count on individual predictions, thus eliminating the necessity of collective inference and cross validation, saving enormous computational time. It only requires one scan of the network with update of some simple statistics. Thus, LBQ can be extremely efficient.

[3]In the figure, we exclude those invalid cases for AC and pAC while computing the average.



(a) Bias (b) Absolute Error (c) KL-Divergence

Figure 5: Performance when the number of negative instances for training is fixed to 100.



(a) Bias (b) Absolute Error (c) KL-Divergence

Figure 6: Performance when the number of negative instances for training is fixed to 500.

Table 2: Computational Time for Different Methods

Method	AC	T50	MS	LBQ
Classification	16 min 20s			0
Quantification	4s	14 s	17s	11s

Table 6.3 lists the average computation time on Industry data for different methods. Since our labeled samples are few, the cross validation for classification-based quantification does not take too much time. However, the major computational cost for classification cannot be evaded. It takes over 16 minutes to obtain individual predictions on a Intel Core 2 Duo 3.33GHz CPU. Comparatively, LBQ takes 11 mere seconds. Note that the benchmark data presented here is relatively small, with only thousands of nodes. Consider a network of much larger size (say millions of actors), the difference of link-based and classification-based quantification methods can be more drastic.

It seems that MS, pAC and LBQ all can correct the biases associated with labels for prevalence estimation. MS, by exhausting all the possible thresholds, tends to yield more accurate prevalence estimation. pAC performs reasonable well, but it sometimes outputs an invalid solution. Both MS and pAC rely on classification, hence their computational time can be tremendous if a given network is huge. Therefore, LBQ is suitable for initial guess when an immediate response of network quantification is preferred.

7. RELATED WORK

Estimating the hidden population in a social network is a classical problem studied in social science. This problem occurs for populations that are normally difficult to reach by random sampling, such as drug injectors [8], sex workers [16] and unregulated workers [2]. Traditional methods take the form of chain referral sampling. The best known approach in this form is snowball sampling [7], which allows the sampled individuals to provide information about their contacting friends. Klovdahl [13] proposes a "random walk" approach in which each wave of snowball sampling contains only one individual. Spreen and Zwaagstra [19] propose a combination of snowball and targeted sampling, termed "targeted personal network sampling," to analyze the structure of cocaine users' personal networks. Frank and Snijders [6] refine the chain-referral model using a one-wave snowball sampling to estimate the population size. Despite the efforts and progress, "How to draw a random (initial) sample" is still a key unsolved problem. As the selection of initial samples in a chain-referral model may cause bias into inferences, many other approaches such as key information sampling [3] and targeted sampling [22] are widely employed response to this deficiency. Recently, respondent driven sampling [8] is proposed for sampling design and population inference in a social network. By tracing the links in the underlying social network, the RDS process exploits the social structure to expand the sample and reduce its independence on the initial sample. All these aforementioned methods focusing on how

to choose qualified samples applicable to a hidden population. Different from quantification studied in this work, the structure of the network, including size and membership, is not available to researchers.

On the other hand, quantification is also calling attention in data mining field [5]. Some machine learning methods for accurately estimating the class distribution of a test set, using a training set that may have a substantially different distribution, have been designed and applied to various domains. Levy and Kass [14] apply a three-population model to derive estimators of the prevalence of bacteria in the population of the test; Zhou et al. estimate the prevalence of a disease in a test population with an imperfect binary diagnostic test with known sensitivity and specificity [24]; George Forman [4] compares 3 methods: classify & count, adjusted count, and mixture model, to count positives accurately despite the substantial bias in estimating class prevalence caused by inaccurate classification. Hopkins et al. develop an automated content analysis method to examine the support of thousands of people through their daily expressed opinions of blog posts about the U.S. presidency [9]. Quantification is also applied to help semi-supervised learning when labeled samples and unlabeled samples follow different class distributions [23]. George Forman [5] describes a variety of quantification methods and evaluates them with a suitable methodology, revealing which methods give reliable estimates when training data is scarce and the positive class is rare. Differently, in this work, we focus on quantification in a social network by dealing with bias of available labels.

