Explaining classification performance and bias via network structure and sampling technique

Espín-Noboa, Lisette; Karimi, Fariba; Ribeiro, Bruno; Lerman, Kristina; Wagner, Claudia

Veröffentlichungsversion / Published Version
Konferenzbeitrag / conference paper

Espín-Noboa *et al. Appl Netw Sci* (2021) 6:78
https://doi.org/10.1007/s41109-021-00394-3

Applied Network Science

RESEARCH Open Access

Explaining classification performance and bias via network structure and sampling technique

Lisette Espín-Noboa[1,2*], Fariba Karimi[1], Bruno Ribeiro[3], Kristina Lerman[4,5] and Claudia Wagner[1,6]

*Correspondence: Lisette.Espin@gesis.org
[2] Computer Science, University of Koblenz-Landau, Koblenz, Germany
Full list of author information is available at the end of the article

Abstract

Social networks are very important carriers of information. For instance, the political leaning of our friends can serve as a proxy to identify our own political preferences. This explanatory power is leveraged in many scenarios ranging from business decision-making to scientific research to infer missing attributes using machine learning. However, factors affecting the performance and the direction of bias of these algorithms are not well understood. To this end, we systematically study how structural properties of the *network* and the *training sample* influence the results of collective classification. Our main findings show that (i) mean classification performance can empirically and analytically be predicted by structural properties such as homophily, class balance, edge density and sample size, (ii) small training samples are enough for heterophilic networks to achieve high and unbiased classification performance, even with imperfect model estimates, (iii) homophilic networks are more prone to bias issues and low performance when group size differences increase, (iv) when sampling budgets are small, partial crawls achieve the most accurate model estimates, and degree sampling achieves the highest overall performance. Our findings help practitioners to better understand and evaluate their results when sampling budgets are small or when no ground-truth is available.

Keywords: Relational classification, Collective inference, Social networks, Network structure, Input bias, Sampling bias, Output bias

Introduction

People connect with others through online social network platforms (Hughes et al. 2012), scientific collaboration networks (Newman 2001), and other peer-to-peer platforms (Larrimore et al. 2011). All these connections are leveraged by certain systems to recommend new content and new connections. In turn, these recommendations are often based on algorithms that rely on individuals' information such as gender, political leaning or credit score. In practice, however, often only partial information about individuals is available due to API quotas (e.g., very large networks). In this scenario, collective classification[1] Neville and Jensen (2000); Getoor and Taskar (2000); Macskassy and Provost (2007) can be used to infer individual's attributes using information from

[1] A technique that combines relational classification (classifier) and collective inference (inference algorithm).

their neighbors and a few *seeds* (i.e., individuals with known attributes). The advantage of collective classification over traditional machine learning techniques[2] 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 in the network.

However, little is known about the impact of network structure on the performance and the direction of bias of collective classification. For instance, many social networks 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 or ethnicity (McPherson et al. 2001). Furthermore, the *class balance* or distribution of individual attributes over the network is often uneven, with coexisting groups of different sizes, e.g., one ethnic group may dominate the other group. A challenge for inference is then to be accurate and unbiased with each individual and group in the network, regardless of its structure. However, the variety of network types—as well as many choices for the sampling method, modeling, and 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.

Therefore, it becomes crucial to understand how these algorithms work and under which conditions they discriminate against certain groups of people (e.g., minorities). To that end, our work aims at providing decision makers with: (i) evaluation guidelines to assess the impact of different network types and sampling techniques on collective classification, and (ii) a reproducible and reusable tool to identify performance and bias issues on new networks, sampling techniques, classifiers and inference algorithms. Our findings also shed light on the design of better algorithms to mitigate biases coming from networked data.

Research questions In this work we systematically study different factors that may influence the performance and bias of collective classification. These factors relate to structural properties of the *network* and the *training sample* (i.e., random sampled subgraph with labeled nodes, so-called seed nodes) involved in the beginning of the inference process (see Fig. 1).

- **RQ1:** How does *network structure* (i.e., *homophily*, *class balance*, and *edge density*) affect the overall performance of collective classification?
- **RQ2:** How does the choice of the *sampling technique* affect the overall performance of collective classification and its parameter estimation?
- **RQ3:** How does network structure and the choice of sampling technique influence the direction of bias in collective classification?

Approach and methods We utilize a network model that allows us to generate scale-free networks with tunable homophily and class balance (Karimi et al. 2018). One advantage of this model is that it generates node-attributed 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

[2] which rely only on node attributes and ignore relationships with other nodes

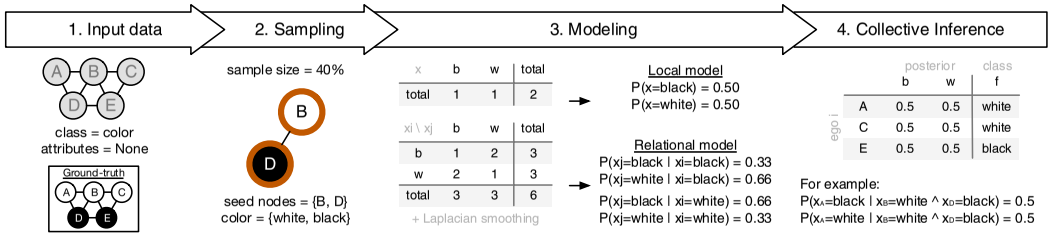

Fig. 1 Classification on networks. This example describes the collective classification workflow. (1) A network $G = (V, E, C)$ is given, and it is defined by a set of nodes V, edges E, and class labels C. Nodes are known to belong to a binary class $color \in \{white, black\}$, and no additional attributes are given. Therefore, the goal of the collective classification is to infer the correct class label of nodes. To achieve this (2) a set of seed nodes is sampled, together with their class labels, to create a subgraph $G_{seeds} \subset G$. Then, (3) the local model learns the probability of each class label, e.g., $P(x = black) = 0.5$, while the relational model learns the probability of neighbor v_j having a particular class label conditioned on the class label of a node v_i, e.g., $P(x_j = white|x_i = black) = 0.66$. Finally, in (4) the collective inference process assigns posterior probabilities to each unlabeled node using their 1-hop neighbors. For example, the probability that node A is black (or white) is conditioned on the color of its neighbors B and D. Notice that the inference is performed collectively – i.e., at the same time for all nodes

and class balance), 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 networks with a broad range of group mixing ranging from heterophilic networks ($0 \leq h < 0.5$) to neutral networks ($h = 0.5$), and homophilic networks ($0.5 < h \leq 1$).

Furthermore, we follow definitions and pseudo-codes from Macskassy and 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, assess the quality of the model parameters using squared estimation errors (SE), and extend the balanced accuracy (Brodersen et al. 2010) to compare the *true positive rates* (TPR) of each class—also known as sensitivity and specificity in binary classification—to assess the direction of bias.

Contributions Our contributions are two-fold: First, we propose a methodology to assess classification performance and bias on real-world networks when no ground-truth is available. Second, we demonstrate analytically and empirically that collective classification mean performance, estimation error, and bias are predictable and mainly depend on homophily, class balance, edge density, and sample size. In particular, we show that: (i) Larger training samples are required in undirected networks with homophilic connections to achieve a similar high classification performance compared to heterophilic networks. (ii) Samples obtained by partial crawls allow to learn the most accurate model estimates, and the most accurate results are obtained when using degree sampling. (iii) Classification results are often less biased in heterophilic networks than in homophilic networks regardless of class imbalance. Last but not least, we make our code and data openly available (Espín-Noboa 2021).

Related work

We cover previous work on analyzing the performance of collective classification from both *algorithmic* and *network* bias perspectives. Additionally, we review literature on *network sampling* to identify characteristics of *seed nodes* (i.e., labeled nodes) that could potentially improve the performance of collective classification.

Collective classification performance From the literature (Macskassy and Provost 2007; Sen et al. 2008; Zeno and Neville 2016), we know that each component in the collective classification pipeline can be implemented in different ways, and it has been shown that different implementations (or combinations) may perform better or worse depending on certain properties of the network and sample size. For instance, Macskassy and Provost (2007) evaluated the influence of relational classifiers (**RC**) together with collective inference algorithms (**CI**) and sample size using random node sampling, and conclude 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 the best among all CIs, whereas weighted-voting (**wvRN**) and class-distribution (**cdRN**) are the best among all RCs. When samples are large, all CIs perform similarly well, and network-only link-based (**nLB**) is the best among all RCs. More recent work by Zeno and Neville (2016) concluded that as the sample size increases it is better to learn a model using nBC than with wvRN. Note that this work utilizes synthetic networks and assumes edge weights to be $w_{ij} = 1$, opposite to the work by Macskassy and Provost (2007) where

they run their experiments on real-world networks and thus utilize real weight values $w_{ij} \in \mathbb{R}$. While all these contributions are very important and relevant, 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 of nBC and RL, and systematically vary some properties of the network and the training sample. We choose this combination because it has been shown that RL outperforms other CIs (Macskassy and Provost 2007), and the model parameters of nBC are easy to interpret, i.e., they reflect the network properties of interest. Besides, nBC outperforms wvRN (Zeno and Neville 2016). Moreover, we use ROCAUC as a standard measure of classification performance and compare the true positive rates of each class to assess the direction of bias.

