Towards Quantifying Sampling Bias in Network Inference

Lisette Espín-Noboa[*]
GESIS & University of Koblenz-Landau
Lisette.Espin@gesis.org

Claudia Wagner
GESIS & University of Koblenz-Landau
Claudia.Wagner@gesis.org

Fariba Karimi
GESIS & University of Koblenz-Landau
Fariba.Karimi@gesis.org

Kristina Lerman
USC Information Sciences Institute
lerman@isi.edu

ABSTRACT

Relational inference leverages relationships between entities and links in a network to infer information about the network from a small sample. This method is often used when global information about the network is not available or difficult to obtain. However, how reliable is inference from a small labeled sample? How should the network be sampled, and what effect does it have on inference error? How does the structure of the network impact the sampling strategy? We address these questions by systematically examining how network sampling strategy and sample size affect accuracy of relational inference in networks. To this end, we generate a family of synthetic networks where nodes have a binary attribute and a tunable level of homophily. As expected, we find that in heterophilic networks, we can obtain good accuracy when only small samples of the network are initially labeled, regardless of the sampling strategy. Surprisingly, this is not the case for homophilic networks, and sampling strategies that work well in heterophilic networks lead to large inference errors. This finding suggests that the impact of network structure on relational classification is more complex than previously thought.

ACM Reference Format:
Lisette Espín-Noboa, Claudia Wagner, Fariba Karimi, and Kristina Lerman. 2018. Towards Quantifying Sampling Bias in Network Inference. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3184558.3191567

1 INTRODUCTION

Networks form the infrastructure of modern life, linking billions of people, organizations and devices via trillions of transactions. Solving today's problems and making critical decisions increasingly calls for mining massive data residing in such networks. Due to their size and complexity, it is often prohibitively costly for analysts to obtain a global view of the network and the data it contains. Instead, they can use relational machine learning methods to infer information about the network from a partial sample, e.g., infer the class of unlabeled nodes from the known classes of a few seed nodes. How reliable is such inference? How much impact does the choice of seeds have on inference error? How much does the structure of the network impact sampling strategy? In this work, we address some of these questions by systematically studying potential sources of bias in the relational inference process: Network Structure, Sampling, Relational Classification, and Collective Inference. Towards that goal, we dig deeper into how sampling strategies can be biased, and when this bias can be beneficial or disadvantageous for the inference process. New insights on how sampling impacts relational classification performance can potentially lead to new unbiased strategies.

Relational Classification. Relational classifiers propagate information through the network from the known labels of nodes to infer unknown labels. The classification performance is measured by how well all the nodes' labels can be recovered when only the labels of a few seed nodes are known. In [9] the authors outline the two main components of collective classification, which are the collective inference method and the relational classifier. They assess how various choices and combinations of components, as well as the percentage of labeled data used for training the method, impact the accuracy of the classification. Their results show that there are two sets of techniques that are preferable in different situations, namely when few versus many labels are known initially, and that link selection plays an important role similar to traditional feature selection. However, the authors did not explore how the network structure impacts performance of collective classification methods. Sen et al. close this gap by comparing four collective classification algorithms with a content-only classifier, which does not take the network into account, on networks that varied in link density and homophily [12]. They found that increasing link density improves the performance of collective classification and clearly outperforms content-only classifiers at all density levels. Moreover, *homophily*, which refers to the tendency of nodes with similar labels to be connected, further helps collective classifiers outperform content-only classifiers, except for very low levels of homophily (< 0.1), where content-only classifiers perform slightly better. While this work explores homophily and density of networks separately, more recent research investigates how these characteristics jointly impact performance. In [17] the authors show that as homophily and link density of the network increase, the accuracy of relational classification also increases. Similar to our study, that work focuses on balanced networks where nodes have a single binary attribute. However, the subgraph used for training is selected via random node sampling only.

[*] This work was part of her summer internship at USC-ISI in 2017.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3191567

Sampling Bias. Previous work has demonstrated that the estimates obtained from network samples collected by various crawlers can be inaccurate with respect to global [8] and local network statistics [14]. Two recent papers showed that the choice of the initial sample of labeled seed nodes can also affect attribute inference [1, 15]. However, these did not explore how properties of networks, such as homophily, affect the choice of seeds and classification performance.

Findings and Contributions. In this work, we focus on the attribute inference task and explore how the accuracy of collective inference in networks depends on the strategy used to create the initial set of labeled nodes. In summary, our main contributions are three-fold: (i) Using synthetic and empirical networks, we provide evidence that homophily plays a decisive role in the collective inference process: First, no sampling technique can beat a random classifier when networks are neutral (i.e., nodes connected at random regardless of their class label). Second, heterophilic networks are easy to classify with any sampling strategy and require a training sample of at least 5% of random nodes to achieve an unbiased classification. Finally, some sampling strategies that work well for heterophilic networks require larger samples for homophilic networks. Only methods that construct samples by selecting highest degree nodes first achieve good classification performance with small samples in both homophilic and heterophilic regimes. (ii) We show that link density influences classification performance under certain conditions: First, sampling methods that rank low-degree nodes first, benefit from networks with high link density. Second, homophilic networks with high link density require larger training samples for edge, mixed degrees, and snowball sampling. (iii) We discuss the impact of sampling strategies on relational classification using collective inference, and demonstrate that inference can be negatively affected by class imbalance.