8. CONCLUSIONS AND FUTURE WORK

This work, to our best knowledge, is the first work to study network quantification problem. That is, given some labeled samples from a network, how can we estimate the prevalence of certain classes in the network? This work is a summary of some ongoing work we are currently pursuing. We show that straightforward baseline approaches are error-prone especially when labeled samples are few and imbalanced. Two kinds of quantification are presented and compared. 1) Classification-based quantification do not model the prevalence directly, but hinge on post-processing to correct the bias with labels. One specific method Median Sweep (MS) is quite robust and outperforms other methods; 2) The link-based approach relies on link analysis to estimate the class prevalence. It does not require classification and prediction, thus saving tremendous computational cost. But its quantification performance is not comparable to MS. It remains an open problem to model the class prevalence more effectively and efficiently and we hope to encourage more research to address this network quantification problem.

As we discussed in the introduction, social media presents both network and attribute data. We plan to harness both types of information for effectively quantify online collective behavior. It requires further research to unify the flood of network data with quantification to estimate rare events such as the click-through rate of one class of advertisement in a population [1]. Recently, we proposed one social dimension based framework for network-based classification [20, 21]. It extracts social dimensions from a network and the classification performance outperforms representative collective inference based methods. It would be interesting to extend social-dimension based approach for quantification as well.

9. ACKNOWLEDGMENTS

This work is, in part, supported by AFOSR and ONR. We thank Professor Sun-Ki Chai for inspiring discussions.

10. REFERENCES

[1] D. Agarwal, A. Z. Broder, D. Chakrabarti, D. Diklic, V. Josifovski, and M. Sayyadian. Estimating rates of rare events at multiple resolutions. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–25, New York, NY, USA, 2007. ACM.

[2] H. D. M. R. T. N. Bernhardt, A. Documenting unregulated work: A survey of workplace violations in new york city. the future ofwork. 2006.

[3] E. Deaux and J. Callaghan. Key Informant Versus Self-Report Estimates of Health-Risk Behavior. *Evaluation Review*, 9(3):365, 1985.

[4] G. Forman. Positives accurately despite inaccurate classification. In *ECML*, 2005.

[5] G. Forman. Quantifying counts and costs via classification. *Data Min. Knowl. Discov.*, 17(2):164–206, 2008.

[6] O. Frank and T. Snijders. Estimating the size of hidden populations using snowball sampling. *JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM-*, 10:53–53, 1994.

[7] L. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 1961.

[8] D. Heckathorn. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems*, 44(2):174–199, 1997.

[9] D. Hopkins and G. King. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science*, 54(1):229–247, 2009.

[10] D. J. Hopkins and G. King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, January 2010.

[11] D. Jensen, J. Neville, and B. Gallagher. Why collective inference improves relational classification. In *KDD*, pages 593–598, 2004.

[12] B. D. Jones and F. R. Baumgartner. *The Politics of Attention: How Government Prioritizes Problems.* The University of Chicago Press, 2005.

[13] A. Klovdahl. âĂŸUrban Social Networks: Some Methodological Problems and Possibilities. *The Small World: A Volume of Recent Research Commemorating Ithiel de Sola Pool, Stanley Milgram, and Theodore Newcombe*, page 176, 1989.

[14] P. S. Levy and E. H. Kass. A three-population model for sequential screening for bacteriuria. *American Journal of Epidemiology*, 91(2):148–154, 1970.

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

[16] M. Malekinejad, L. Johnston, C. Kendall, L. Kerr, M. Rifkin, and G. Rutherford. Using respondent-driven sampling methodology for HIV biological and behavioral surveillance in international

settings: a systematic review. *AIDS and Behavior*, 12:105–130, 2008.

[17] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[18] D. C. Mutz. *Impersonal Influence: How Perceptions of Mass Collectives Affect Political Attitudes*. Cambridge University Press, 1998.

[19] M. Spreen and R. Zwaagstra. Personal network sampling, outdegree analysis and multilevel analysis: Introducing the network concept in studies of hidden populations. *International Sociology*, 9(4):475, 1994.

[20] L. Tang and H. Liu. Relational learning via latent social dimensions. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 817–826, New York, NY, USA, 2009. ACM.

[21] L. Tang and H. Liu. Scalable learning of collective behavior based on sparse social dimensions. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 1107–1116, New York, NY, USA, 2009. ACM.

[22] J. Watters and P. Biernacki. Targeted sampling: options for the study of hidden populations. *Social Problems*, 36(4):416–430, 1989.

[23] J. C. Xue and G. M. Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 897–906, New York, NY, USA, 2009. ACM.

[24] X.-H. Zhou, D. K. McClish, and N. A. Obuchowski. *Statistical Methods in Diagnostic Medicine*. Wiley, 2002.