Network bias The influence of homophily and edge density on both RC and CI was studied in Sen et al. (2008) on synthetic networks. They found that, as homophily (or density) increases, the accuracy of classification improves drastically over non-relational classifiers. It has also been shown that certain RC such as wvRN perform poorly on heterophilic networks (Dong et al. 2019). Besides, when networks are neutral (i.e., when nodes are connected at random), no classifier—even with the largest training dataset—can beat a random classifier (Espín-Noboa et al. 2018). We build upon these findings and broaden the spectrum of network types by varying not only homophily and density, but also class balance (i.e., fraction of minorities). Note that a new metric called *monophily* shows that similarity among friends-of-friends (a.k.a., 2-hop neighbors) can improve relational classification results, especially in the neutral case (Altenburger and Ugander 2018). However, we focus on 1-hop neighbors to better understand the role of homophily and class balance in collective classification.

Sampling bias on networks Previous research has studied the robustness of network samples from different angles. For instance, a range of network properties such as degree and betweenness centrality have been found to be sensitive to the choice of sampling methods (Leskovec and Faloutsos 2006; Galaskiewicz 1991; Costenbader and Valente 2003; Huisman 2009; Borgatti et al. 2006; Kossinets 2006; Wang et al. 2012; Lee and Pfeffer 2015; Li and Yeh 2011; Wagner et al. 2016). These efforts have shown that network estimates become more inaccurate with lower sample coverage, but there is a wide variability of these effects across different measures, network structures and sampling errors. In terms of benchmarking network sampling strategies, Coscia and Rossi (2018) show that it is not enough to ask which method returns the most accurate sample (in terms of statistical properties); one also needs to consider API constraints and sampling budgets. In the context of collective classification, Yang et al. (2017) demonstrate that certain sampling techniques such as snowball sampling and random walks could lead to biased parameter estimates, and then correct such bias by exploiting a general crawling method (Avrachenkov et al. 2016) that produces unbiased model estimates. We leverage the nature of these estimates to verify whether perfect estimates always lead to perfect classification performance and unbiased results while varying the sampling budget.

Fairness in classification In recent years, there has been an increase of research focusing on mitigating bias (Raghavan et al. 2020; Dixon et al. 2018; Krasanakiset al. 2018) and guaranteeing individual and group fairness while preserving accuracy in classification algorithms (Dwork et al. 2018; Binns 2020; Kallus et al. 2019; Zafar et al. 2017). Many

definitions of fairness have been proposed (Verma and Rubin 2018; Mehrabi et al. 2019) and the most used are equalized odds (Hardt et al. 2016), equal opportunity (Hardt et al. 2016), counterfactual fairness (Kusner et al. 2017), demographic parity and fairness through awareness (Dwork et al. 2012). While all this body of research focuses on fairness influenced by the attributes of the individuals, recent research proposes a new notion of fairness that is able to capture the relational structure of individuals (Farnadi et al. 2018; Zhang et al. 2020). The main difference between this line of research and our methodology is that instead of ensuring a fair algorithm, we focus on *explaining discrimination* (Mehrabi et al. 2019) via input and sampling bias. By doing so, we gain a better understanding of the direction of bias (i.e., why and when collective classification discriminates against certain groups of people). Consequently, we simplify the classification task to work with only one (protected) binary attribute (e.g., gender) which in turn plays the role of a target and a membership class.

To our best knowledge, there is no systematic study that explores the interplay between sampling, network structure, performance and bias in collective classification, and we aim to fill this gap.

Methods: classification on networks

We focus on *classification on networks* as a *semi*-supervised machine learning technique, where categorical class labels of records are predicted by exploiting both the labeled and the unlabeled part of the data (Marinho et al. 2009). In particular, we study *relational classification* together with *collective inference* (a.k.a., collective classification Sen et al. 2008), two techniques used to infer missing attributes of nodes using information from their neighbors. Figure 1 shows the four requirements of collective classification: (i) Data: a network with unlabeled nodes. (ii) Training sample: a subgraph with known labels sampled from the network. (iii) Models: local and relational models learned from the training sample to encode class priors and conditional probabilities, respectively. (iv) Collective Inference: A systematic process where models are fitted to the ego networks of each unlabeled node to infer their posterior class probabilities.

Next, we describe (i) networks of interest, (ii) network sampling, and (iii) the modeling and inference processes utilized in this work.

Input data: an attributed network

We define the input network as: Let $G = (V, E, C)$ be an attributed unweighted graph with $V = \{v_1, ..., v_n\}$ being a set of N nodes, $E \subseteq V \times V$ a set of undirected edges, and $C = \{c_1, ..., c_n\}$ a list of binary class labels where each element c_i represents the class membership of node v_i.

The homophily parameter H is the probability of nodes with the same class label to be connected. Homophily values range from 0 to 1. 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 minority nodes—with respect to C—in the network. A network is *balanced* when all class labels have the same number of nodes

($B = 0.5$), otherwise it is *unbalanced* ($B < 0.5$). Edge density $d = \frac{2|E|}{N(N-1)}$ represents the fraction of existing edges out of all possible edges in G.

To generate such networks, we refer to the preferential attachment-based model with adjustable homophily proposed in Karimi et al. (2018). In this model, each node is assigned one binary class label, e.g., $color \in \{white, black\}$[3]. The probability of node v_i to connect to node v_j is given by:

$$\Pi_{ij} = \frac{h_{ij} k_j}{\sum_l h_{il} k_l} \tag{1}$$

where k_i is the degree of node v_i and h_{ij} is the homophily between nodes v_i and v_j. For simplicity, in our synthetic networks, we assume that homophily is symmetric and complementary: $h_{aa} = h_{bb} = H$ and $h_{ab} = h_{ba} = 1 - H$.

Example. Figure 1 shows an attributed network (see the ground-truth in "1. Input data"), where nodes are assigned one *color*, either *white* or *black*. Since only 3 out of 7 edges are same-color connections, this network is heterophilic ($H \approx 0.43$). This network is also unbalanced ($B = \frac{2}{5} = 0.4$) because the number of black nodes ($N_b = 2$) is different from the number of white nodes ($N_w = 3$).

Note that in practice, often the list of class labels C is unknown or incomplete. Therefore, values for B and H are either not available or inaccurate. However, in our experiments we assume that the ground-truth is given (see section "Discussion and future work" for a real use case).

Sampling: the observed network

The goal of sampling is to split the network into *training* and *testing* samples. First, a subgraph $G_{\text{seeds}} = (\hat{V}, \hat{E}, \hat{C})$ is extracted from G in order to learn the model parameters (see section "Modeling and collective inference: estimates"). Nodes $\hat{V} \subset V$ that belong to the training sample G_{seeds} are called *seed nodes*, and they are a percentage *pseeds* of nodes selected by the sampling method. Similarly, edges $\hat{E} \subseteq \hat{V} \times \hat{V}$ are all links ($\hat{E} \subset E$) in the induced subgraph between seed nodes $\hat{V}$. Class labels $\hat{C} \subset C$ are automatically known by the classification algorithm after sampling. The testing sample includes all nodes and edges of G, but only nodes $v_i \in V - \hat{V}$ are target for classification. In this paper we explore four widely used sampling methods: random nodes, random edges, degree ranking, and partial crawls.

Random nodes This is the most used and basic sampling method where a percentage *pseeds* of random nodes is selected. The training sample then contains the selected nodes and all edges among them. Note that in the case of unbalanced networks, this sample will be biased towards the majority class.

Random edges This technique randomly selects edges (and their nodes) until it reaches a specific percentage *pseeds* of nodes. The training sample then contains this percentage of nodes and the selected edges. Note that when sampling by edges, the resulting sample will be biased towards hubs since they get higher chances to be picked multiple times through their multiple connections.

[3] Minority group always refers to black nodes.

Degree rank We rank all nodes by their degree in descending order and select the top *pseeds*% of nodes. While computing the degree of all nodes is expensive, the idea is to verify whether high degree nodes improve the inference (Lin and Cohen 2010). The main difference compared to edge sampling is that degree sampling selects hubs without their neighbors (which may have low degree).

Partial crawls Yang et al. (2017) proposed a crawl-aware parameter estimator of peer-effect based on random walk tours. The procedure is detailed in Avrachenkov et al. (2016), but roughly described as follows. First, a fraction p_{sn} of nodes in V is sampled by some arbitrary sampling procedure (the method is insensitive to sampling biases in this phase). These sampled nodes are called a *super node* S. Second, t different random walks are performed starting at nodes in S, ending when the random walker encounters any node also in S. The starting node is chosen from S proportional to the number of edges from nodes in S to nodes outside S. The random walk progresses by moving to a random neighbor of the current visited node. Note that all nodes in the super node together with the crawled nodes in every tour belong to G_{seeds}[4]. The random walker stops once G_{seeds} contains a percentage *pseeds* of the total number of nodes in G. It has been shown that partial crawls can provide unbiased estimates of network statistics (Avrachenkov et al. 2016). Thus, we verify whether perfect model estimates always lead to perfect classification.

In RQ2 we compare the influence that each of these sampling techniques has on classification performance and parameter estimation. In RQ3 we examine their bias with respect to minority and majority groups. In RQ1 we explore the impact of the network structure on classification performance. Thus, to ensure that the sampling method does not influence the performance we only use random node sampling.

Example. Following the example in Fig. 1 (see "2. Sampling"), the subgraph extracted via random nodes consists of: 40% randomly selected nodes $\hat{V} = \{B, D\}$, their class labels (color) $\hat{c}_B = white$ and $\hat{c}_D = black$, and all edges between them $\hat{E} = \{(B, D)\}$.

