Applied Network Science

RESEARCH

Simulating systematic bias in attributed social networks and its effect on rankings of minority nodes

Leonie Neuhäuser[1*†], Felix I. Stamm[1†], Florian Lemmerich[2], Michael T. Schaub[1] and Markus Strohmaier[1,3]

*Correspondence:
neuhaeuser@cs.rwth-aachen.de
†Leonie Neuhäuser and Felix I.Stamm are contributed equally to this work
[1] RWTH Aachen University, Aachen, Germany
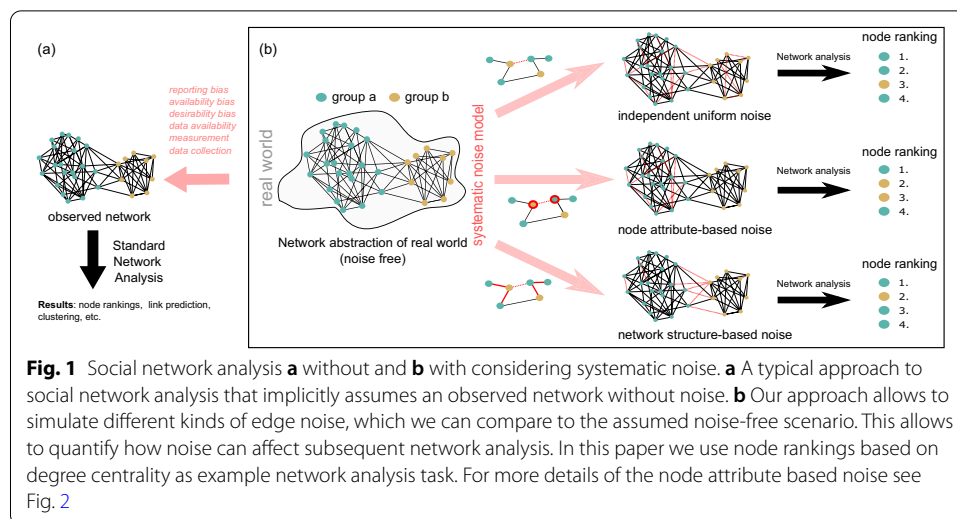Full list of author information is available at the end of the article

Abstract

Network analysis provides powerful tools to learn about a variety of social systems. However, most analyses implicitly assume that the considered relational data is error-free, and reliable and accurately reflects the system to be analysed. Especially if the network consists of multiple groups (e.g., genders, races), this assumption conflicts with a range of systematic biases, measurement errors and other inaccuracies that are well documented in the literature. To investigate the effects of such errors we introduce a framework for simulating systematic bias in attributed networks. Our framework enables us to model erroneous edge observations that are driven by external node attributes or errors arising from the (hidden) network structure itself. We exemplify how systematic inaccuracies distort conclusions drawn from network analyses on the task of minority representations in degree-based rankings. By analysing synthetic and real networks with varying homophily levels and group sizes, we find that the effect of introducing systematic edge errors depends on both the type of edge error and the level of homophily in the system: in heterophilic networks, minority representations in rankings are very sensitive to the type of systematic edge error. In contrast, in homophilic networks we find that minorities are at a disadvantage regardless of the type of error present. We thus conclude that the implications of systematic bias in edge data depend on an interplay between network topology and type of systematic error. This emphasises the need for an error model framework as developed here, which provides a first step towards studying the effects of systematic edge-uncertainty for various network analysis tasks.

Keywords: Edge uncertainty, Rankings, Attributed networks, Bias, Social networks

Introduction

Networks are a simple but powerful model for the complex interplay of elements in systems from various areas, including natural, social and technical systems (Strogatz 2001; Easley and Kleinberg 2010; Newman 2018a).

Computational network analysis typically processes observed networks as if these networks were free of uncertainties and accurately represent the system to be analysed. While researchers have investigated how network analysis tasks are affected

Fig. 1 Social network analysis **a** without and **b** with considering systematic noise. **a** A typical approach to social network analysis that implicitly assumes an observed network without noise. **b** Our approach allows to simulate different kinds of edge noise, which we can compare to the assumed noise-free scenario. This allows to quantify how noise can affect subsequent network analysis. In this paper we use node rankings based on degree centrality as example network analysis task. For more details of the node attribute based noise see Fig. 2

by inaccurate observations of the node set (Wagner et al. 2017; Kossinets 2006; Borgatti et al. 2006) and examined the consequences of random erroneous observation of edges (Borgatti et al. 2006; Frantz et al. 2009; Lee et al. 2006; Murai and Yoshida 2019; Wang et al. 2012; Almquist 2012), non uniform and systematic erroneous edge observations (Adiga and Vullikanti 2013) have been largely neglected. This seems in contrast with the large body of literature on various reporting biases, data collection biases, and other systematic errors which arise when data on social systems is collected (Sen et al. 2019; Holland and Leinhardt 1973; Feld and Carter 2002; Brewer 2000; Marsden 1990, 2003). In this work we thus provide a model to study the impact of systematically missing edges on subsequent network analysis tasks.

Whether collected via surveys or through automated processes such as crawling, the vast majority of social networks describe connections that are reported or initiated by people and are thus unlikely to be completely bias free.

This presence of biases in the observation or reporting of network edges can lead to erroneous conclusions drawn from subsequent analysis (see Fig. 1), which becomes particularly problematic if results of the network analysis are used to drive real-world decisions, e.g. rankings of people in recruiting, marketing or other contexts. In social networks, edge errors may specifically depend on the presence of different social groups, which can be modeled as node labels in attributed networks (Newman and Clauset 2016; Yang et al. 2013; Wagner et al. 2017). Understanding how systematic edge noise may alter the results of network analyses is thus important from at least two perspectives: (1) from a scientific point of view we need to reach robust conclusions about the (social) systems we study; (2) from a societal point of view we want to ensure, e.g., that our network analysis does not discriminate against specific groups. Indeed, previous studies have shown that minorities can be disadvantaged in rankings (Karimi et al. 2018) due to network effects.