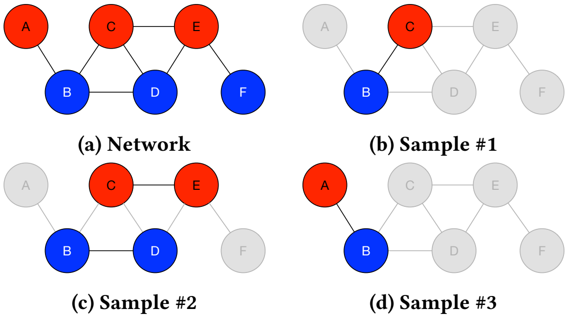
The remainder of this paper is organized as follows: In Section 2 we present background knowledge. Experiments, datasets and results are described in Section 3 followed by a discussion in Section 4. Finally, we present future work and conclusions in Section 5.

2 BACKGROUND

In this work, we are primarily interested in studying the influence of different sampling techniques on relational classification. We describe (i) networks of interest, (ii) the classification process and (iii) used network sampling techniques.

2.1 Attributed Networks

We formally define this as: Let $G = (V, E, F)$ be an attributed unweighted graph with $V = (v_1, ..., v_n)$ being a set of nodes, $E = \{(v_i, v_j)\} \in (V \times V)$ a set of either directed or undirected edges, and a set of feature vectors $F = (f_1, ..., f_n)$. Each feature vector $f_i = (f_i[1], ..., f_i[t])^T$ maps a node v_i to t (binary, numeric or categorical) attributes. A class label is defined as $c \in t$, and represents the attribute to be inferred in the classification process. Link density is described as the fraction of potential connections in G that are actual connections, that is $d = \frac{|E|}{N(N-1)}$ for directed, and $d = \frac{2|E|}{N(N-1)}$ for undirected networks. The average degree of the network captures the average number of edges per node: $\langle k \rangle = \frac{|E|}{N}$ for directed and $\langle k \rangle = \frac{2|E|}{N}$ for undirected networks. The network homophily H is captured by the fraction of same-class connections among the total number of edges $|E|$. Homophily values range from $H = 0.0$ to $H = 1.0$. Networks with homophily $H = 0.5$ are referred to as *neutral*, otherwise they are *heterophilic* if $H < 0.5$, or *homophilic* when $H > 0.5$. Class balance B captures the fraction of nodes under each class value. A network is *balanced* when all class values have the same number of nodes, otherwise it is an *unbalanced* network.



(a) Network (b) Sample #1 (c) Sample #2 (d) Sample #3

Figure 1: Example. This figure illustrates an unweighted undirected node-attributed network and three different samples. (a) Shows a heterophilic network with seven edges and six nodes. Each node is coloured either red (A, C, E) or blue (B, D, F). (b) Sample #1 shows a subgraph extracted by sampling two nodes. This sample includes nodes B and C, which reflect perfect heterophily. (c) Sample #2 shows a homophilic subgraph sampled by randomly picking two edges, C-E and B-D. (d) This sample is similar to Sample #1 as it reflects perfect heterophily. However in this case node F is 3-HOPs away from the nearest seed node (compared to 2-HOPs in Sample #1).

In this work, we focus on attributed networks whose edges are unweighted and undirected, and whose nodes are balanced along a binary feature (e.g., $c = color \in \{blue, red\}$). Figure 1a shows an example of such network, where nodes are assigned to one color, either blue or red, and since only 2 out of 7 edges are same-color connections this network is heterophilic ($H \approx 0.3$). This network is also balanced ($B = \frac{3}{6} = 0.5$) since the number of blue nodes ($n_b = 3$) is equal to the number of red nodes ($n_r = 3$).

Notice that in a realistic scenario, class balance B is often unknown, as well as homophily H. However, these two values can be inferred. For instance, balance can be approximated by randomly picking a set of nodes to extract their true class value and then infer class balance. Similarly, homophily can be approximated by randomly picking a set of edges. These approximations go beyond the scope of this work. For evaluation purposes we assume total knowledge of the network.

2.2 Relational Classification

Classification in networked data [5, 9, 12] learns correlations between attribute values of linked nodes from observed data and then uses them in a collective inference process that propagates predictions through the network. This process can be divided into four phases. First, a subgraph needs to be *sampled* from the network.

Second, a *local model* is learned by using information from nodes only (e.g., nodal attributes) and can be used as class priors later in the inference. Third, the *relational model* in contrast to the local model, learns information from nodes and their 1-HOP neighbourhood. Finally, once the models' parameters (i.e., probabilities) have been learned, the collective inference phase determines how the unknown values are estimated.

Every phase can be implemented in different ways [9]. Since our work focuses on the sampling phase, we keep fixed the other modules by (i) learning the local model as class priors from the nodes in the training sample, (ii) learning the relational model from the nodes and edges in the training sample using Bayesian statistics, and (iii) inferring estimate values using Relaxation labeling.

For simplicity, we focus on uni-variate network classification, which means that the linkage structure in the network is modeled with the class label and no information from other attributes. This setup is referred to as network-only Bayes classifier (NBC) in [9] to emphasize that additional local attributes of a node are ignored.

2.3 Network Sampling

The goal of sampling is to split the network into a *training* and a *testing* sample. First, a subgraph $\hat{G} = (\hat{V}, \hat{E}, \hat{F})$ is extracted from the network G in order to learn the model parameters. Nodes $\hat{V} \subset V$ that belong to the training sample $\hat{G}$ are called *seed nodes*, and we assume that their edges $\hat{E} = \{(\hat{v_i}, \hat{v_j})\} \in (\hat{V} \times \hat{V}) \subset (V \times V)$ and attributes $\hat{F} \subset F$ are known by the classification algorithm. For example, based on the information shown in Figure 1, if we choose the sample in Figure 1b, node A would be correctly classified as red, due to the fact that A is connected to a blue seed node, and the sample (B-C) reflects perfect heterophily. However, if we choose the sample in Figure 1c, node A would be classified as blue, because it is connected to a blue seed node and the sample (C-E, B-D) reflects perfect homophily. A different sample is shown in Figure 1d, in this case nodes A and B are selected as seed nodes, and regardless of the learned model parameters (i.e., probability of connecting blue-blue, blue-red, red-blue, red-red), notice that node F is not connected to any seed node. Thus, the inferred attribute of node F will depend on the inferred attribute of node E, which in turn also depends on the estimates of unlabeled nodes, C and D. If those estimates are wrong the inference for node F will probably also be wrong.