Modeling and collective inference: estimates

Collective classification in networked data (Macskassy and Provost 2007; Getoor and Taskar 2000; Jensen et al. 2004) learns correlations between attributes of linked nodes from observed data, and transfers this knowledge simultaneously to the unseen nodes. This process consists of three components: local model, relational model, and collective inference. To isolate the effects of network and training sample, we fix the classification algorithm as follows. We (i) learn the local model LC from the nodes in the training sample, (ii) learn the relational model RC from the nodes and edges in the training sample using *Bayesian* statistics, and (iii) infer class values using *relaxation labeling* as the collective inference process CI. Therefore, the probability of a node $v_i \in V - \hat{V}$ with neighbors $\mathcal{N}_i$ taking on class $x_i = c$ is given by:

$$\overbrace{P(x_i = c|\mathcal{N}_i)}^{\text{posterior}} = \frac{\overbrace{P(x = c)}^{\text{prior}} \cdot \overbrace{P(\mathcal{N}_i|x_i = c)}^{\text{likelihood}}}{\underbrace{P(\mathcal{N}_i)}_{\text{marginal likelihood}}} \tag{2}$$

[4] This is an adaptation of the original algorithm to work with semi-supervised learning. That is, instead of crawling an unknown network, we extract a subgraph by sampling class labels of nodes from a known network.

where $P(\mathcal{N}_i|x_i = c) = \prod_{v_j \in \mathcal{N}_i} P(x_j = \widetilde{x}_j|x_i = c)$ and $\widetilde{x}_j$ is the actual class observed at node v_j. Parameters in the local and relational model include the prior probability $P(x = c)$ of any node being of class c, and conditional probabilities $P(x_j = \widetilde{x}_j|x_i = c)$ that a neighboring node v_j has class $\widetilde{x}_j$ given that node v_i has class c.

Parameter estimation The model parameters are inferred from the nodes and edges in the training sample. The estimates learned using the partial crawls algorithm are defined in Yang et al. (2017). The estimates learned using random nodes, random edges, and degree ranking are calculated as follows:

$$P(x = c) = \frac{1}{|\hat{V}|} \sum_{\hat{v}_i \in \hat{V}} \mathbb{1}\{\hat{c}_i = c\} \tag{3}$$

$$P(x_j = a|x_i = c) = \frac{\sum_{(\hat{v}_i,\hat{v}_j)\in\hat{E}} \mathbb{1}\{\hat{c}_i = c\} \cdot \mathbb{1}\{\hat{c}_j = a\}}{\sum_{(\hat{v}_i,\hat{v}_j)\in\hat{E}} \mathbb{1}\{\hat{c}_i = c\}} \tag{4}$$

Inference (relaxation labeling) Once the model parameters are learned, relaxation labeling initializes each unlabeled node with the prior probabilities. Then, rather than estimating one node at a time and updating the graph right away, the current estimations are frozen so that at step $t + 1$ all vertices will be updated based on the estimations of step t. The updating step takes into consideration a decay constant to regulate the influence of neighboring nodes in every iteration (Macskassy and Provost 2007).

Example. During the modeling phase in Fig. 1 (see "3. Modeling") we learn the prior probabilities (e.g., $P(x = black) = 0.5$) and the conditional probabilities (e.g., $P(x_j = white|x_i = black) = 0.66$). Continuing to the inference phase in Fig. 1 (see "4. Collective Inference"), the *relaxation labeling* first initializes the posterior probabilities of all unlabeled nodes using the class priors, and then iterates through all unlabeled nodes simultaneously to infer their posterior probabilities using the Bayes theorem, see Equation 2.

Experimental setup

To explore the interplay between network structure, sampling techniques, and the performance and bias of classification, we systematically vary structural properties of the *network* and the *training sample* by fixing the classification algorithm as explained below.

Synthetic networks We generate 330 undirected networks G using the model by Karimi et al. (2018), referred to as BA-Homophily[5], and adjust four parameters: number of nodes $N = 2000$, class balance $B \in \{0.1, 0.3, 0.5\}$, homophily $H \in \{0.0, 0.1, \ldots, 1.0\}$, and edge density $d \in \{0.004, 0.02\}$[6]. Networks are generated 5 times in each configuration to control for random fluctuations. We omitted results using smaller and larger networks, i.e., $N \in \{500, 10000\}$. The main difference across network sizes is that the variance of ROCAUC reduces with larger networks. However, their mean ROCAUC

[5] The acronym for Barabási-Albert Homophilic network.

[6] Density in the BA-Homophily model was originally adjusted (indirectly) via minimum degree $m \in \{4, 20\}$. Since degrees are power-law distributed and nodes have minimum degree of m, the larger the value of m the higher the density d.

scores are very similar to the ones obtained with $N = 2000$ (see Figure A2 in the Additional file 1). Due to this similarity we decided to show only results with $N = 2000$.

Training samples Subgraphs G_{seeds} contain a percentage *pseeds* of nodes from G that are selected by one of the following sampling methods: random nodes (*nodes*), random edges (*edges*), degree rank (*degree*) and partial crawls (*partial_crawls*). We assume that G_{seeds} is completely observed, which means that we know the class labels of nodes and all or some edges among them. We vary *pseeds* $\in \{5\%, 10\%, 20\%, ..., 90\%\}$ to measure the impact of sample size on classification. In the particular case of the partial crawls, we set the size of the super node to $|S| = p_{sn} \times N$, where $p_{sn} = 0.01$, and the number of tours t to as many as necessary until reaching *pseeds* $\times N$ of nodes in G_{seeds}. For each *pseeds*, we run the classification algorithm 10 times.

Classification algorithm We focus on uni-variate network classification, which means that the linkage structure in the network is modeled with the class label of the nodes and no information from additional node attributes. In particular, we choose the *network-only Bayes classifier* (nBC) as the relational model (RC), and apply *relaxation labeling* (RL) as the collective inference algorithm (CI). We use this combination for two reasons. First, it has been shown that RL outperforms other CIs when training samples are small, and when training samples are large any CI performs equally well (Macskassy and Provost 2007). Second, the nBC model parameters are easy to interpret since they are based on network structure (i.e., class priors relate to class balance, and conditional probabilities to homophily). Additionally, we show that the overall trend of classification performance vs. network structure does not vary with a different RC, namely *LINK classifier* (Zheleva and Getoor 2009) (see "To what extent do these results depend on the algorithm?").

Evaluation We quantify the performance of the classification using three different metrics: (i) **ROCAUC score**[7] to estimate the overall performance of the collective classification, (ii) **squared estimation errors (SE)** between global and sample parameters to assess the quality of the parameter estimation, and (iii) a comparison between the **true positive rates** of each class to measure the direction of bias. Note that when working with unbalanced data, a classifier may achieve high overall performance even if it often misclassifies instances of the minority class. By comparing the positive rates—sensitivity and specificity in binary classification—we disentangle the direction of bias and assess how well the algorithm classified both, minority and majority classes.

Results

Using *naive Bayes* and *relaxation labeling*, we classify nodes as either *white* or *black* using different sample sizes and different evaluation metrics (see section "Experimental setup"). Next, we will discuss our results and answer the three research questions which we raised before.

RQ1: How does *network structure* affect the overall performance of collective classification?

We analyze to what extent the structure of the network (i.e., *homophily*, *class balance* and *edge density*) impacts classification performance. We measure performance of

[7] Area under the receiver operating characteristic curve.

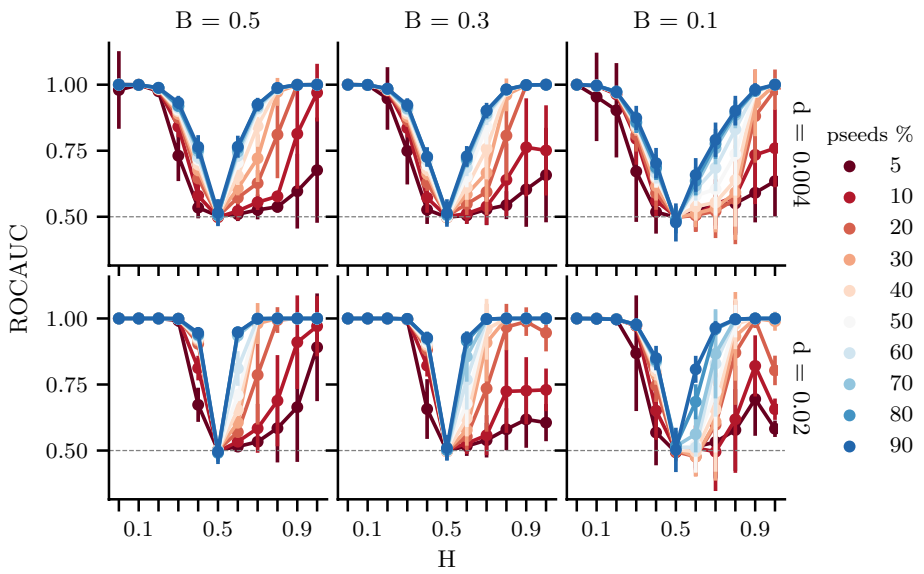

Fig. 2 RQ1: Network structure, sample size and overall performance using random node sampling and network-only Bayes classifier (nBC). Classification performance, measured by ROCAUC scores, is shown on the y-axis for networks with $N = 2000$ nodes and different levels of homophily (x-axis), edge density (rows), class balance (columns), and sample size (colors). Dots represent mean ROCAUC scores over 50 runs, and error bars their respective standard deviation. In general we see that: (i) neutral networks $H = 0.5$ cannot be classified better than a random classifier; (ii) heterophilic networks $H < 0.5$ require smaller samples to achieve high and stable classification performance compared to homophilic networks $H > 0.5$; (iii) Dense networks $d = 0.02$ achieve higher ROCAUC compared to sparse networks $d = 0.004$

collective classification using ROCAUC scores, where each value can be interpreted as the probability of distinguishing between classes.