Research question We thus ask: how can systematic edge errors in attributed networks be simulated, what are its effects on subsequent network analysis, and how can such effects be assessed?

To illustrate these effects we focus on the question of minority representations in terms of degree-based node rankings, as a simple, but important prototypical network centrality analysis task. More technically, we focus on a sensitivity analysis of degree centrality for various (systematic) edge error models. Such degree-centrality based rankings of the nodes are frequently used to access the relevance and importance of entities in networks.

Contributions and findings

(1) We introduce a general model for simulating systematic edge errors in attributed networks. The model assumes the presence of an unobserved true network, whose edge set is altered by various error conditions, resulting in a distorted observed network. We differentiate between (1) node attribute-based errors which are a function of the labels associated to the nodes an edge is incident on (edges between groups or attached to a specific group have a higher probability of going unobserved) and (2) network structure-based errors which are independent of the node meta data and derive from the structure of the latent social network (edges between lower degree nodes or edges between nodes with few neighbors in common have a higher probability of going unobserved). To illustrate how this error model can be used to assess the potential impact of systematic edge errors on network analysis, we focus on the representation of minorities in centrality-based rankings. We consider this a first step to study the effects of systematic noise, as simulated here by our error model, as centrality measures are frequently used to access the relevance of entities in social networks. Therefore, an altered representation of minorites in centrality-based rankings, caused by systematic noise in data, serves well as an indicator for misleading real-world conclusions.

(2) We apply our method to synthetic and empirical networks with binary node labels. We find that systematic noise affects minority representation in rankings while uniform noise does not. The effect depends on both the type of edge error and the group-label assortativity in the network: In heterophilic networks, minority ranking positions are very sensitive to the type of systematic edge error. In contrast, in homophilic networks the minority's disadvantage is independent of the type of error present. While our modeling approach is general and can be easily extended to networks with multiple groups represented through nominal attributes, or to continuous node-covariates, we restrict our analysis here to the case of binary attributes representing two groups as an important but comparably simple setup.

Implications With our model we enable researchers to simulate errors due to different data collection and reporting biases. As our numerical experiments show, the effects of such systematic errors on subsequent network analyses can vary depending on the network structure. This emphasizes the utility of a framework such as ours, which enables researchers to simulate systematic error hypothesis and examine the effects on particular data sets. While it may seem desirable to not rely on

some hypothesis about the errors present in the data, we remark that unless specific assumptions are made about the generative process creating the observed network (Peixoto 2018; Newman 2018b; Young et al. 2020), it is generally impossible to extract (and correct) errors in network data based on a single network observation.

Background and related work

Networks with edge uncertainty Most empirical network analyses assume that the collected data accurately represent the underlying social system, even though such network data are known in general to be inherently noisy due to various social, cognitive and technical biases and errors. Especially if these network representations are used in subsequent data analysis tasks with real-world implications—consider, e.g., recommendation systems, importance rankings via centrality, etc.—neglecting such errors can have potentially detrimental consequences. Focusing on the issue of edge-errors, the literature related to this work may be divided into two strands.

First, for *unlabeled* networks a number of studies, e.g., Borgatti et al. (2006), Frantz et al. (2009), Lee et al. (2006), Murai and Yoshida (2019), Wang et al. (2012), Almquist (2012) have investigated the effect of random (non systematic) errors on certain analysis tasks. Some recent work has focused on inferring network errors, which is however only possible if the exact form of the network model and the error (unbiased, random) is assumed (Peixoto 2018; Newman 2018b; Young et al. 2020; Guimerá and Sales-Pardo 2009). Similar to these studies we consider that there exists an underlying network, from which we can only obtain a noisy observation. While there are some initial attempts at exploring non-uniform edge error models (Adiga and Vullikanti 2013), we provide here a more complete characterization of errors and associated interpretations. In addition, our error model allows for attribute dependent noise, which appears to have not been considered before in the literature. We argue that systematic errors that are dependent on node attributes such as group membership are common and that attributed networks provide a unique opportunity to study the impact of these systematic uncertainties.

Second, noise in network data has been considered in the context of the generating process of the network. For instance, Moore et al. study the addition or deletion of edges during the growth of networks with uniform or preferential attachment (Moore et al. 2006). Considering downstream analysis, recent research has started to investigate the effects of sampling networks from random models on node centrality analysis (Dasaratha 2020; Avella-Medina et al. 2020), thus providing a possible notion of confidence of these metrics. In contrast to our work the edge-uncertainty derives purely from the network generating process and no systematic edges errors are considered.

Note that there has also been a lot of work related to link predictability. Link prediction considers a setup in which the set of observed links in a network is used to estimate the likelihood that an unobserved link will start to exist in the future (Liben-Nowell and Kleinberg 2007; Clauset et al. 2008; Lü et al. 2015). This is different to the simulation of systematic edge errors where we are not interested in predicting the temporal evolution of a network, or to reconstruct a partially observed network, Instead we aim to quantify the implications of erroneous edges on subsequent network analysis tasks.

Centrality in attributed and sampled networks The centrality of nodes is frequently used to access the relevance and importance of entities in social networks. For attributed

social networks, the relative positioning of a group is a major determinant of how much access this group has to resources and information (Calvó-Armengol and Jackson 2004; Nilizadeh et al. 2016; Hannák et al. 2017). In this context, Karimi et al. show that a minority group in a network can become strongly (dis-)advantaged in terms of the number of network connections it can accumulate (measured in terms of node degrees), based merely on a combination of homophily and a preferential attachment mechanism (Karimi et al. 2018).

Lerman et al. (2016) demonstrate that the local visibility of nodes can lead to a so called "majority illusion", a situation where individuals systematically overestimating the prevalence of a group in the network. Wagner et al. (2017) observed significant differences between sampling techniques when trying capture the centrality of nodes and concluded a potential impairment of group visibility. In contrast to our work, both studies focus on uncertainty related to having access only to parts of the (node set of the) network, rather than the impact of systematic edge errors in the network. Other relevant studies on the effect of node-subsampling in networks include (Kossinets 2006; Borgatti et al. 2006; Smith and Duggan 2013).