Notice the importance of the sampling method. The selected nodes should not only reflect the global properties of the network such as balance and homophily but should also be as close as possible to the unlabeled nodes to avoid long label propagation chains that may potentially be erroneous. Next, we describe ten different sampling methods that we evaluate in this work.

Random Nodes. This is the most basic sampling method where a random fraction p of nodes is selected. The sampled network then contains the selected nodes and all edges among them.

Random Edges. This technique randomly selects edges from the set of all edges E. In order to make a fair comparison among other sampling techniques (based on number of nodes), we select edges randomly until we reach a specific fraction p of nodes. That is why this sampling method is referred to as nedges.

SnowBall. Snowball sampling [2, 6] randomly selects a starting node and all its neighbours as well as their neighbours' neighbours (similar to breadth-search-first). The algorithm continues until it has gathered a fraction p of nodes.

Degree. We rank all nodes by their degree in descendant (degreeDESC) and ascendant (degreeASC) order. The idea is to verify whether high (or low) degree nodes are good seeds for classification. Therefore, the fraction p of selected nodes includes the top $p \times 100\%$ of nodes in the ranking. We also provide a mix of degrees (degreeMIX) by selecting $\frac{p}{2} \times 100\%$ of both top high and top low degree nodes.

PageRank. Similar to sampling by degree, we rank nodes by their PageRank (PR) [11] in descendant (pagerankDESC) and ascendant (pagerankASC) order. By using highest PR first, (pagerankDESC) we test whether the most important nodes in the network are good samples for the learning and testing inference. We expect pagerankASC to work poorly since their top low PR nodes are not well connected, and often have low degree.

Optimal Percolation. The motivation behind optimal percolation [10], is to find a minimal set of nodes, called influencers, which, if activated, would cause the spread of information through the whole network. Therefore, we rank nodes based on their *collective influence* in descendant (percolationDESC) and ascendant (percolationASC) order. By taking into account collective influence effects, strategic influencers are identified, also called weak-nodes, which outrank the hubs in the network.

3 EXPERIMENTS

The classification process can be summarized into three steps. First, it splits the nodes of the network into training and testing sets, then it learns the local and relational models using the subgraph extracted from the training sample, and finally it runs the classification on the testing set. The selection of nodes in the training sample varies depending on the sampling method. However, to compare sampling methods, the size of the training sample is kept constant and contains between $5 - 90\%$ of nodes from V.

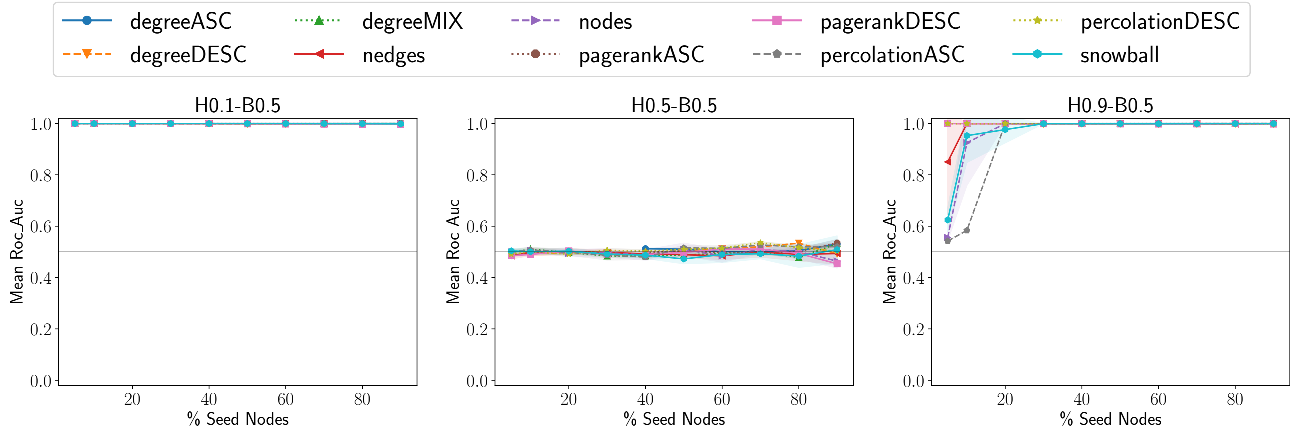
3.1 Synthetic Networks

Datasets. We generate 11 networks with given class balance $B = 0.5$, homophily $H \in \{0.0, 0.1, \ldots, 1.0\}$ and starting degree $m = 4$, using the preferential attachment-based algorithm proposed by Karimi et al. in [7]. Every network consists of $N = 2000$ nodes, $|E| = 7984$ edges, average degree $\langle k \rangle = 8$, and link density $d = 0.0039$[1]. Table 1 shows more network properties for some heterophilic, neutral and homophilic networks. In general, each node is assigned one binary attribute, i.e., $color \in \{blue, red\}$, which defines its class membership. The probability of node v_j to connect to node v_i is given by:

$$\Pi_i = \frac{h_{ij} k_i}{\sum_l h_{lj} k_l} \tag{1}$$

where k_i is the degree of node v_i and h_{ij} is the homophily between the two nodes [7]. Considering that homophily is symmetric and complementary, we can assume that the probability of connections between nodes that belong to class blue (h_{bb}) and nodes that belong to class red (h_{rr}) is identical ($h_{bb} = h_{rr} = H$) and is complementary to intra-class link probability $h_{br} = h_{rb} = 1 - H$. We vary

[1] Since link density is very small, we refer to this set of networks as sparse networks.