Overall performance vs. network structure

Figure 2 shows the classification performance on synthetic networks with number of nodes $N = 2000$ and edge density $d \in \{0.004, 0.02\}$ (rows)[8]. Class balance is defined by the parameter B (columns). Homophily H ranges from 0 to 1 (x-axis). Sample size, using random node sampling is shown as the percentage *pseeds* of nodes (colors), and the overall performance as ROCAUC scores (y-axis). At first glance, from Fig. 2 we notice four main patterns. **(i)** As expected, classification performance on neutral networks ($H = 0.5$) is always similar to a random classifier. **(ii)** Surprisingly, heterophilic networks ($H < 0.5$) require smaller samples to achieve high and stable classification performance compared to homophilic networks ($H > 0.5$). **(iii)** ROCAUC scores are neither stable nor consistent (i.e., high variance) in the homophilic regime when samples are very small. In other words, classification performance varies widely. **(iv)** Dense networks (d=0.02) achieve higher classification performance compared to sparse networks (d=0.004) around $H = 0.5 \pm 0.3$, i.e., $\overline{\text{ROCAUC}}_{d=0.02, H=0.5\pm0.3} = 0.82 > \overline{\text{ROCAUC}}_{d=0.004, H=0.5\pm0.3} = 0.74$.

[8] A different visualization can be found in Figure A1 in the Additional file 1.

Why is heterophily easier to predict?

In Fig. 2 we see an asymmetry between homophilic ($H > 0.5$) and heterophilic ($H < 0.5$) regimes for small samples (red lines) and all class balance levels B (columns). To explain this discrepancy, we turn to the properties of the sampling error[9] and the network structure: Undirected networks only contain three types of edges, e.g., black-white, white-white, and black-black. In the heterophilic regime, only one type of edge is prevalent (black-white), while in the homophilic regime two types are equally prevalent (white-white, black-black). In general, for small training samples (e.g., $pseeds \leq 30\%$), the probability of correctly observing each type of edge is very low. Consequently, the parameter estimation is prone to be wrong. However, its impact depends on the class balance and homophily of the network.

Balanced networks, B=0.5 First, note that the probability of observing a black-black edge in the synthetic network can be calculated analytically given the homophily (H), the class balance (B), and the degree exponents of the groups (β) as follows:

$$P_{bb} = \frac{B^2 H(1-\beta_w)}{Z} \tag{5}$$

where, Z is a normalization constant, and β_b and β_w are the exponents of the degree distribution for the *black* and *white* nodes, respectively. For the detailed analytical derivations and values of β see (Karimi et al. 2018). Similarly, the probability of observing a black-white edge is given by:

$$P_{bw} = \frac{B(1-B)(1-H)[(1-\beta_b)+(1-\beta_w)]}{Z} \tag{6}$$

In the heterophilic case ($H = 0.2$), the probability of observing a black-white edge in the whole graph is 0.8. Thus, the sampling error in a small sample follows $(0.8|\hat{E}|)^{-\frac{1}{2}}$, where $|\hat{E}|$ is the total number of edges in the sample. In the homophilic case ($H = 0.8$), the probability of observing a black-black edge is 0.4 and a white-white edge is also 0.4. The sampling error for *each* homophilic class is then $(0.4|\hat{E}|)^{-\frac{1}{2}}$ which individually are smaller than the error in the heterophilic case but adding them together they are larger. These sampling errors are reflected in the *estimation error* calculated here as the squared distance between the model parameter inferred from the training sample ($P\{.\}$) and the global network ($\theta\{.\}$):

$$SE\{.\} = (P\{.\} - \theta\{.\})^2 \tag{7}$$

We see these errors in the left-most column of Fig. 3, where the x-axis refers to $SE_{maj|maj}$, and the y-axis to $SE_{min|min}$. Note that large errors in homophilic networks ($H = 0.8$) lead to low overall performance (brown). However, there are some cases where performance is also low even though such errors are small. This means that homophilic networks are more sensitive to the precision of the parameter estimation because it requires: $P_{maj|maj} = P_{min|min}$.

Unbalanced networks, B<0.5 In addition to the sampling error explained above, the group size differences and the inherent structure of the network add additional

[9] The error caused by observing a sample instead of the whole population.

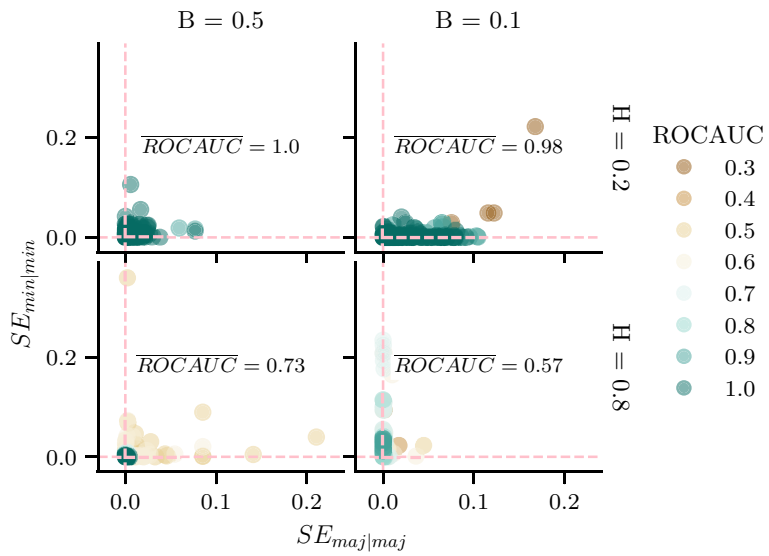

Fig. 3 RQ1: Estimation errors on small samples using random node sampling. These are the squared estimation errors (SE) of the conditional probabilities $P_{maj|maj}$ (x-axis) and $P_{min|min}$ (y-axis) learned from small samples (*pseeds* $\leq$ 30%) using random nodes on sparse networks ($N = 2000, d = 0.004$) with different levels of homophily (rows) and class balance (columns). Mean ROCAUC scores within each type of network is shown as $\overline{ROCAUC}$. In general we see that homophilic networks require lower estimation errors—especially within majorities—to achieve high performance (green)

complexity to the learning process. This happens because of the interplay between homophily and preferential attachment which enables the formation of all different types of connections. For instance, in *homophilic networks* ($H = 0.8$), minority nodes will be mainly attracted by other minority nodes. However, due to the preferential attachment, minority nodes will also be partly attracted to majority nodes. On the other hand, majority nodes will be mostly connected to other majority nodes due to both mechanisms. Therefore, the estimation error of the conditional probability $P_{maj|maj}$ is on average lower than the estimation error for $P_{min|min}$, as shown at the bottom-right plot in Fig. 3. The same principle applies to *heterophilic networks* ($H = 0.2$). In this case, even though most edges are heterophilic, networks will also contain edges between nodes of the same type but in significantly different proportions. Since there is only a very limited number of minority nodes, there can only be a very limited number of edges between them. That is not the case for majorities because they can connect to many more majorities. Therefore, though locally they connect to a few other majorities, globally there are many edges within this group. This gives an advantage to small samples because the randomly selected majority nodes are likely to be either disconnected[10] or connected to other minority nodes that are in the training sample. Thus, the classifier learns that the network is heterophilic. This explains why heterophilic networks can achieve high overall performance even when estimation errors are high for $P_{maj|maj}$ as shown in the top-right plot in Fig. 3. This holds as long as $\frac{P_{maj|maj}}{P_{min|maj}} \times \frac{P_{min|min}}{P_{maj|min}} < 1$, otherwise the classifier believes that the network is extremely homophilic.

[10] When their neighbors belong to the majority group but are not in the sample.

Finally, besides these conditional probabilities, class priors are also important in the collective inference. Thus, in the balanced case ($B = 0.5$), we expect the class priors to be the same: $P_{min} = P_{maj} = 0.5$; if this condition is not fulfilled, the classifier initially believes that one group is more prevalent than the other[11]. In the unbalanced case ($B = 0.1$), however, it is enough to identify the minority group correctly, regardless of its actual group size.

To what extent do these results depend on the algorithm?

For interpretability reasons we chose the network-only Bayes classifier (nBC) as relational model, since its model parameters correlate with the homophily and class balance of the network. However, it is unclear whether the results shown in Fig. 2 are to some extent a product of the relational classifier. Therefore, we run the classification algorithm on the same networks by changing the relational model. We choose the LINK classifier (Zheleva and Getoor 2009; Altenburger and Ugander 2018), which learns a regularized logistic regression. The features of a node are the entire row of the adjacency matrix and the outcome variable is the node's class. In this case, the model parameters are not based on the classes of the nodes (as in nBC), but purely on all nodes in the network. Results using this new setup are shown in Figure A3 in the Additional file 1. We see that the main patterns—compared to the results using nBC—persist. Classification performance achieves its best scores in the extreme levels of homophily, and it drops when networks are neutral. Also, classification on heterophilic networks is just slightly better than classification on homophilic networks. However, the most notorious difference between LINK and nBC is the performance across sample sizes. First, we notice that when using nBC, performance drops drastically when using small training samples on homophilic networks. Second, in this regime performance is not stable (i.e., high variance), see Fig. 2. These two issues do not appear in the results when using LINK, see Figure A3 in the Additional file 1. Therefore, we can conclude that performance, in terms of ROCAUC scores, is mainly driven by the type of network (i.e., the interplay between homophily, class balance, edge density and preferential attachment). When it comes to sample size, nBC gets penalized by small samples since their fluctuations introduce noise in the model parameters, while the parameters of LINK never change.