Social biases, cognitive biases and systematic errors There are multiple biases, associated with beliefs, decision-making, memory and social desirability, which influence reporting about social connections (Sen et al. 2019; Holland and Leinhardt 1973; Feld and Carter 2002; Brewer 2000; Marsden 1990, 2003). In the field of self-reported ego networks, research on measurement bias in respondent reports of personal networks has shown that the reported social network is usually much smaller than the actual number of relationships and that it can be influenced by the interviewer (Feld and Carter 2002; van Tilburg 1998; Marsden 2003). A particularly timely example for a systemically biased reporting of contacts is the case of manual contact tracing in the Covid19 pandemic. Having a high number of contacts is socially condemned, consequently individuals with a lot of interactions to others might systematically under report their number of contacts to health departments which perform contact tracing. On the contrary, individuals with a low number of contacts are more likely to report these correctly. For mitigation of outbreaks and effective quarantining, it is vital to capture these networks correctly which is a key challenge in current research (Braithwaite et al. 2020).

Additionally, the data collection process itself can be structurally focusing on a specific group of actors and therefore lead to a biased data availability driven by peer effects (Wiese et al. 2015; Yang et al. 2017). For example, González-Bailón et al. have found that the Twitter search API over-represents the more central users and does not offer an accurate picture of peripheral activity (González-Baión 2014). Similarly in Wikipedia clickstream data (Rodi et al. 2017) only transitions exceeding 10 requests are present. We would argue both of these effect are poorly modelled by random uniform missing edges.

Systematic error model

The main contribution of our work is the introduction of a model for simulating systematic edge noise, with which we explore different (hypothetical) biases in data collection and interpretation. For instance, due to a social desirability bias one social group might over- report or under report connections with another group. Alternatively, due

to cognitive availability biases, people may remember interactions with closely related others more (Bell et al. 2007; Smieszek et al. 2012; Calloway et al. 1993). They might also tend to report more connections to nodes with a high visibility in their social network. We model such social, cognitive and technical biases by systematically adding or removing edges to a network.

We distinguish between two types of critical information for systematic edge errors: (1) node attribute data, which is extrinsic to the network, and (2) local network structure properties, which is information that is intrinsic to the network.

In the following, we describe the different ways in which we model systematic noise.

Mathematical model

In this work we consider networks represented by simple undirected graphs $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with a vertex set $\mathcal{V} = \{1, \ldots, N\}$ and an edge-set $\mathcal{E} = \{\{i, j\} : i, j \in \mathcal{V}\}$. We can represent such a graph algebraically by an adjacency matrix $A \in \mathbb{R}^{N \times N}$ with entries $A_{ij} = 1$ if $\{i, j\} \in \mathcal{E}$ and $A_{ij} = 0$ else. As the graph is undirected A is a symmetric matrix.

When observing a network, we will generally obtain an (observed) adjacency matrix A^o, which is usually an error prone observation of an underlying latent network A. If we observe an edge in A^o we may have (i) a true positive observation, meaning that the edge was indeed present in the latent network or (ii) a false positive observation, where we observe an edge, when there should have been none. We model these cases by defining the probability of an observed true positive edge A^o_{ij} as $P_{ij} = \mathbb{P}(A^o_{ij} = 1 | A_{ij} = 1)$.

The true positive probability P_{ij} can be interpreted as the probability for a latent edge to be retained in the observed network, and we thus alternatively refer to it as retain probability. Analogously, if we do not observe an edge, we may have (i) a true negative observation, where we observe no edge and there was indeed none, or (ii) a false negative observation, i.e., we may not observe an edge when there is one in the latent network.

For edge errors described by statistically independent $\{0, 1\}$ Bernoulli random variables, we can encode the marginal probabilities of the observed true positives in a symmetric $n \times n$ matrix P. We denote a sample from this distribution over matrices by S_P. The adjacency matrix of the observed network may then be expressed as compactly as:

$$A^o = A \odot S_P, \tag{1}$$

where $\odot$ denotes the element-wise (Hadamard) product. In other words, each edge (i, j) present in the latent network is retained with probability P_{ij}. This means that once the probabilities are fixed the edges are dropped independently.

In this work we focus on the particular case of systematic missing edges modelled by being drawn independently. Generalizations such as edge addition or edges being dropped in a correlated way are also possible with minor adaptation. For instance, we can describe the observed adjacency matrix of a model with both independent edge additions and deletions analogously to above as:

$$A^o = A \odot S_P + A^c \odot S_Q, \tag{2}$$

where A^c denotes the adjacency matrix of the complement graph and S_Q a $\{0, 1\}$ sample drawn according the matrix of false positives.

In the following subsections we discuss in detail how (systematic) edge-errors can shape the error probability matrix *P*.

Baseline model

Removing edges uniformly at random has previously been studied (Borgatti et al. 2006; Kossinets 2006) and serves as our baseline model. In the baseline model, errors are independent and identical for each edge. This kind of measurement error may be relevant in settings in which all observed noise is technical, e.g., if we measure social contacts via some proximity sensors, which may independently fail to detect a contact with a certain probability (DuBois et al. 2012; Martin et al. 2006). Moreover, the baseline model enables us to better gauge and compare the effects of systematic errors.

In the baseline model, every edge that is present in the latent network is retained with constant probability *p*. This leads to a (constant) noise probability matrix:

$$P_{ij} = p \tag{3}$$

Note that low values of *p* correspond to a large impact on the observed network.