(a) Heterophilic **(b) Neutral** **(c) Homophilic**

Figure 2: Results on synthetic (sparse) networks ($\langle k \rangle = 8$, $d = 0.0039$). This figure shows the mean ROC-AUC values of classification for 10 sampling methods on (a) heterophilic, (b) neutral and (c) homophilic networks generated using the preferential attachment-based algorithm proposed by Karimi et al. in [7]. Sample size is shown on the x-axis. Values are averages of 5 runs; shaded areas depict standard deviations.

homophily from $H = 0.0$ (completely heterophilic) to $H = 1.0$ (completely homophilic). When $H = 0.0$, only nodes that do not share the same attribute are connected. In contrast, in the complete homophilic case, only nodes that share the same attribute are interlinked. In neutral networks ($H = 0.5$), a node is equally likely to link to nodes with either label. That means, in neutral networks the formation of edges is statistically independent from node attributes.

Results. For simplicity, we report on networks with special cases of homophily, i.e., $H \in \{0.1, 0.5, 0.9\}$. These results are shown in Figure 2. From Figure 2b we see that classification performance across all sampling methods is uniform in neutral networks since the formation of links is independent of the node attributes. Therefore, relational classifiers cannot detect any pattern in the network structure that helps to guess the correct attributes. The comparison between heterophilic (Figure 2a) and homophilic (Figure 2c) networks shows that regardless of the sampling technique and sample size, heterophilic networks are easier to classify (i.e., most ROC-AUC values are 1.0), whereas homophilic networks (in some cases) require larger training samples to achieve perfect classification. For instance, the overall classification performance is worse (ROC-AUC ≈ 0.6) for sampling by nodes, percolationASC, snowball and nedges, if sample sizes are very small (5%). However, once sample sizes increase, ROC-AUC values quickly converge to 1.0.

Table 1: Synthetic (sparse) network properties. This table shows properties of the networks analyzed in this work. These networks contain two balanced groups of nodes (i.e., blue, red). Each numeric column represents a single network with a specific level of homophily.

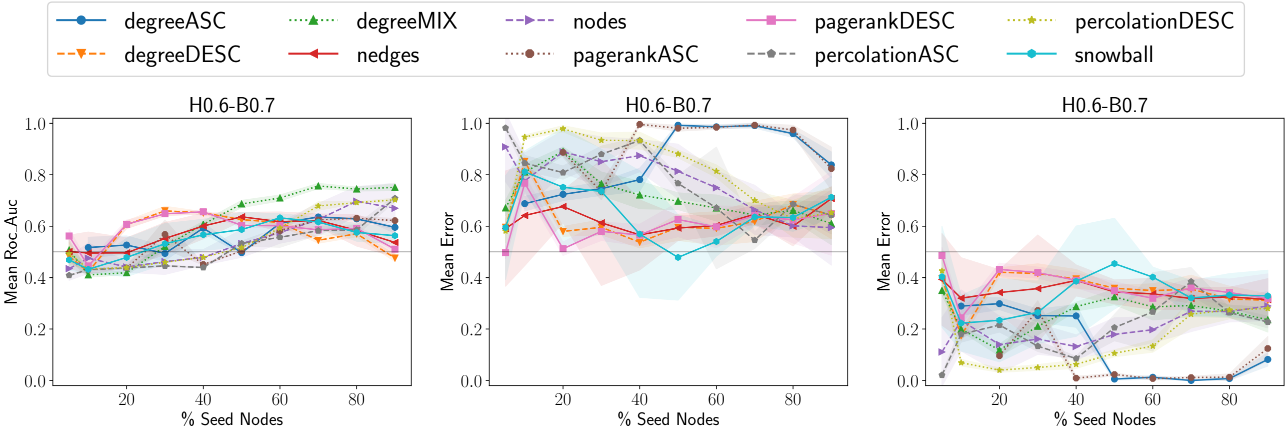
Property	Homophily (H)		
	0.1	0.5	0.9
B	0.5	0.5	0.5
$\langle k \rangle$	8	8	8
Link Density	0.0039	0.0039	0.0039
Node Connectivity	4	4	4
Degree Assortativity	-0.06	-0.06	-0.05
Attribute Assortativity	-0.8	0.01	0.8
Clustering Coefficient	0.01	0.02	0.03

This asymmetry between homophilic and heterophilic (sparse) networks is clear in Figures 6a and 6b, which summarizes the classification performance on samples drawn from balanced networks using different sampling strategies. Color represents different sampling methods, and bars the minimum sample size required so that a classifier achieves an error below 20% for both classes. Homophilic networks require larger sample sizes to achieve good classification performance when sampling by nodes, nedges, snowball, and all metrics that rank nodes in ascendant order (generally, low degree nodes first). Hence, sampling methods that are biased towards high degree nodes (DESC) outperform the other techniques. Heterophilic networks on the other hand are easier to classify, since 8 out of 10 sampling methods achieve good performance for both classes with small training samples that contain only 5-10% of the nodes. Notice that for the particular cases of degreeASC and pagerankASC, all networks require at least 40% of seed nodes to achieve good performance. This occurs because both sampling techniques rank lowest-degree nodes first, and these nodes do not necessary link to each other[2]. Hence, such samples contain disconnected nodes (i.e., $\hat{E} = \emptyset$) that do not help learning the model's parameters.

3.2 Real-World Networks

Dataset. We choose one of the 100 Facebook networks extracted back in 2005 [13]. We focus on the Caltech University network

[2] In fact, degree assortativity[4] for these networks is around -0.06 (i.e., no degree correlation).



