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

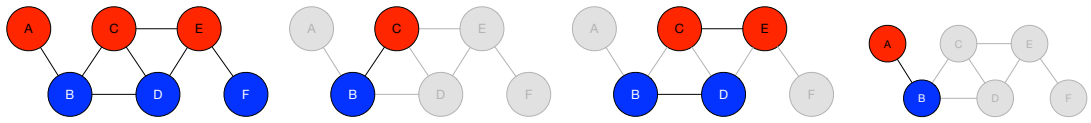
Abstract

Relational Classification together with Collective Inference are used to infer missing attributes of nodes using information from their neighbors. First, the model parameters are learned from a sample of nodes and edges and then that knowledge is applied to the rest of unseen nodes. 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? 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. 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 which require larger training samples to obtain good accuracy. This finding suggests that the impact of network structure on relational classification is more complex than previously thought.

Research Problem. 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 machine learning methods to infer information about the network from a partial sample. 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 [1], we address some of these questions by systematically studying potential sources of bias in the relational inference process. New insights on how sampling impacts relational classification performance can potentially lead to new unbiased strategies.

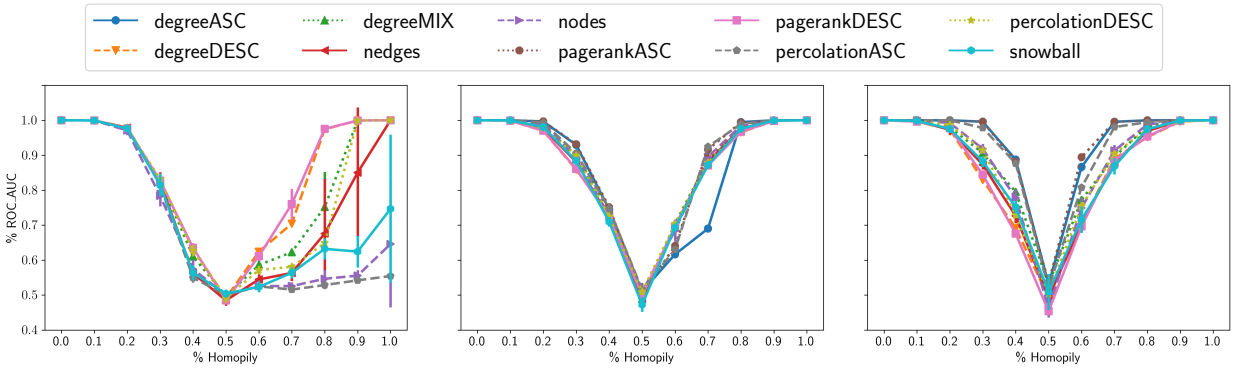
Motivation. The goal of sampling is to split the network into *training* and *testing* samples. First, a subgraph is extracted from the original network to learn the model parameters. Nodes that belong to the training sample are called *seed nodes*, and we assume that their edges and attributes are known by the classification algorithm. For example, based on the information shown in Fig. 1, if we choose the sample in Fig. 1b, node A would be correctly classified as red, since A is connected to a blue seed node, and the sample (B-C) reflects perfect heterophily. However, if we choose the sample in Fig. 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 Fig. 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 label propagation chains that may potentially be erroneous.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.



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

Figure 1: **Example**. (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.



(a) 5% seed nodes (b) 50% seed nodes (c) 90% seed nodes

Figure 2: **Results on synthetic networks**. This figure shows the mean ROC-AUC values (y-axis) of classification using three different sample sizes: (a) 5%, (b) 10% and (c) 90% seed nodes. Values are averages of 5 runs; error bars depict standard deviations.

Experiments. We generate 11 networks with given class balance $B = 0.5$ (i.e., 50% red and 50% blue nodes), 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 [2]. Every network consists of $N = 2000$ nodes, $|E| = 7984$ edges, and average degree $\langle k \rangle = 8$. We run the relational classification algorithm proposed by Macskassy and Provost in [3] 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. Results are shown in Fig. 2. We see that classification performance across all sampling methods is uniform in neutral networks (H=0.5) since the formation of links is independent of the node attributes. The comparison between heterophilic ($H < 0.3$) and homophilic ($H > 0.7$) 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. These differences are due to the fact that probabilities within blue and red nodes are identical ($h_{bb} = h_{rr}$). This means that the homophily of the network is the sum of homophily within groups $H = h_{bb} + h_{rr}$. Thus, fluctuations can be higher within groups when samples are small and networks homophilic (e.g., $h_{bb} = 0.1 \; ! = h_{rr} = 0.7$). This leads to a biased classification towards the homophilic group with higher likelihood.

Findings and Contributions. In this work [1], 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 two-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.

References

[1] Lisette Espín-Noboa, Claudia Wagner, Fariba Karimi, and Kristina Lerman. Towards quantifying sampling bias in network inference. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 1277–1285. International World Wide Web Conferences Steering Committee, 2018.

[2] Fariba Karimi, Mathieu Génois, Claudia Wagner, Philipp Singer, and Markus Strohmaier. Homophily influences ranking of minorities in social networks. *Scientific Reports*, 8(1):11077, 2018.

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