Node attribute-based noise

In the context of social networks, systematic errors are often driven by node attributes, e.g. membership in certain groups (based on ethnicity, gender, etc.). For a network split into multiple groups, we may for instance observe certain forms of in-group/out-group bias or social desirability bias, where a connection to a particular group may be less or more often reported (e.g. in-group favoritism (Everett et al. 2015)). We encode node attributes in the form of continuous or discrete node labels l_i. For instance, the label l_i may correspond to the height of a person or its membership in a particular social group. We will focus on categorical node attributes representing two groups in the following discussion, although our models can be applied to other discrete or continuous node attribute data as well (cf. Section "Discussion" for additional discussion).

Given that each node *i* in the network has an associated categorical label $l_i = 1, \ldots, \ell$, we now assume that noise is solely a function of the pair of labels involved in each edge. We thus write
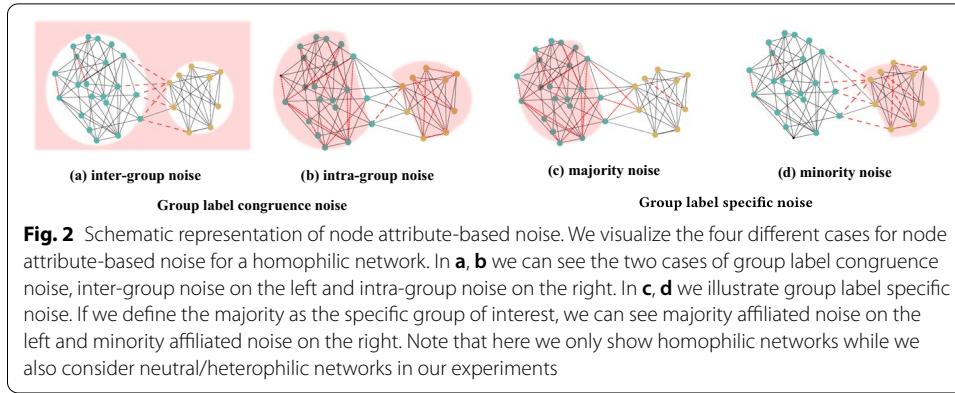
$$P_{ij} = \Omega_{l_i l_j} \tag{4}$$

with Ω_{mn} encoding the probability that an edge between nodes with label *m* and *n* is retained.

Here we are particularly interested in two scenarios, which we call *group label congruence* and *group label specific* noise. In the following we investigate how the matrix Ω can be chosen to account for these different kinds of noise.

Group label congruence noise (intra/inter group) This setting describes classical intergroup (between groups) vs. intra-group (within group) effects. The characteristic that determines the noise level is whether the nodes incident to an edge have the same (intra) or distinct labels (inter), i.e., belong to the same group or not.

In this case, the matrix Ω will be structured as:

Fig. 2 Schematic representation of node attribute-based noise. We visualize the four different cases for node attribute-based noise for a homophilic network. In **a**, **b** we can see the two cases of group label congruence noise, inter-group noise on the left and intra-group noise on the right. In **c**, **d** we illustrate group label specific noise. If we define the majority as the specific group of interest, we can see majority affiliated noise on the left and minority affiliated noise on the right. Note that here we only show homophilic networks while we also consider neutral/heterophilic networks in our experiments

Table 1 Specific parametrization of label-based error matrix Ω used for numerical experiments

Noise type	Intra	Inter	Majority	Minority
Ω	$\begin{pmatrix} \rho & 0.9 \\ 0.9 & \rho \end{pmatrix}$	$\begin{pmatrix} 0.9 & \rho \\ \rho & 0.9 \end{pmatrix}$	$\begin{pmatrix} \rho & \rho \\ \rho & 0.9 \end{pmatrix}$	$\begin{pmatrix} 0.9 & \rho \\ \rho & \rho \end{pmatrix}$

We model four different noise types. We use the retain parameter ρ to control the noise probability of edges. For inter-group noise, e.g., we retain an edge that connects the two classes with probability ρ, while we retain edges within a class with 90% probability

$$\Omega_{mn} = \begin{cases} p_{\text{intra}} & m = n, \\ p_{\text{inter}} & \text{otherwise.} \end{cases} \tag{5}$$

where p_{intra} describes the probability of retaining an edge if two nodes are in the same group (have the same node label), and p_{inter} describes the probability of retaining an edge if two nodes are in different groups. We can distinguish between $p_{\text{intra}} < p_{\text{inter}}$ in which intra-group connections are less likely to be retained, and $p_{\text{intra}} > p_{\text{inter}}$ in which the same is true for inter-group edges.

Group label specific noise (majority/minority) This setting models a scenario where edges attached to nodes of a particular group and thus to one particular label l are significantly more or less affected by noise. For instance, edges connecting to a minority group may be under-reported.

The matrix Ω will be structured as:

$$\Omega_{mn} = \begin{cases} p_l & m = l \vee n = l, \\ p_{\text{standard}} & \text{otherwise.} \end{cases} \tag{6}$$

This means, edges that are on either end connected to nodes of that particular class l are retained with probability p_l while all other edges are retained with a probability p_{standard}.

Figure 2 provides illustrations for group label congruence and group label specific edge noise, for the common scenario in which the network consist of two groups: a majority and a minority. This setup will be the main focus for our numerical experiments.

Parameter specification of Ω in numerical experiments

In our simulations in section "Experimental results", we explore four different node attribute-based noise types (1) inter-group noise, (2) intra-group noise, (3) minority specific noise, and (4) majority specific noise. The exact parametrization of the introduced label-based error matrix Ω which captures these four settings can be found in Table 1.

Network structure-based noise

Certain systematic errors are intrinsically coupled to the structure of the latent network itself. For instance, humans tend to recall information better that is more readily available in their memory, e.g., because the information is more recent or associated to an event of higher (perceived) importance (Bell et al. 2007; Smieszek et al. 2012; Calloway et al. 1993).

To model the aforementioned biases and systematic errors, we consider the structured error matrix P, such that the probability P_{ij} of an edge $\{i, j\}$ to be retained is a function $f(\{i, j\}, \mathcal{G})$ of the position of the specific edge in the network.