(a) ROC-AUC

(b) Mean Error gender 1 (minority)

(c) Mean Error gender 2 (majority)

Figure 3: Results on Caltech Facebook dataset. From left to right, this figure shows the performance of 10 different sampling techniques on the Caltech dataset for different sample sizes: (a) mean ROC-AUC, (b) classification mean error of class gender 1, and (c) classification mean error of class gender 2. Values are averages of 5 runs; shaded areas depict standard deviations. This network is unbalanced $B = 0.7$ towards gender 2, and almost neutral $H = 0.6$. Thus, it is not surprising that classification performance is around $ROC - AUC = 0.5$. From (a) we can see that a small training sample of 30% of highest degree nodes (degreeDESC) can achieve $ROC - AUC = 0.66$. However, the best model is degreeMIX, which achieves $ROC - AUC = 0.76$ with 70% of top mix degree nodes (i.e., 35% top high degree, and 35% top low degree nodes). In general, all sampling methods improve the classification performance with higher sample sizes. From figures (b,c) we can see the class imbalance problem. Since gender 2 is the majority class, it has lower classification error compared to the minority class gender 1.

which includes only intra-school links (i.e., friendship links between user's FB pages). Every node represents a member of the school, and it is described by several attributes: a student/faculty status flag, gender, major, second major/minor, dorm/house, year, and high school. For the purpose of our experiments we choose the attribute $gender \in \{1, 2\}$ as the class label (and only attribute) for the classifier. After removing nodes without gender information (i.e., $gender = 0$), and nodes with no edges we end up with 701 nodes and 15464 edges. The final network is almost neutral ($H = 0.6$), and unbalanced ($B = 0.7$) towards gender 2. Properties of this network are shown in Table 2. For instance, we can see that people are highly connected (i.e., 44.12 friendships on average).

Notice that this network differs from the synthetic network examples, regarding not only class imbalance, but also link density, average degree and clustering coefficient.

Although this network goes beyond the scope of our work (i.e., it is an unbalanced network), we include it in this report for two reasons: (i) to highlight the importance of further research on minimizing the class imbalance problem, and (ii) to show whether homophily has an impact on classification regardless of class imbalance.

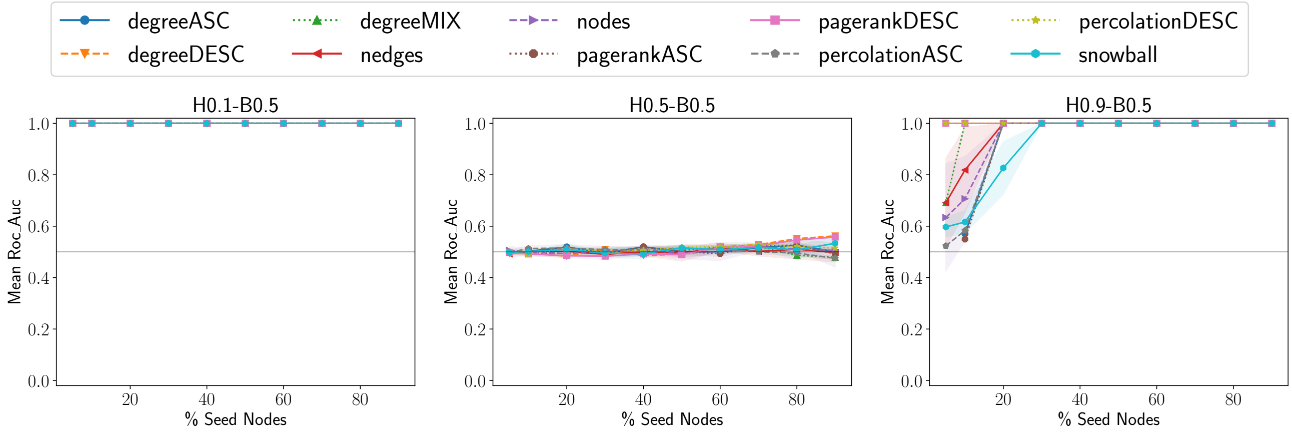
Results. Figure 3 shows the classification results of the Caltech network. Since this network is almost neutral ($H = 0.6$) we expect its performance to be similar to a uniform classifier (i.e., random guessing). Figure 3a confirms this expectation, since it shows that most sample techniques achieve a ROC-AUC value around 0.5, especially for small samples ($5 - 50$% seed nodes). We conclude that degreeMIX outperforms the other sampling methods, although it requires at least 70% of the total number of nodes N to achieve a ROC-AUC=0.76.

ROC-AUC values give us an overall performance of the classification, but they do not give us the whole picture within classes. In Figures 3b and 3c, the classification mean error values for minority and majority are shown respectively. Here we observe the *class imbalance* problem, where classification estimates tend to lean towards the majority class: mean error values for gender 2 (i.e., majority 70%) are lower than for gender 1 (i.e., minority 30%).

These results show the importance of further research on the understanding of relational classification in unbalanced networks with different levels of homophily.

Table 2: Caltech 2005. Properties of the Caltech university Facebook network.

Property	Value	Property	Value
N	701	$\langle k \rangle$	44.12
|E|	15464	$\langle k_{minority} \rangle$	51
gender 1 (min.)	228 (33%)	$\langle k_{majority} \rangle$	41
gender 2 (maj.)	473 (67%)	Node Connectivity	0
B	~0.70	Degree Assortativity	-0.0617
H	0.6	Attribute Assortativity	0.054
Link Density	0.063	Clustering Coefficient	0.39



(a) Heterophilic

(b) Neutral

(c) Homophilic

Figure 4: Results on synthetic (dense) networks ($\langle k \rangle = 40$, $d = 0.019$). Similar to Figure 2, this figure shows the mean ROC-AUC values of classification for 10 sampling methods on (a) heterophilic, (b) neutral and (c) homophilic networks. The difference relies on link density. These networks posses higher density. Sample size is shown on the x-axis. Values are averages of 5 runs; shaded areas depict standard deviations.