RQ2: How does the choice of the sampling technique affect the overall performance of collective classification and its parameter estimation?

In section "RQ1: How does *network structure* affect the overall performance of collective classification?" we learned that certain properties of the network structure help in the parameter estimation even when training samples are very small. Now, we compare random node sampling with three other sampling methods, two of them are *biased* towards high degree nodes (*random edge sampling* and *degree ranking*), and one is *unbiased* (*partial crawls*); more details in section "Sampling: the observed network".

Since the focus is on the sampling techniques, we fix the number of nodes and edge density of networks to $N = 2000$ and $d = 0.004$, respectively. We also omit results on

[11] In fact, larger fluctuations are more likely in small samples

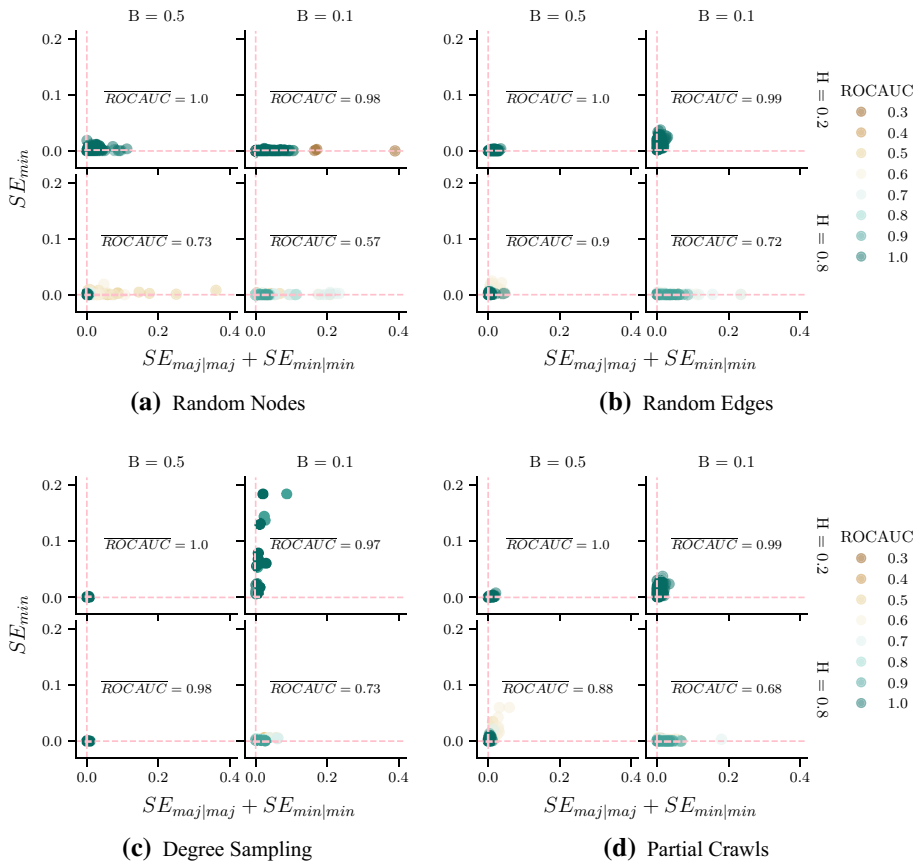

Fig. 4 RQ2: Parameter estimation and overall performance on small samples. These are the squared estimation errors of class priors (y-axis) and conditional probabilities (x-axis) with their respective overall performance (colors) on small samples (*pseeds* $\leq$ 30%) using different sampling techniques: **a** random nodes, **b** random edges, **c** degree sampling, and **d** partial crawls. Class balance and homophily values are shown as columns *B* and rows *H*, respectively. Mean ROCAUC scores per network type (B,H) are shown as $\overline{ROCAUC}$. Conditional probabilities obtained by partial crawls ($SE_{maj|maj} + SE_{min|min} < 0.18$) are the most accurate, followed by random edges, degree sampling and random nodes. However, the most accurate class priors are obtained by random nodes ($SE_{min} < 0.02$), followed by random edges, partial crawls, and degree rank. In terms of performance, on average degree sampling achieves the highest ROCAUC followed by random edges, partial crawls, and random nodes. Surprisingly, perfect estimates ($\sum SE = SE_{min} + SE_{maj|maj} + SE_{min|min} \approx 0$) do not guarantee perfect performance (ROCAUC $\approx$ 1.0), and vice versa

neutral networks, and large sample sizes since their performance is either consistent or often very high. Results are shown in Fig. 4. The x-axis represents the sum of the squared estimation errors of conditional probabilities $P_{maj|maj}$ and $P_{min|min}$, the y-axis shows the squared estimation error of the class prior P_{min}, and colors represent the overall performance.

Random nodes vs. other sampling techniques: First, if we look at the estimation errors from the class prior and the conditional probabilities separately (as shown in Fig. 4) we notice that random edges, degree sampling, and partial crawls are better at estimating conditional probabilities than random nodes. This is because conditional probabilities are based on connections between nodes and all three sampling methods exploit these

connections during the sampling. Second, not surprisingly, random node sampling is on average better at estimating class priors since it observes a random sample of nodes, and the class prior only depends on the prevalence of node attributes. Third, on average degree sampling achieves the highest performance ($\overline{\text{ROCAUC}} \approx 0.91$) followed by random edges, partial crawls, and random nodes ($\overline{\text{ROCAUC}} \approx 0.81$). On the other hand, partial crawls sampling provides the most accurate estimates followed by random edges, random nodes, and degree ranking. However, depending on the structure of the network these sampling techniques may improve or worsen their overall performance and parameter estimation as described below.

Trade-off between homophily and class balance: In terms of overall performance, all sampling techniques perform equally well in heterophilic networks in both the balanced and unbalanced regimes ($\overline{\text{ROCAUC}}_{H=0.2} \approx 0.97$). Similarly, all sampling techniques perform equally well in homophilic networks ($\overline{\text{ROCAUC}}_{H=0.8} \approx 0.76$). However, this performance is proportional to the class balance: low for unbalanced networks ($\overline{\text{ROCAUC}}_{H=0.8,B=0.1} \approx 0.67$), and high for balanced networks ($\overline{\text{ROCAUC}}_{H=0.8,B=0.5} \approx 0.85$). Last but not least, we also see in Fig. 4 that the most accurate estimates across sampling techniques are obtained in balanced networks, especially when they are also heterophilic (more details in Figure A4 in the Additional file 1).

Which sampling technique should we use?

If the goal is to achieve high overall performance ($\overline{\text{ROCAUC}} \approx 1.0$) with a small sample, random edge sampling or partial crawls should be used in *heterophilic* networks[12], and degree sampling in *homophilic* networks, as long as the degree of nodes is available, otherwise random edges should be considered. However, if the goal is to achieve good quality of estimates ($\sum SE = SE_{min} + SE_{maj|maj} + SE_{min|min} \approx 0$) with a small sample, then the most accurate estimates are obtained by degree ranking (followed by partial crawls) when networks are *balanced*, and partial crawls when networks are *unbalanced* (see Figure A4 in the Additional file 1).

RQ3: How does network structure and the choice of sampling technique influence the direction of bias in collective classification?

Now, we explore how classification mistakes are distributed across both classes. If mistakes are concentrated in one class, the classifier is biased against that class. For example, when data is unbalanced, a majority class classifier will be highly accurate, but misclassify—or be biased against—the minority class. To disentangle *how well the algorithm classifies both minority and majority* classes, we extend the balanced accuracy (Brodersen et al. 2010) to assess the direction of bias. We then compare the true positive rates (TPR) of each class as follows:

$$bias = \frac{TPR_{min}}{TPR_{min} + TPR_{maj}} \tag{8}$$

[12] In this regime these sampling techniques are on average slightly better than the other methods, though all of them perform equally well.