We argue that local properties are of particular interest for modelling social and cognitive biases as most global information is typically not available to individuals. Consequently, the specific functions $f(\cdot)$ considered below depend only on information that is locally available to the two nodes adjacent to the edge (or can be computed based on local information).

We propose two forms of systematic noise (Jaccard and Centrality noise) which could be used to model these effects as a function of network topology.

Jaccard noise (availability bias) When working with data from self-reported, ego centered networks, frequent interactions will be recalled more likely than rare interactions. If two nodes share a lot of neighbours in a social network, i.e., their social circles overlap significantly, this can indicate more frequent interactions between the two nodes. Specifically, the probability of an edge i, j to be retained or falsely included is a function of the overlap of the neighborhoods between the two endpoints i, j. In line with the idea of availability bias, if the two endpoints have a very similar neighborhood, then the probability of retention is high. We model the availability bias by retaining an edge with a probability that is given by the Jaccard similarity of the (augmented) neighborhoods of the two endpoints of the edge:

$$P_{ij} = \left(\frac{|\mathcal{N}_i \cap \mathcal{N}_j|}{|\mathcal{N}_i \cup \mathcal{N}_j|} \right)^\alpha =: \mathcal{J}^\alpha \tag{7}$$

To avoid removing all edges in the case of almost bipartite networks, here we consider each node i to be part of its own neighborhood as well, i.e., $\mathcal{N}_i = \{i\} \cup \{k : \{k, i\} \in \mathcal{E}\}$. α is a scaling parameter that we can choose to modulate how aggressive edges will be dropped. For $\alpha = 0$ all edges will be retained, for larger α we more aggressively remove edges whose endpoints have a low product of degrees (relative to other node pairs in the graph). We call the noise of type Eq. (7) *Jaccard noise*. In addition we define the *inverse Jaccard noise* via $P_{ij} = (1 - \mathcal{J})^\alpha$, which models a situation in which an edge between two similar nodes is more likely to be dropped. This is captured by the opposite of the relative overlap of the neighborhoods, $(1 - \mathcal{J})$, which is then again scaled by a parameter α which controls the noise strength.

Centrality noise (desirability bias) Alternatively, interactions with highly central nodes in the network might be perceived as more important, and interactions to these individuals would accordingly be remembered better when reporting ties. Social desirability may amplify this even further, leading to a higher probability of observing interactions with central nodes at the core of the network.

We thus model a setting in which the probability for edges to be omitted will be larger in the periphery than in the core of the network, where we define core/periphery based on the degrees of the nodes. Specifically let $d_i = \sum_j A_{ij}$ denote the degree of node i. As the degree parameters are very heterogeneous, we use logarithmic degrees to define $d_{ij} = \log(d_i) + \log(d_j)$ and then use a normalization to define the matrix of retain probabilities via:

$$P_{ij} = \left(\frac{d_{ij} - \min_{ij \in \mathcal{E}} d_{ij}}{\max_{ij \in \mathcal{E}} d_{ij} - \min_{ij \in \mathcal{E}} d_{ij}} \right)^{\alpha} =: \mathcal{D}_{ij}^{\alpha}. \tag{8}$$

The scaling parameter α has the same function here as for Jaccard noise. We call (8) *centrality-based noise*. As before, we define the *inverse centrality-based noise* via $P_{ij} = \left(1 - \mathcal{D}_{ij} \right)^{\alpha}$.

Note that the two different types of bias in our model differ in the type of information they are based on, but not necessarily in their effect on the network data. This can imply that a node attribute-based bias and a structure-based bias have the same effect, even if two different biases are modelled. Such an effect arises, e.g., if the node labels are aligned with the respective structural properties.
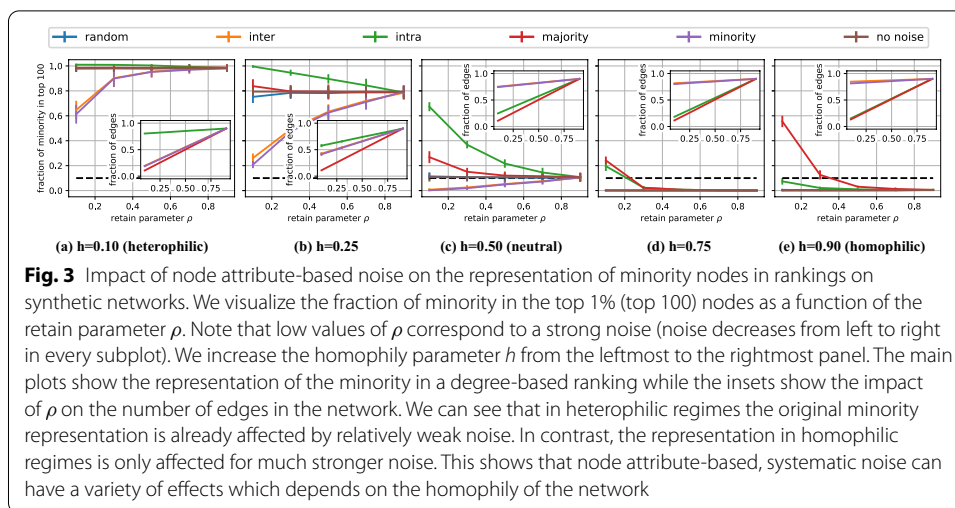
Experimental results

In our experiments, we focus on attributed networks with binary labels which represent two groups with unequal group sizes. We investigate how the systematic errors we introduced in section "Systematic error model" affect the representation of a minority in degree-based node rankings in networks with different degree of homophily. Our overall goal here is to elucidate possible effects of interactions of specific systematic bias with the network topology. Specifically, we consider the fraction of minority nodes in the top k nodes of the ranking. We perform experiments on both synthetic and real-world networks.

Experiments on synthetic networks

Experimental setup

We first investigate the effects of the different noise types on synthetic networks in which we can control the homophily of the two groups. We use an adaptation of the Barabasi–Albert (BA) model, introduced in Karimi et al. (2018).