4 DISCUSSION

The current work presents a descriptive study of the effect of network sampling on the performance of relational classification. In the following we present a detailed discussion of the factors we explored.

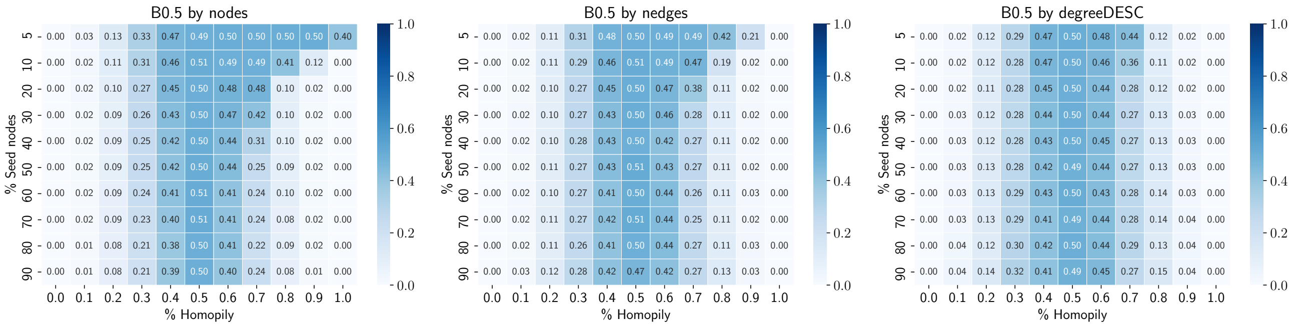
4.1 Network Structure

Link Density. The synthetic networks used in our experiments were generated with $N = 2000$ nodes, $|E| = 7984$ edges, average degree $\langle k \rangle = 8$, and link density $d = 0.0039$. Previous work found that link density impacts performance of relational classifiers [12, 17]. To test this finding, we increased link density to $d = 0.019$, which resulted in $|E_{dense}| = 39600$ edges, and average degree $\langle k_{dense} \rangle = 40$. We refer to this set of networks as dense networks. As we see in Figure 4, higher link density did not improve the classification performance. For neutral (Figure 4b) and heterophilic (Figure 4a) networks, results were similar to the ones using networks with lower link density. Classification of homophilic networks (Figure 4c) on the other hand, worsened ROC-AUC values for very small sample sizes (i.e., 5-20% seed nodes) for degreeMIX, nedges and snowball sampling. Only the performance of classifier based on degreeDESC, pagerankDESC and percolationDESC samples were not affected by increasing link density and outperformed other techniques when only small samples (5% of nodes) were available. Notice that in both variants of link density, ROC-AUC values converged to 1.0 for all sampling strategies with at least 30% of seed nodes. From Figure 6 we see that high link density helps sampling techniques that rank low degree nodes first.

Class Imbalance. Real-world networks can be highly unbalanced, with dissimilar proportions of nodes of each type. For example, the network we studied in Section 3.2 has homophily $H = 0.6$, and class balance $B = 0.7$. We showed that class balance affects classification results. Due to the almost neutral homophily, the collective inference performed only slightly better than random baseline (i.e., random guessing), regardless of the sampling technique. However, due to class imbalance the error for each class (e.g., minority vs. majority) differed drastically. For instance, using random node sampling and 30% seed nodes, class gender 1 achieved a classification error of 0.85, whereas class gender 2 only 0.16. Further research is needed to understand dynamics of relational classification in unbalanced networks.

Homophily. Our work shows that homophily clearly impacts the performance of relational classifiers (cf. Figure 2). When networks are balanced, ROC-AUC curves vary depending on the level of homophily and sample size. As expected, in neutral networks (homophily $H = 0.5$) all sampling methods perform equally well and sample size does not impact classification performance. Thus, it is not surprising that the classifier cannot learn any pattern, since no pattern exists, no matter the size of the sample. However, in heterophilic ($H = 0.1$) or homophilic ($H = 0.9$) networks, the classification accuracy varies based on the sampling strategy and number of labeled seeds present in the training sample. To understand this, let us focus on the performance of three different sampling methods over networks with different levels of homophily shown in Figure 5. Each heatmap shows the classification mean error (averaged over 5 runs) for each of the 11 synthetic networks (x-axis) described in Section 3.1, and the amount of seed nodes in the training sample (y-axis). Darker cells represent higher errors, i.e., worse performance.

Overall, we can see that in the heterophilic regime (H ≤ 0.2, leftmost columns), the classifier works very well with a small fraction of seed nodes. For homophilic networks (H ≥ 0.8, rightmost columns), the classification error is high when the training sample is small for node and nedge sampling. DegreeDESC on the other hand, performs best with only 5% of seed nodes. This seems counterintuitive, as one would expect perfect classification in both cases,



(a) nodes Sampling **(b) nedges Sampling** **(c) degreeDESC Sampling**

Figure 5: Overall mean error of synthetic (sparse) networks. These heatmaps illustrate the overall classification mean error using (a) random node sampling, (b) random edge sampling, and (c) degree descendant sampling. Columns represent networks of different homophily, from heterophilic ($H = 0.0$) to homophilic ($H = 1.0$). Every row shows the percentage of nodes collected in the sampling. Neutral and almost neutral networks ($0.4 \leq H \leq 0.6$) perform uniformly using either sampling technique. The more heterophilic/homophilic the network the more accurate the classification in all cases. However, homophilic networks require larger samples compared to heterophilic networks, especially when using nodes and nedges sampling. In general, in both regimes classification error decreases as the size of training samples increases. Values are averages of 5 runs.

since the strong homophily and strong heterophily should help the classifier learn the relationships between links and attributes. Moreover, global properties of both networks seem to be almost identical, as shown in Table 1. At first glance, it seems that heterophilic networks are easier to classify and their performance do not vary across sampling techniques. Furthermore, high degree seed nodes help the classifier to not only learn the correct parameters, but also to propagate the correct inference among nodes in both heterophilic and homophilic networks. Notice that in both regimes classification error reduces with larger training samples.