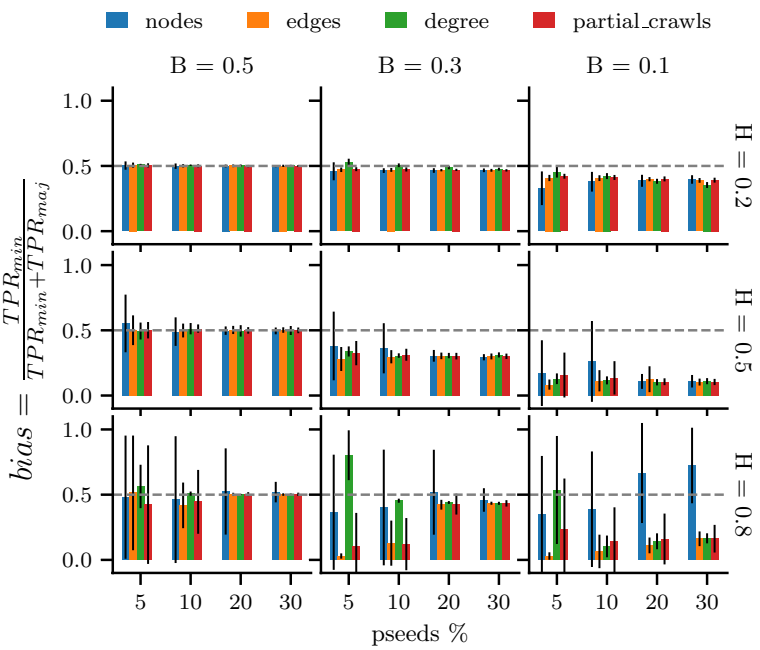

Fig. 5 RQ3: Direction of bias. We measure the direction of bias by comparing the true positive rates of each class. The unbiased case is *bias* = 0.5, when both classes have the same true positive rates. Otherwise, results are biased towards majority nodes *bias* < 0.5, or towards minority nodes *bias* > 0.5. These are observations on small samples (*pseeds* ≤ 30%) where fluctuations are high, after this point bias scores converge or get better (i.e., towards *bias* = 0.5). We see that classification on heterophilic networks is less biased than in neutral and homophilic networks

Since our classification task is on a binary attribute, TPR_{min} refers to *sensitivity* and TPR_{maj} refers to TNR_{min}[13] or *specificity*. This bias score ranges from 0 to 1. Depending on its value, classification can be interpreted as: (a) bias < 0.5: biased towards majorities (or against minorities), (b) bias > 0.5: biased towards minorities (or against majorities), and (c) bias $= 0.5$: unbiased.

Results on networks with fixed number of nodes ($N = 2000$) and fixed density ($d = 0.004$), using the four sampling techniques, are shown in Fig. 5. Large samples (*pseeds* > 30%) are not shown since bias scores converge at that point for almost all cases[14].

On average, classification results are unbiased in balanced networks ($B = 0.5$). Additionally, when class balance decreases ($B < 0.5$), classification results are often biased towards majority nodes. However, depending on the level of homophily of the network, the bias score decreases considerably in neutral and homophilic networks ($H \in \{0.5, 0.8\}$), or just slightly in heterophilic networks ($H = 0.2$). Notice as well that in the homophilic regime the standard deviation is high. This means that the variation with respect to which group is classified correctly is high. These results are consistent across all sampling methods, and indicates that unbiased results are more robust to changes in group-size and sampling choice in heterophilic networks than in neutral and homophilic networks.

[13] True negative rates of the minority class.

[14] For larger samples in networks with $B = 0.1$ and $H = 0.8$, the bias score slightly increases (but it is still biased against the minority nodes).

Table 1 *Empirical networks* Structural properties of five real-world networks: Escorts, Swarthmore42, Caltech36, Wikipedia, and GitHub

Dataset	Escorts	Swarth.	Caltech	Wiki.	GitHub
N	16730	1519	701	2132	37700
m	1	1	1	1	1
class	role	gender	gender	gender	dev
minority	escort	2 (m)	1 (f)	female	1 (ML)
B	0.40	0.49	0.33	0.15	0.26
E	39044	53726	15464	3143	289003
d	0.0003	0.05	0.06	0.001	0.0004
β	2.87	5.50	4.90	2.87	2.54
H	0.00	0.52	0.54	0.64	0.84
N_{fit}	14338	208	179	2893	9830
m_{fit}	2	2	2	2	2

In addition to the properties of interest, we report β, the power-law exponent of the degree distribution computed as described in Karimi et al. (2018). N_{fit} and m_{fit} represent the number of nodes and minimum degree utilized to generate synthetic networks, respectively

Surprisingly, there are a few cases where classification is biased against majority nodes (bias > 0.5). Specifically in homophilic networks when nodes are sampled randomly (blue) or by degree (green). Thus, their classification performance is low $ROCAUC < 0.6$ (see Figure A5(a) and Figure A5(c) in the Additional file 1). On the other hand, when classification results are unbiased (bias $= 0.5$) or biased against minority nodes (bias < 0.5), their classification performance is high $ROCAUC > 0.8$ and inversely proportional to the sum of estimation errors (see Figure A5 in the Additional file 1).

Future work should investigate new sampling methods or classifiers that focus on overcoming the bias issue of collective classification especially in homophilic and neutral networks. For instance, one promising direction that has been proven to improve performance, especially for neutral networks, is to look at friends-of-friends similarities in the parameter estimation (Altenburger and Ugander 2018).

Empirical networks

Finally, we focus on five real-world networks, described in Table 1, and show that the utilized network model (Karimi et al. 2018), allows for computing a baseline for the performance that a collective classifier can achieve on empirical social networks.

Real-world networks: Sexual contact network: The Escorts dataset represents a network of sexual contacts from Brazil (Rocha et al. 2010). Nodes are of two types: client or escort. *Friendship networks:* Swarthmore42 and Caltech36 are University networks which include friendship links between user's Facebook pages (Traud et al. 2012). Every node in each network represents a member of the school. For the purpose of our experiments we choose the attribute *gender* $\in \{1(\text{female}), 2(\text{male})\}$ as class label. *Hyperlink network:* This is a hyperlink network of American politicians in Wikipedia (Wagner 2017). We consider reciprocal edges in order to treat it as an undirected network. We use the politician's gender as class label. *Following network:* The GitHub dataset is a large social network of GitHub developers (Rozemberczki et al. 2019). Nodes are developers who have starred at least 10 repositories and edges are mutual follower relationships

between them. Nodes possess a binary attribute that describes whether the user is a web or a machine learning developer.

We remove nodes without class label, and nodes with no edges. Note that all these networks are scale-free (i.e., power-law degree distributed). Table 1 shows the value of the power-law exponent β of each dataset, and Figure A6 (see the Additional file 1) shows their degree distributions.

BA-Homophily networks: For each dataset we generate 5 synthetic networks using the BA-Homophily model with symmetric homophily (e.g., $h_{aa} = h_{bb}$), and adapt the algorithm to fulfill the edge density condition (Espín-Noboa 2021). We pass to the algorithm five parameters shown in Table 1: number of nodes N_{fit}, class balance B, homophily H, edge density d, and minimum degree m_{fit}[15].

Results on empirical networks: As in section "Experimental setup", we run the collective classification algorithm 10 times for each sample size, and report its performance using mean ROCAUC scores. In Fig. 6 we see the classification performance (y-axis) on five real-world networks (columns) for different sample sizes (x-axis). The (expected) ROCAUC using each synthetic network is shown as "BA-Homophily" (light blue), and the (observed) ROCAUC using the real networks is shown as "empirical" (dark blue). Note that results have been sorted by H (homophily) in ascending order. We see a similar pattern in Fig. 2, where ROCAUC scores besides getting higher with larger sample sizes, they also get higher upper-bounds with values of homophily far from $H = 0.5$. Moreover, we see that the synthetic networks are able to mimic the collective classification performance of real-world networks. As expected, the Escorts network has the best fit between data and model, because it is extremely heterophilic (see section "RQ1: How does *network structure* affect the overall performance of collective classification?"). The Caltech and Wikipedia networks also show very good fit. Note that Swarthmore42 has a perfect fit only for small samples ($pseeds < 50\%$). Conversely, when training samples are small, GitHub shows high variance. These discrepancies might be due to other network properties that the model does not capture. For instance, rich mixing patterns might get ignored by summarizing homophily with a global statistic (Peel et al. 2018; Peel 2017). In other words, while the global behavior is captured in the sample, some nodes can exhibit local differences. Similarly, real-world networks might exhibit asymmetric homophily ($h_{aa} \neq h_{bb}$) and high clustering (Holme and Kim 2002).

Discussion and future work

Many popular applications rely on peer information and other types of relational data. For instance, peer-to-peer lending systems (Bachmann et al. 2011) allow people to borrow/lend money to connected friends. Such systems utilize machine learning algorithms to infer credit scores of individuals using the credit score of their friends (i.e., high/low risk) (Lin et al. 2013; Hadji Misheva et al. 2019; Liet al. 2020). In such cases, it is extremely important to understand and explain the overall performance of these algorithms, as well as their impact on different groups, especially minorities. Our results highlight that especially in homophilic and neutral networks, minorities may be at a disadvantage when the classifier is trained on a small sample.

[15] Although m is required for the BA-Homophily model, it does not affect the classification results because the behavior of the network is independent of m (Karimi et al. 2018).