The model uses a preferential attachment mechanism where the attachment probability for a new node of label a to attach with a node of label b is additionally skewed by their homophily h_{ab}. Note that we only consider the case of symmetric homophily for binary labels, which means that $h_{00} = h_{11} = h$ and $h_{01} = h_{10} = 1 - h$. Now $h = 0$ indicates an entirely heterophilic network, i.e., nodes in a group a only connect to nodes with group label $b \neq a$. The case of $h = 0.5$ is equivalent to the standard BA model

Fig. 3 Impact of node attribute-based noise on the representation of minority nodes in rankings on synthetic networks. We visualize the fraction of minority in the top 1% (top 100) nodes as a function of the retain parameter ρ. Note that low values of ρ correspond to a strong noise (noise decreases from left to right in every subplot). We increase the homophily parameter h from the leftmost to the rightmost panel. The main plots show the representation of the minority in a degree-based ranking while the insets show the impact of ρ on the number of edges in the network. We can see that in heterophilic regimes the original minority representation is already affected by relatively weak noise. In contrast, the representation in homophilic regimes is only affected for much stronger noise. This shows that node attribute-based, systematic noise can have a variety of effects which depends on the homophily of the network

where node labels play no role in the formation of the network and $h = 1.0$ is a completely homophilic network where nodes of type l only connect to nodes with similar label l.
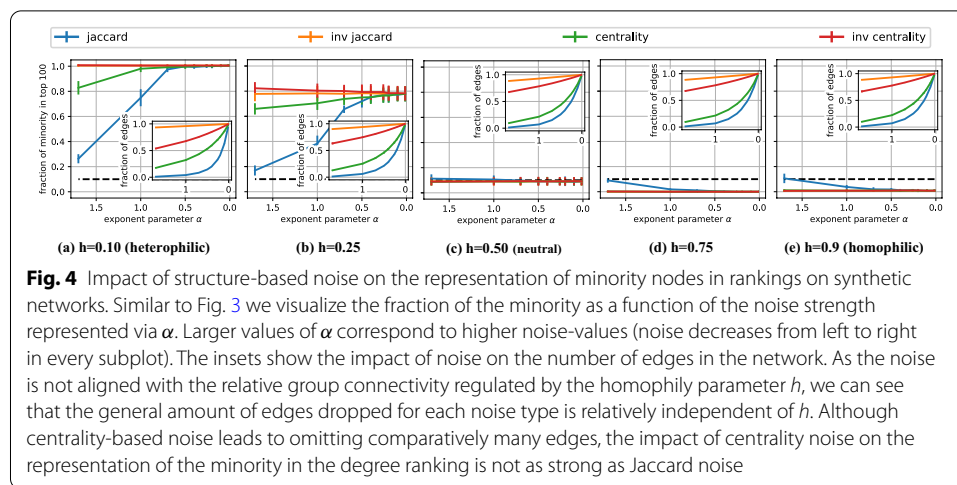
In our numerical experiments, we focus on binary minority/majority labels with a fixed minority size of 10% of the nodes. In Section A in the "Appendix" we assess the robustness of our results with respect to this choice. We generate networks with $N = 10,000$ nodes by adding $m = 5$ edges in the modified preferential attachment scheme (Karimi et al. 2018). The different noise types are then applied to each network. We include both simulations of independent uniform noise and calculations without edge-errors as baselines for the systematic error scenarios.

We explore four different node attribute-based noise types (1) inter-group noise, (2) intra-group noise, (3) minority specific noise, and (4) majority specific noise. We use a retain parameter ρ to control the noise probability of edges. For inter-group noise, we retain an edge that connects the two classes with probability ρ, while we retain edges within a class with 90% probability, vice versa for intra-group noise. For minority/majority specific noise, the retain probability of an edge connecting with at least one node of the specific group given by ρ, and only if both nodes belong to another group, the edge is retained with a probability of 90%. Mathematically, this is modeled through different structures of a label-based error matrix (see section "Node attribute-based noise" for further details).

Additionally we investigate network structure-based noise types: Jaccard and inverse Jaccard noise, as well as centrality-based and inverse centrality-based noise. The strength of the noise is controlled by an exponent parameter α (see Eqs. 7, 8 in section "Network structure-based noise"). For low values of α edges are retained more likely while for large α fewer and fewer edges are retained. This in contrast with ρ where low values correspond to strong noise. For each parameter setting, we generate 10 networks.

Results for synthetic networks

In Figs. 3 and 4 we display the simulation results of the effects of attribute-based and structure-based noise. In all simulations, we already observe interesting effects for

(a) h=0.10 (heterophilic)　　**(b) h=0.25**　　**(c) h=0.50 (neutral)**　　**(d) h=0.75**　　**(e) h=0.9 (homophilic)**

Fig. 4 Impact of structure-based noise on the representation of minority nodes in rankings on synthetic networks. Similar to Fig. 3 we visualize the fraction of the minority as a function of the noise strength represented via α. Larger values of α correspond to higher noise-values (noise decreases from left to right in every subplot). The insets show the impact of noise on the number of edges in the network. As the noise is not aligned with the relative group connectivity regulated by the homophily parameter h, we can see that the general amount of edges dropped for each noise type is relatively independent of h. Although centrality-based noise leads to omitting comparatively many edges, the impact of centrality noise on the representation of the minority in the degree ranking is not as strong as Jaccard noise

the noise-free setting, which are caused by the homophily of the network independent of noise and have been discussed by Karimi et al. (2018): In a homophilic regime (low h, see Figs. 3, 4d, e), we find a general under-representation of the minority. This effect is due to the relative size of the majority which prefers interactions among itself. In contrast, in a heterophilic regime ($h \leq 0.25$, see Figs. 3, 4a, b), the number of inter-group edges is high, resulting in a strong over-representation of the minority in the degree ranking in the noise free case, relative to the actual group-sizes (Karimi et al. 2018). Throughout our analysis we interpret this over/under representation as the baseline which is indicated through the "no-noise" line and will refer to it as *original* over/under representation in contrast to the *noise-based* over/under representation which is caused by systematic errors.