4.2 Sampling Networks

As described in Section 2.3, we used ten different network sampling strategies. In Figure 6 we rank sampling methods based on the sample size that is required to train a classifier that achieves a classification error below 20% for both classes[3]. Figures 6a and 6b refer to the sparse networks shown in Section 3.1 (link density $d = 0.0039$). Figures 6c and 6d refer to the respective dense versions (link density $d = 0.019$).

The rightmost figures show the results for classifiers trained on homophilic networks. One can see that sampling strategies that include nodes which have a central position in the network (degreeDESC, pagerankDESC, percolationDESC) work best, and require only 5% of seed nodes. However, these sampling strategies require knowledge of the full network and may not be appropriate in cases where this information is difficult to obtain. In those cases, the second best option for sparse networks (cf. Figure 6b) is to sample 10% of nodes by nedges, which randomly selects edges from the network. Notice that this leads to hubs (i.e., high degree nodes) being preferred. For dense networks (cf. Figure 6d), the second best option is to sample 20% of random nodes.

The rest of sampling techniques require larger training samples. Next, we explain their poor performance. Snowball sampling is a technique that randomly picks a node, then its neighbours, and the neighbours' neighbours, similar to breadth-first search. It is expected to require a larger sample of seed nodes, since in a homophilic network, a snowball sample that starts with a red node, will then select its neighbours—who are likely to be mostly red—without capturing enough blue nodes. Degree, pagerank and percolation ascendant sampling methods (degreeASC, pagerankASC, percolationASC) are usually ranked the worst. This is not surprising, since low degree and low pagerank nodes tend to connect to fewer nodes, leaving a vast majority of nodes disconnected from them (especially in sparse networks).

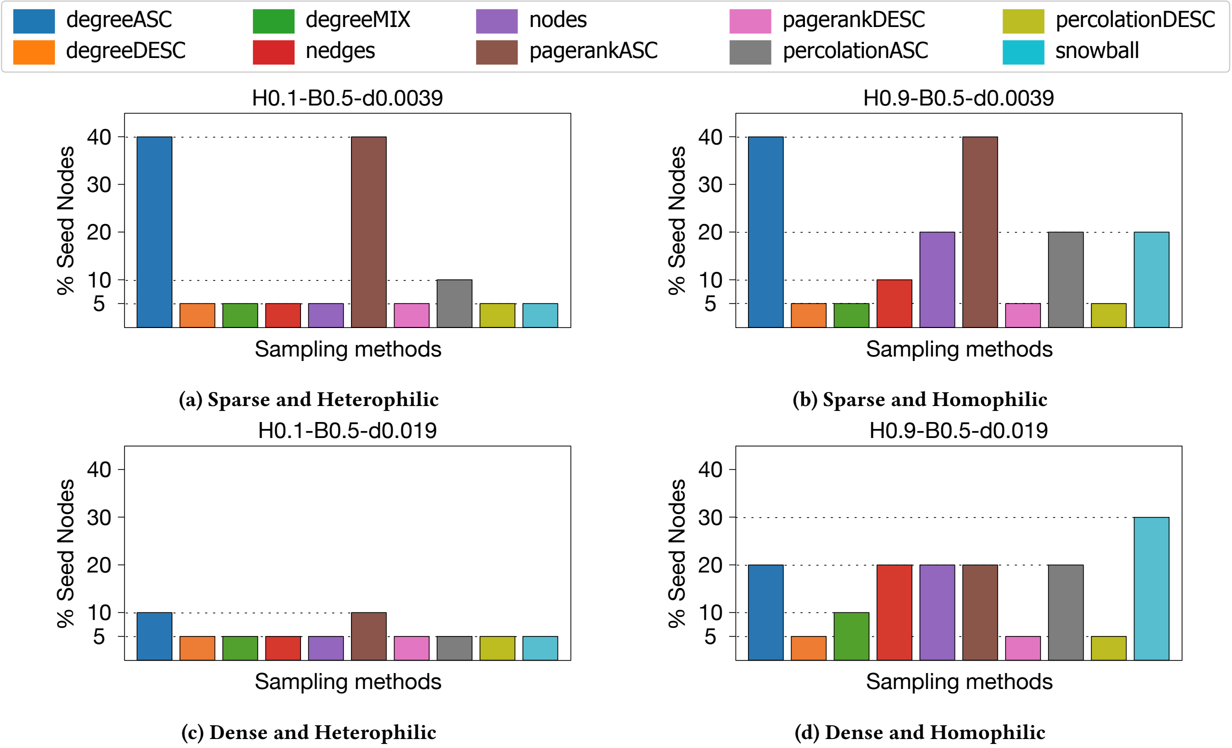
In the heterophilic regime (leftmost figures), degreeASC and pagerankASC are also ranked last. This is also due to the fact that low degree nodes connect to only a few nodes. However, eight out of ten sampling techniques require a training sample containing only 5-10% of seed nodes. In other words, in a heterophilic network, it is enough to collect only 5% of random nodes in order to achieve good classification performance[4].