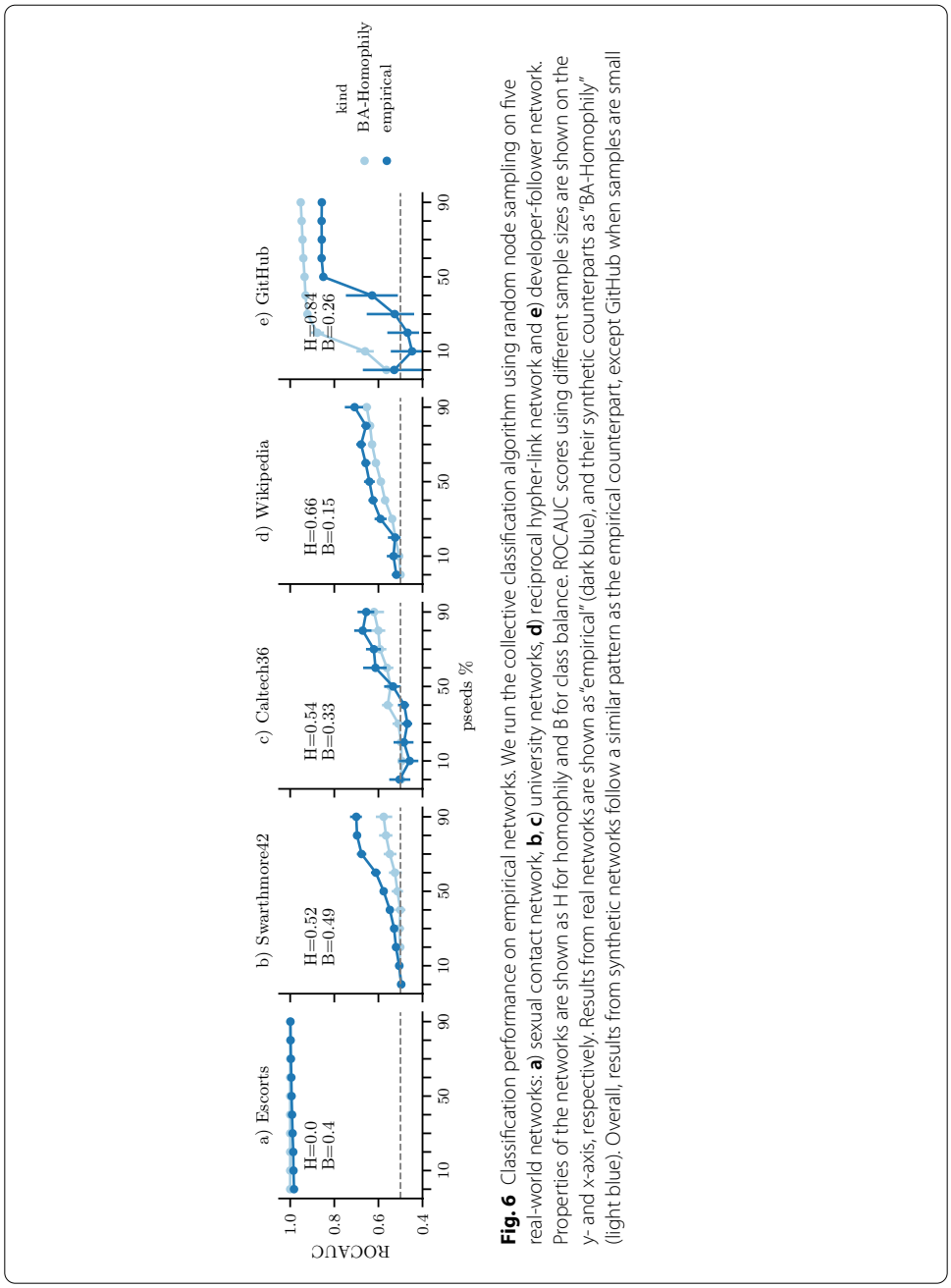

Fig. 6 Classification performance on empirical networks. We run the collective classification algorithm using random node sampling on five real-world networks: **a**) sexual contact network, **b**, **c**) university networks, **d**) reciprocal hypher-link network and **e**) developer-follower network. Properties of the networks are shown as H for homophily and B for class balance. ROCAUC scores using different sample sizes are shown on the y- and x-axis, respectively. 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, except GitHub when samples are small

Additionally, we show that homophily, class balance, edge density, and sample size impact the performance and the direction of bias of collective classification. Interestingly, we find that larger training samples are required in undirected networks with homophilic connections to achieve a similar high classification performance compared to heterophilic networks. This fundamental difference between heterophilic and homophilic networks can be explained by the sampling error and the network structure; in particular undirected edges, class balance, homophily, and preferential attachment (see section "RQ1: How does *network structure* affect the overall performance of collective classification?"). This suggests that the structure of a social network can help to infer the needed sample size to achieve high classification performance, and to be aware of potential bias issues.

Our comparison of sampling techniques suggests that although partial crawls sampling provides the most accurate estimates (i.e., the prior and conditional probabilities learned from their training samples are closest to what we observe in the full network), the inherent bias of degree and edge sampling (towards high degree nodes), help the classifier to achieve very high performance. This suggests that accurate estimates do not always lead to perfect classification performance, and vice versa. Concretely, we observed that when sampling by degree on heterophilic and unbalanced networks, perfect classification performance can also be achieved based on imperfect estimates, in particular the class prior. This can be explained by the fact that in heterophilic networks, minority nodes have high degree, and heterophilic edges are predominant. Thus, a very small sample (sampled by degree) will mostly contain minority nodes and a few majority nodes that are linked to the minorities (see section "RQ1: How does *network structure* affect the overall performance of collective classification?"). Surprisingly, we also find that perfect estimates do not guarantee perfect classification performance, especially in homophilic and unbalanced networks across all sampling techniques. One explanation could be that only a few unlabeled nodes are connected to the seed nodes as 1-hop neighbors. Thus, long cascades of unlabeled nodes (in k-hops; $k > 1$) might propagate erroneous information. Further studies are needed to explain this behavior.

Use case. Due to the design of our experiments, we always had access to the class balance and homophily of the network. In a real scenario, however, these properties might be unknown. Then, how can we evaluate results on new datasets if the structure of the network is unknown? It has been shown in Avrachenkov et al. (2016) that partial crawls may obtain unbiased and reliable node and edge statistics. We corroborated that in section "RQ2: How does the choice of the sampling technique affect the overall performance of collective classification and its parameter estimation?". Therefore, in cases when these properties are unknown, a small sample taken by partial crawls can already capture accurately the structure statistics of the whole network, e.g., class priors and conditional probabilities (see section "Modeling and collective inference: estimates").

Limitations. Our results are limited to undirected networks to simplify our exposition. However, results may and should be extended to directed networks including larger sets of attributes per node, and different levels of asymmetric homophily in both directed and undirected networks. Additionally, while in this work we focused on social networks with preferential attachment and homophily, further research may investigate the impact of other mechanisms of edge formation on relational classification. For instance, by disentangling the effects of homophily and triadic closure (Holme and Kim 2002; Asikainen et al.

2020). Finally, we focus on one specific collective inference method (relaxation labeling) and mostly on one relational model (network-only Bayes classifier, nBC). The comparison of multiple methods is a potential avenue to disentangle algorithmic bias. For instance, when we compared results using nBC and LINK (see section "To what extent do these results depend on the algorithm"), we learned that their main differences rely on the training sample size. In particular, we found that the model parameters of nBC get distorted when training samples are small. These fluctuations affect performance since the nBC parameters must learn the correct class balance and homophily of the network in order to increase the chances of achieving high performance. Even when these parameters are correct, if the training sample is too small, unlabeled nodes might not reach sufficient seed nodes and this might lead to erroneous collective inference. This does not occur in LINK since its parameters only depend on the presence or absence of a node as a neighbor.

Conclusions

Collective classification is often used to infer missing attributes of nodes in networks. However, which factors impact its performance? And under which conditions is inference biased towards minority or majority groups? This paper provides answers to these questions by systematically analyzing the impact of network structure and sampling technique on the performance of collective classification. Our findings suggest that (i) mean classification performance can empirically and analytically be predicted by homophily, class balance, edge density, and sample size, (ii) networks with homophilic connections require larger training samples than heterophilic networks to achieve comparable performance, (iii) when sampling budgets are small, on average, partial crawls and edge sampling achieve the most accurate model estimates, and (iv) classification results are often less biased in heterophilic networks than in homophilic networks regardless of class imbalance.

Abbreviations

LC: Local classifier or local model.; RC: Relational classifier or relational model.; CI: Collective inference.; nBC: Network-only Bayes classifier.; wvRN: Weighted-vote relational neighbor classifier.; cdRN: Class-distribution relational neighbor classifier.; RL: Relaxation labeling inference.; nLB: Network-only link-based classification.; w_{ij}: Edge weight.; $\mathbb{R}$: Real numbers.; ROCAUC: Area under the receiver operating characteristic curve.; SE: Squared estimation error.; TPR: True positive rate; G: Attributed unweighted graph.; V: Set of nodes in G.; N: Number of nodes in V.; E: Set of undirected edges in G.; C: List of class labels for each node in V.; min: Minority group of nodes in V.; maj: Majority group of nodes in V.; H: Homophily in G: the probability of nodes with the same class label to be connected.; B: Class balance or fraction of minority nodes in G.; d: Edge density.; pseeds: Sample size: the percentage of nodes to be sampled from V.; G_{seeds}: Subgraph extracted from G.; $\hat{V}$: Set of nodes in G_{seeds}.; $\hat{E}$: Set of undirected edges in G_{seeds}.; $\hat{C}$: List of class labels for each node in $\hat{V}$.; p_{sn}: Super-node size: the initial percentage of nodes to be sampled from V when using the partial crawls sampling.; S: Super-node: the initial subset of nodes sampled from V required by the partial crawls sampling..

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-021-00394-3.

Additioanl file 1. Figure A.1. RQ1: Overall performance vs. Network structure. **Figure A.2. RQ1:** Overall performance vs. Network size. **Figure A.3. RQ1:** Network structure, sample size and overall performance using random node sampling and LINK classifier. **Figure A.4. RQ2:** Overall performance vs.overall parameter estimation on small samples. **Figure A.5 RQ1, RQ2, RQ3:** Overall performance, bias, and parameter estimation on small samples. **Figure A.6** Degree distribution of empirical networks.