Node attribute-based noise

The results for the effects of attribute-based noise in Fig. 3 show contrasting effects for homophilic and heterophilic networks for different types of noise. We find that in the case of a heterophilic network (Fig. 3a, b), systematic noise can have large and opposing effects compared to the error-free and random baselines. Noise types that remove preferably inter-group edges reduce the original representational advantage of the minority. Vice versa, dropping intra-class edges further increases the original over-representation of the minority, as the significance of the minority in the degree-based ranking is not due to intra-class edges, but due to a large number of inter-class edges.

In contrast, we find that the under-representation of the minority in homophilic networks is stable with regard to different noise types and over a broad range of retain probabilities (Fig. 3d, e). The minority can only gain some noise-based significance in the ranking for very low edge retain probabilities in the majority noise and intra-group noise scenarios. In these two scenarios the majority will proportionally loose many more edges than the minority.

If the minority and majority nodes are distributed randomly in the network (corresponding to no homophily, $h = 0.5$) (see Fig. 3c), the relative representation of the minority/majority in the noise-free degree-based ranking is only dependent on the relative group sizes (10% minority). Accordingly, we find opposing effects of different noise types with respect to this baseline for $h = 0.5$: intra-group and majority noise lead to a noise-based over representation of the minority, inter-group noise and minority noise lead to a noise-based under representation.

Network structure-based noise

In our second set of experiments, shown in Fig. 4, we focus on network structure-based noise.

For the extended BA model, Jaccard noise mostly penalizes edges that connect low degree with high degree nodes. In a heterophilic setup (Fig. 4a, b), most of these edges are inter-group edges and Jaccard noise thus behaves similarly to inter-group noise in the case of attribute-based noise. In contrast, in homophilic regimes (Fig. 4d, e), most of the edges that connect low degree with high degree nodes are intra-group edges and the Jaccard-noise is thus similar to the case of intra-group noise. Depending on the homophily parameter, we therefore either penalise the minority or majority, as this type of noise is not dependant on the labels as in the case of attribute-based noise, but derived from the network structure itself.

Inverse Jaccard noise mostly drops edges within strong communities i.e. within densely connected subsets of nodes which are not present in the extended BA model. This leads to an omission of very few edges and almost no shift in the original representation.

This highlights an important difference of our two types of noise: attribute-based noise is derived from groups which are based on the meta data of the nodes, whereas the structure-based noise (Jaccard noise in particular) is dependant on the community structure of the network. As community structure of the network and the metadata of different groups do not necessarily have to align (Peel et al. 2017), this can lead to different effects depending on how the structure and meta-data correspond in the particular network of interest.

For centrality noise, the effects we observe are not as strong as for Jaccard noise. Centrality noise mostly drops edges connecting low to low/mid degree nodes. This leaves the high degree nodes represented in rankings mostly untouched. Thus, dropping comparatively many edges has no effect in homophilic regimes where the majority has originally an advantage. We remark that the above finding does not imply that centrality-based noise has no effect on the network. Rather, for centrality-based noise the periphery-periphery edges are mostly affected, and these do not show up in the ranking measure we consider here. In heterophilic regimes a substantial amount of the in-edges of top nodes in either group is supplied by nodes of the other group. Since minority nodes have a higher *average* degree this implies that links from the minority to high degree nodes in the majority are more likely being retained. This leads to a noise-based reduction of the original over-representation of minority nodes. This effect is not reversed for inverse centrality noise. Due to the specific structure of the BA model there are comparably less edges that connect two very high-degree nodes

Table 2 Overview of the empirical datasets

Name	Topic	Nodes	Edges	Q_{mod}	f	A
brazil (Rocha et al. 2010)	Sexworker	15k	38k	−0.500	39%	−1.00
pok (Holme et al. 2004)	Online dating	25k	25k	−0.349	43%	−0.84
github (Karimi 2019)	Followers	119k	248k	0.004	5%	−0.03
dblp (Karimi et al. 2016)	Coauthorship	185k	619k	0.027	21%	−0.15
aps (Karimi 2019)	Citation	1k	3k	0.346	32%	0.74

We use Q_{mod}/A to denote the modularity/assortativity of the partition based on minority and majority label. The modularity is not obtained from an optimisation. Both metrics may be interpreted as a proxy for the homophily in the dataset. The fraction of the minority in the dataset is denoted by *f*. Our datasets cover a wide range of sizes, minority fraction and modularity/assortativity

and thus very few edges have a high probability of getting omitted (cf. the insets of Fig. 4). Thus, the inverse centrality noise leaves the network mostly untouched.

We can conclude from our analysis on synthetic networks that systematic biases can have very different effects on the ranking outcomes depending on the joint effect of degree of homophily and type of systematic error. We find that in heterophilic networks, majority/minority representations in rankings are very sensitive to the type of edge error present. In contrast, in homophilic networks we find that minorities are at a disadvantage regardless of the type of error present.