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented a step towards quantifying sampling bias in network inference. Precisely, we studied the influence of network structure and sampling on relational classification. Our findings are as follows. (i) Heterophilic networks are easier to classify than homophilic networks. The former require only 5-10% of seed nodes to be able to classify correctly most unlabeled nodes. This holds for 8 out of 10 sampling methods. (ii) Link density and degree assortativity influence the performance of sampling methods that rank low-degree nodes first. Low-degree nodes are less likely to connect to each other if the network has low link density

[3] While this measure might be arbitrary, the goal is to show an unbiased classification, where both classes get correctly inferred

[4] This holds for $0.0 \leq H \leq 0.2$ as shown in Figure 5a.



(a) Sparse and Heterophilic

(b) Sparse and Homophilic

(c) Dense and Heterophilic

(d) Dense and Homophilic

Figure 6: Ranking of Sampling methods. This figure depicts the minimum sample size (in terms of percentage of nodes) required for all sampling techniques in order to achieve a classification error below 20% for both classes (blue and red) in balanced scale-free networks. We summarize from left to right, (a,c) heterophilic and (b,d) homophilic networks, and from top to bottom, (a,b) sparse and (c,d) dense networks. Each color represents a different sampling method. The lower the bars the better. For instance, sampling by degreeDESC (second bar from left to right in each plot) works very well by picking the top 5% of highest degree nodes in all cases. In general, classification works best for heterophilic networks since they require very small training samples regardless of sampling strategy and density.

and negative or neutral degree assortativity. Therefore, sampling methods that rank low degree nodes first require larger samples to include not only more nodes, but also more edges (within seeds, and from seeds to unlabeled nodes). (iii) Link density also influences the performance of homophilic networks. The higher the link density, the larger the training samples for degreeMIX, nedges and snowball sampling in order to achieve good accuracy for all classes. (iv) High degree nodes are the best seed nodes, since only 5% of them achieve optimal classification performance (ROC-AUC=1.0) for homophilic, heterophilic, dense and sparse networks. However, these sampling strategies require knowledge of the full network and may not be appropriate in cases where this information is difficult to obtain. Therefore, in those cases it is sufficient to sample: 5% of random nodes in heterophilic networks, 10% of nedges if the network is homophilic and has low link density, and 20% nodes if the network is homophilic and has high link density.

How to select a sample that reflects the global properties of the original network and allows accurate label propagation is still an open research question. In future work, we plan to investigate the trade-off between constructing well-connected samples that help the classifier to learn patterns between link formation, attributes, and seed nodes that are as distant as possible to gain information about different parts of the original network. We also plan to investigate: (i) networks with multiple attributes where attribute distributions are skewed (i.e., imbalanced classes exist), (ii) more sophisticated sampling techniques such as the one in [3], and (iii) other relational models such as relational logistic regression[16].

Acknowledgements. The authors would like to thank Dr. Peter Fennell for his time and helpful advice on various technical topics, and Prof. Dr. Markus Strohmaier, Dr. Mathieu Génois and Reinhard Munz for their valuable comments and suggestions to improve the quality of the paper. This work was supported in part by the ARO (contract W911NF-16-1-0306) and by the AFOSR (contract FA9550-17-1-0327).

REFERENCES

[1] Nesreen K Ahmed, Jennifer Neville, and Ramana Rao Kompella. 2012. Network Sampling Designs for Relational Classification. In *ICWSM.*

[2] Rowland Atkinson and John Flint. 2001. Accessing hidden and hard-to-reach populations: Snowball research strategies. *Social research update* 33, 1 (2001), 1–4.

[3] Konstantin Avrachenkov, Bruno Ribeiro, and Jithin K Sreedharan. 2016. Inference in OSNs via Lightweight Partial Crawls. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer*

Science. ACM, 165–177.

[4] Albert-László Barabási. 2016. *Network science.* Cambridge university press.

[5] Lise Getoor and Ben Taskar. 2007. *Introduction to statistical relational learning.* MIT press.

[6] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

[7] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. 2017. Visibility of minorities in social networks. *arXiv:1702.00150* (2017).

[8] Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 631–636.

[9] Sofus A. Macskassy and Foster Provost. 2007. Classification in Networked Data: A Toolkit and a Univariate Case Study. *J. Mach. Learn. Res.* 8 (May 2007), 935–983. http://dl.acm.org/citation.cfm?id=1248659.1248693

[10] Flaviano Morone and Hernán A Makse. 2015. Influence maximization in complex networks through optimal percolation. *Nature* 524, 7563 (2015), 65.

[11] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[12] Prithviraj Sen, Galileo Mark Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. 2008. Collective Classification in Network Data. *AI Magazine* 29, 3 (2008), 93–106. http://www.cs.iit.edu/~ml/pdfs/sen-aimag08.pdf

[13] Amanda L Traud, Peter J Mucha, and Mason A Porter. 2012. Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications* 391, 16 (2012), 4165–4180.

[14] Claudia Wagner, Philipp Singer, Fariba Karimi, Jürgen Pfeffer, and Markus Strohmaier. 2017. Sampling from Social Networks with Attributes. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1181–1190. https://doi.org/10.1145/3038912.3052665

[15] Jiasen Yang, Bruno Ribeiro, and Jennifer Neville. 2017. Should We Be Confident in Peer Effects Estimated From Social Network Crawls?. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.* 708–711. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15696

[16] Jiasen Yang, Bruno Ribeiro, and Jennifer Neville. 2017. Stochastic Gradient Descent for Relational Logistic Regression via Partial Network Crawls. *arXiv preprint arXiv:1707.07716* (2017).

[17] Giselle Zeno and Jennifer Neville. 2016. Investigating the impact of graph structure and attribute correlation on collective classification performance. (2016).