Acknowledgements
We would like to thank all anonymous reviewers as well as Juhi Kulshrestha, Indira Sen, and Mattia Samori who have helped improving this paper. Special thanks to the LXAI@ICML2020 and NetSci2020 communities for their fruitful discussions around this paper. Last but not least, our gratitude to the "Complex networks and their Applications" organizers for their support.

Authors' contributions
Conceived and designed the experiments: LE KL CW FK BR. Performed the experiments: LE BR. Analyzed the data: LE. Contributed reagents/materials/analysis tools: LE. Wrote the paper: LE FK BR KL CW. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials
All data and code are made publicly available on GitHub at Espín-Noboa (2021).

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Computational Social Science, GESIS, Cologne, Germany. [2]Computer Science, University of Koblenz-Landau, Koblenz, Germany. [3]Computer Science, Purdue University, West Lafayette, IN, USA. [4]Computer Science, University of Southern California, Los Angeles, USA. [5]Artificial Intelligence, Information Sciences Institute, Los Angeles, USA. [6]Computational Social Science, RWTH Aachen University, Aachen, Germany.

Received: 13 January 2021 Accepted: 1 July 2021
Published online: 21 October 2021

References

Altenburger KM, Ugander J (2018) Monophily in social networks introduces similarity among friends-of-friends. Nat Hum Behav 2(4):284
Asikainen A, Iñiguez G, Ureña-Carrión J, Kaski K, Kivelä M (2020) Cumulative effects of triadic closure and homophily in social networks. Sci Adv 6(19):7310
Avrachenkov K, Ribeiro B, Sreedharan JK (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, pp 165–177
Bachmann A, Becker A, Buerckner D, Hilker M, Kock F, Lehmann M, Tiburtius P, Funk B (2011) Online peer-to-peer lending-a literature review. J Internet Bank Commerce 16(2):1
Barabási A-L (2009) Scale-free networks: a decade and beyond. Science 325(5939), 412–413
Binns R (2020) On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 514–524 (2020)
Borgatti SP, Carley K, Krackhardt D (2006) Robustness of centrality measures under conditions of imperfect data. Soc Netw 28(1):124–136
Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) The balanced accuracy and its posterior distribution. In: 2010 20th international conference on pattern recognition, pp 3121–3124. IEEE
Coscia M, Rossi L (2018) Benchmarking api costs of network sampling strategies. In: 2018 IEEE international conference on big data (Big Data), pp 663–672. IEEE
Costenbader E, Valente TW (2003) The stability of centrality measures when networks are sampled. Soc Netw 25(4):283–307. https://doi.org/10.1016/s0378-8733(03)00012-1
Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, pp 67–73
Dong S, Liu D, Ouyang R, Zhu Y, Li L, Li T, Liu J (2019) Second-order markov assumption based bayes classifier for networked data with heterophily. IEEE Access
Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference, pp 214–226
Dwork C, Immorlica N, Kalai AT, Leiserson M (2018) Decoupled classifiers for group-fair and efficient machine learning. In: conference on fairness, accountability and transparency, pp 119–133 (2018)
Espín-Noboa L (2021) Discrimination-in-relational-classification. GitHub Repository. https://github.com/gesiscss/Discrimination-in-Relational-Classification
Espín-Noboa L, Wagner C, Karimi F, Lerman K (2018) Towards quantifying sampling bias in network inference. In: Companion of the the web conference 2018 on the web conference 2018, pp 1277–1285. International World Wide Web Conferences Steering Committee
Farnadi G, Babaki B, Getoor L (2018) Fairness in relational domains. In: Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society, pp 108–114
Galaskiewicz J (1991) Estimating point centrality using different network sampling techniques. Soc Netw 13(4):347–386
Getoor L, Taskar B (2007) Introduction to statistical relational learning. MIT Press, Cambridge

Hadji Misheva B, Spelta A, Giudici P (2019) Network based scoring models to improve credit risk management in peer to peer lending platforms. Front Artif Intell 2:3
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in neural information processing systems, pp 3315–3323
Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E 65(2):026107
Hughes DJ, Rowe M, Batey M, Lee A (2012) A tale of two sites: Twitter vs. facebook and the personality predictors of social media usage. Comput Hum Behav 28(2):561–569
Huisman M (2009) Imputation of missing network data: some simple procedures. Soc Struct 10(1):1–29
Jensen, D., Neville, J., Gallagher, B.: Why collective inference improves relational classification. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, pp 593–598 (2004). ACM
Kallus N, Mao X, Zhou A (2019) Assessing algorithmic fairness with unobserved protected class using data combination. arXiv preprint arXiv:1906.00285
Karimi F, Génois M, Wagner C, Singer P, Strohmaier M (2018) Homophily influences ranking of minorities in social networks. Sci Rep 8
Kossinets G (2006) Effects of missing data in social networks. Soc Netw 28:247–268
Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., Kompatsiaris, Y.: Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In: Proceedings of the 2018 world wide web conference, pp 853–862 (2018)
Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Advances in neural information processing systems, pp 4066–4076
Larrimore L, Jiang L, Larrimore J, Markowitz D, Gorski S (2011) Peer to peer lending: the relationship between language features, trustworthiness, and persuasion success. J Appl Commun Res 39(1):19–37
Lee J, Pfeffer J (2015) Estimating centrality statistics for complete and sampled networks: Some approaches and complications. In: 48th Hawaii international conference on system sciences, HICSS 2015, Kauai, Hawaii, USA, January 5–8, 2015, pp 1686–1695. https://doi.org/10.1109/HICSS.2015.203
Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 631–636. ACM
Lin F, Cohen WW (2010) Semi-supervised classification of network data using very few labels. In: 2010 international conference on advances in social networks analysis and mining, pp 192–199. IEEE
Lin M, Prabhala NR, Viswanathan S (2013) Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. Manage Sci 59(1):17–35
Li J-Y, Yeh M-Y (2011) On sampling type distribution from heterogeneous social networks. In: Proceedings of the 15th pacific-asia conference on advances in knowledge discovery and data mining - volume Part II. PAKDD'11, pp 111–122. Springer, Berlin, Heidelberg. http://dl.acm.org/citation.cfm?id=2022850.2022860
Li Y, Ning Y, Liu R, Wu Y, Hui Wang W (2020) Fairness of classification using users' social relationships in online peer-to-peer lending. In: Companion proceedings of the web conference 2020. WWW '20, pp. 733–742. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3366424.3383557
Macskassy SA, Provost F (2007) Classification in networked data: a toolkit and a univariate case study. J Mach Learn Res 8:935–983
Marinho LB, Preisach C, Schmidt-Thieme L et al (2009) Relational classification for personalized tag recommendation. ECML PKDD Discov Chall 2009(DC09):7
McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27(1):415–444. https://doi.org/10.1146/annurev.soc.27.1.415
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2019) A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635
Neville J, Jensen D (2000) Iterative classification in relational data. In: Proceedings of the AAAI-2000 workshop on learning statistical models from relational data, pp 13–20
Newman ME (2001) The structure of scientific collaboration networks. Proc Natl Acad Sci 98(2):404–409
Peel L (2017) Graph-based semi-supervised learning for relational networks. In: Proceedings of the 2017 SIAM international conference on data mining, pp 435–443 (2017). SIAM. http://hdl.handle.net/2078.1/182929
Peel L, Delvenne J-C, Lambiotte R (2018) Multiscale mixing patterns in networks. Proc Natl Acad Sci 115(16):4057–4062
Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 469–481
Rocha LEC, Liljeros F, Holme P (2010) Information dynamics shape the sexual networks of Internet-mediated prostitution. Proc Natl Acad Sci 107(13):5706–5711
Rozemberczki B, Allen C, Sarkar R (2019) Multi-scale attributed node embedding (2019). arXiv:1909.13021
Sen P, Namata G, Bilgic M, Getoor L, Galligher B, Eliassi-Rad T (2008) Collective classification in network data. AI Mag 29(3):93–106
Traud AL, Mucha PJ, Porter MA (2012) Social structure of facebook networks. Physica A 391(16):4165–4180
Verma S, Rubin J (2018) Fairness definitions explained. In: 2018 IEEE/ACM international workshop on software fairness (FairWare), pp 1–7. IEEE
Wagner C Politicians on Wikipedia and DBpedia (Version: 1.0.0) (2017) https://doi.org/10.7802/1515. GESIS - Leibniz-Institute for the Social Sciences
Wagner C, Graells-Garrido E, Garcia D, Menczer F (2016) Women through the glass ceiling: gender asymmetries in wikipedia. EPJ Data Sci. 5(5). https://doi.org/10.1140/epjds/s13688-016-0066-4
Wang DJ, Shi X, McFarland DA, Leskovec J (2012) Measurement error in network data: a re-classification. Soc Netw 34(4):396–409
Yang J, Ribeiro B, Neville J (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, pp 708–711. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15696

Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness beyond disparate treatment and disparate impact: learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web, pp 1171–1180 (2017)
Zeno G, Neville J (2016) Investigating the impact of graph structure and attribute correlation on collective classification performance
Zhang Y, Ramesh A (2020) Learning fairness-aware relational structures. ECAI (2020). arXiv preprint arXiv:2002.09471
Zheleva E, Getoor L (2009) To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In: Proceedings of the 18th international conference on world wide web, pp 531–540

Publisher's Note