Quantifying Biases in Attribute Inference on Social Networks

Lisette Espín-Noboa [1 2] **Fariba Karimi** [1 2] **Bruno Ribeiro** [3] **Kristina Lerman** [4] **Claudia Wagner** [1 2]

Abstract

Scientists working with observational data often need to know attributes of nodes in social networks. When these attributes are partly missing, collective classification can be used to infer them. However, factors affecting its performance are not yet well understood, nor the conditions that bias the performance against minorities. To this end, we systematically study how structural properties of the *network* and the *training sample* influence the performance of collective classification. Our main finding shows that mean classification performance can empirically and analytically be predicted by structural properties such as homophily, fraction of minorities, and edge density. Our results establish evaluation benchmarks, especially helpful when no ground-truth is available.

1. Introduction

In many scientific fields, such as social science or web science, as well as in industry applications, there is a major need to have access to the attributes of individuals in social networks; for instance, to explore the relationships between socio-demographic attributes of people and their behavior, or to investigate segregation and information diffusion across different groups of people. In practice, however, often only partial information about individuals is available due to API quotas or privacy settings. In this scenario, collective classification (Neville & Jensen, 2000; Getoor & Taskar, 2007; Macskassy & Provost, 2007) can be used to infer individual's attributes using information from their neighbors and a few *seeds* (i.e., individuals with known attributes). The advantage of collective classification over traditional machine learning techniques—which rely only on node attributes and ignore relationships with other nodes—is that the former does not require the data to be independent and identically distributed, which is important when dealing with networked data, as the class label of a node may depend on the class label of its neighbors.

A challenge for inference is that the distribution of individual attributes over the network is often uneven, with coexisting groups of different sizes: for example, one ethnic group or gender may dominate the other group. Machine learning methods often struggle with unbalanced data, and as a result, may misclassify the minority class more often than the majority class. Many social networks also demonstrate a property known as homophily, which is the tendency of individuals to associate with others who are similar to them (e.g., with respect to gender) (McPherson et al., 2001).

Despite its importance, little is known about the impact of network structure—in particular *homophily* and the *fraction of minorities*—on the performance of collective classification. The variety of network types—as well as many choices for the graph sampling method, relational model, and collective inference—make it difficult to choose the best combination of methods for a particular problem. A further complication is that ground truth data is not always available to evaluate results.

Research Questions. In this work we systematically compare different factors that may influence the performance of collective classification. These factors relate to structural properties of the *network* and the *training sample* involved in the inference process.

- **RQ1:** How does *network structure* (i.e., *number of nodes*, *edge density*, *fraction of minorities*, *homophily*) affect performance of collective classification?
- **RQ2:** How does the choice of the *sampling technique* affect the performance of collective classification and its parameter estimation?
- **RQ3:** How does network structure and the choice of the sampling technique bias inference against the minority or majority groups?

2. Related Work

(Macskassy & Provost, 2007) evaluated the influence of relational classifiers (**RC**) together with collective inference algorithms (**CI**) and sample size using random node sampling,

[1]GESIS [2]University of Koblenz-Landau [3]Purdue University [4]USC-ISI. Correspondence to: Lisette Espín-Noboa <Lisette.Espin@gesis.org>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 108, 2020. Copyright 2020 by the author(s).

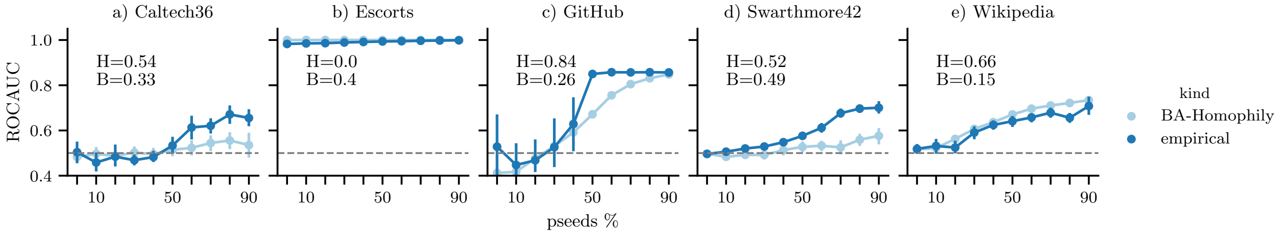

Figure 1. **Classification performance on empirical networks:** (a,d) university networks, (b) sexual contact network, (c) mutual-follower developer network, and (e) reciprocal hypher-link network. Properties of the networks are shown as H (homophily) and B (fraction of minorities). Sample size using random node sampling is shown on the x-axis, and the mean ROCAUC score on the y-axis. Results from real networks are shown as "empirical" (dark blue), and their synthetic counterparts as "BA-Homophily" (light blue). Overall, results from synthetic networks follow a similar pattern as the empirical counterpart.

and concluded that the network-only Bayes classifier (**nBC**) is almost always significantly and often substantially worse than other RCs. When samples are small, relaxation labeling (**RL**) is best among all CIs, whereas weighted-voting (**wvRN**) and class-distribution (**cdRN**) are best among all RCs. When samples are large, all CIs perform similarly well, and network-only link-based (**nLB**) is best among all RCs. More recent work by (Zeno & Neville, 2016) concluded that as the sample size increases it is better to learn a model using nBC than with wvRN. While all these contributions touch upon important points, they have mostly focused on the performance of RCs and CIs. Besides, their findings are not comparable since they use different datasets, different configurations of RC and CI, and different evaluation metrics. In our work we focus on the performance and fairness of nBC and RL by systematically varying some properties of the network and the training sample.

3. Approach and Methods

We utilize *BA-Homophily*, a simple model that allows to generate scale-free undirected networks with tunable homophily and group size (Karimi et al., 2018). One advantage of this model is that it generates networks with power-law degree distributions which have been observed in many large-scale social networks (Barabási, 2009). More importantly, it only requires two main input parameters (homophily and the fraction of minorities), and thus the behavior of the model is analytically tractable (Karimi et al., 2018). The homophily parameter ranges from 0 to 1, and it allows us to generate heterophilic networks ($0 \leq H < 0.5$), neutral networks ($H = 0.5$), and homophilic networks ($0.5 < H \leq 1$).

Furthermore, we follow definitions and pseudo-codes from (Macskassy & Provost, 2007) to implement the *network-only Bayes* classifier (nBC) and the *relaxation labeling* inference algorithm (RL). We measure classification performance in terms of ROCAUC[1], assess the quality of the model parameters using squared estimation errors (SE), and compute the overall accuracy equality score (Berk et al., 2018) to asses the fairness of the classifier with respect to minority and majority classes.

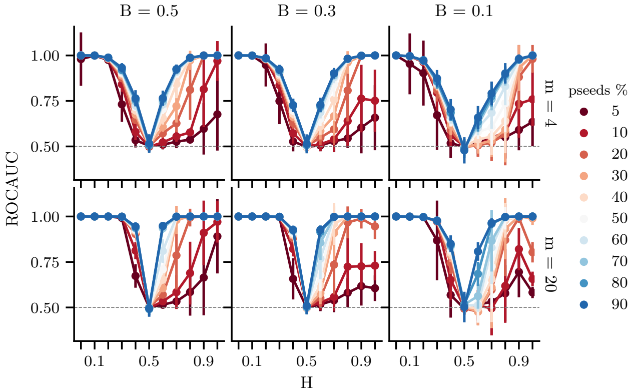

Figure 2. **Network structure vs. classification performance**. Results on "BA-Homophily" networks with $N = 2000$ nodes and different levels of: homophily (x-axis), minimum degree (rows), fraction of minorities (columns), and sample size (colors). Dots represent mean ROCAUC scores (y-axis) over 50 runs, and error bars their respective standard deviation. We see that (i) heterophilic networks $H < 0.5$ achieve higher ROCAUC than homophilic networks $H > 0.5$, especially when samples are small (red), and (ii) the denser the network $m = 20$, the higher the ROCAUC.

Contributions. We demonstrate analytically and empirically that classification performance, estimation error, and fairness are predictable and mainly depend on homophily, fraction of minorities, and sample size, see Figures 1 and 2. In particular, we show that: (i) small training samples are enough for heterophilic networks to achieve high and fair classification performance, even with imperfect estimates, (ii) when sampling budgets are small, partial crawls (Avrachenkov et al., 2016) and edge sampling achieve the most accurate model estimates, and (iii) homophilic networks are more prone to fairness issues and low performance, especially when samples are small and the fraction of minorities decreases. Last but not least, we make our code and data openly available (Espin-Noboa, 2019).

[1] Area under the receiver operating characteristic curve: It measures how well the classifier can distinguish between classes.

References

Avrachenkov, K., Ribeiro, B., and Sreedharan, J. K. Inference in osns via lightweight partial crawls. In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, pp. 165–177. ACM, 2016.

Barabási, A.-L. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 0049124118782533, 2018.

Espin-Noboa, L. Discrimination-in-relational-classification. https://github.com/gesiscss/Discrimination-in-Relational-Classification, 2019.

Getoor, L. and Taskar, B. *Introduction to statistical relational learning*. MIT press, 2007.

Karimi, F., Génois, M., Wagner, C., Singer, P., and Strohmaier, M. Homophily influences ranking of minorities in social networks. *Scientific reports*, 8, 2018.

Macskassy, S. A. and Provost, F. Classification in networked data: A toolkit and a univariate case study. *J. Mach. Learn. Res.*, 8:935–983, May 2007. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1248659.1248693.

McPherson, M., Smith-Lovin, L., and Cook, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001. doi: 10.1146/annurev.soc.27.1.415. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.soc.27.1.415.

Neville, J. and Jensen, D. Iterative classification in relational data. In *Proc. AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, pp. 13–20, 2000.

Zeno, G. and Neville, J. Investigating the impact of graph structure and attribute correlation on collective classification performance. 2016.