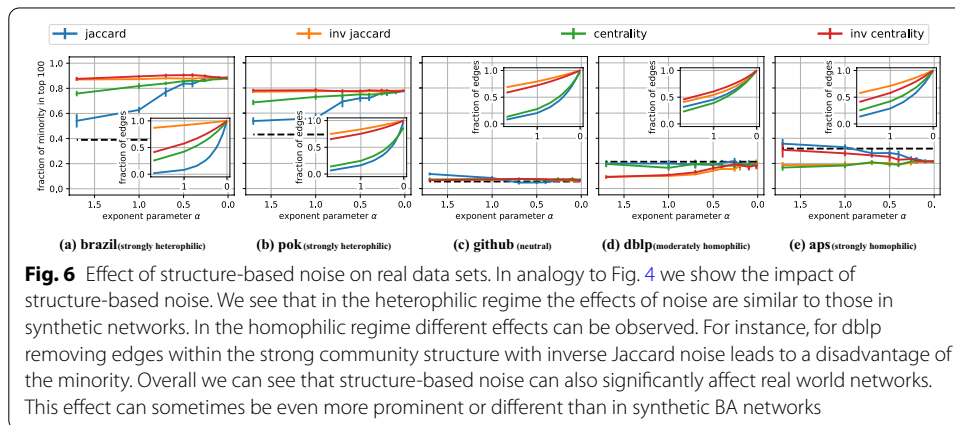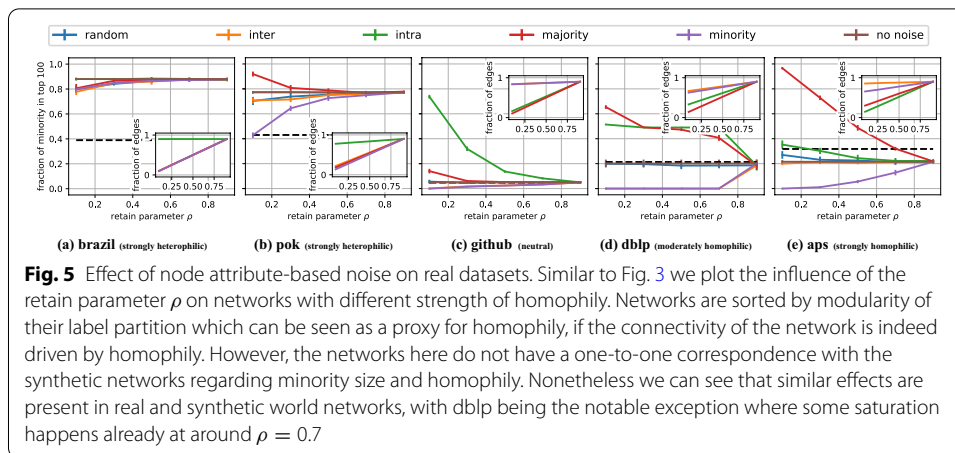## Experiments on real-world networks

In this section, we investigate whether the observations made for synthetic datasets hold with empirically observed networks. Even if these data are most likely not error-free to start with, they provide us with real topologies of complex networks that may not be captured by a synthetic model. For these empirically observed networks, we can also test whether systematic noise based on our model would lead to significant effects. Using real-world networks in this way, we follow standard practice in the literature of assessing the effects of non-systematic errors (Murai and Yoshida 2019).

Real world networks

We assess the impact of noise on a wide range of social networks with available binary attributes, see Table 2 for an overview. We report assortativity and modularity for all networks as a proxy for the homophily parameter. Note again the difference between node attribute- and structure-based groups: In the case of the synthetic networks, the correspondence was controlled by the generation process. This is different in the case of empirical networks. We therefore have to note that the modularity parameter of the label partition is only a proxy for homophily.

For the datasets brazil, github and dblp the labels correspond to gender[1] For the pusokram (POK) network the labels were previously inferred using a maxcut approach, based on the assumption of a predominance of heterosexual relationships. We can

---

[1] We consider gender labels to be binary here due to data availability, being aware of the deficit of not explicitly considering non-binary gender identities.

Fig. 5 Effect of node attribute-based noise on real datasets. Similar to Fig. 3 we plot the influence of the retain parameter ρ on networks with different strength of homophily. Networks are sorted by modularity of their label partition which can be seen as a proxy for homophily, if the connectivity of the network is indeed driven by homophily. However, the networks here do not have a one-to-one correspondence with the synthetic networks regarding minority size and homophily. Nonetheless we can see that similar effects are present in real and synthetic world networks, with dblp being the notable exception where some saturation happens already at around $\rho = 0.7$

Fig. 6 Effect of structure-based noise on real data sets. In analogy to Fig. 4 we show the impact of structure-based noise. We see that in the heterophilic regime the effects of noise are similar to those in synthetic networks. In the homophilic regime different effects can be observed. For instance, for dblp removing edges within the strong community structure with inverse Jaccard noise leads to a disadvantage of the minority. Overall we can see that structure-based noise can also significantly affect real world networks. This effect can sometimes be even more prominent or different than in synthetic BA networks

therefore only label the two groups as majority/minority, as in Karimi et al. (2018). For the APS network, labels reflect scientific sub-fields.

Note that all of the here considered data sets are large networks. As we only consider degree centrality-based rankings in this work, and degree is a local network statistic, the results are largely unaffected by the network size. Nevertheless, we included simulations on networks with smaller node sets in "Appendix" which are coherent with the real world networks presented here. This confirms that the effects of our error simulations also apply to smaller networks. Note that the sensitivity of other networks, e.g., different centrality measures, may have a stronger network-size dependent dependency.

Results on empirical networks

Figures 5 and 6 summarize the effects of applying different noise types to the empirical networks. Overall, we observe similar effects as for the synthetic networks when we consider them in order of their modularity of the label partition. This suggests that key factors for the implications of edge uncertainty on node centrality are already captured well by the comparatively simple modified BA model.

However, there are some notable exceptions: In the dblp co-authorship network, a saturation already sets in for a retain parameter of around $\rho = 0.7$. Beyond that value, there are no more significant effects to be observed when dropping more edges, independent of the noise type. As another main discrepancy, the difference between Jaccard-based and centrality-based noise types with respect to the number of dropped edges is much weaker in empirical to synthetic networks.

This might be due to the fact, that centrality-noise mostly affects low-low degree connections. Such low-low degree connections occur more often in the empirical datasets due to local clustering, which is absent in Barabasi–Albert-based models.

The results of our simulations on real-world networks thus lead to similar results as the analysis of synthetic BA-networks. This suggests that the homophily of the network is responsible for some of the effects of systematic bias on minority representation in degree-based rankings.

Discussion

The main result of our work is that systematic errors can give rise to significant effects on subsequent network analysis. This emphasises that researchers should account for the possibility of such errors more carefully, e.g., by checking certain bias hypotheses using our framework, and including knowledge on systematic edge errors into their analysis, where possible. Interestingly, we observed similar results when investigating the effects of systematic edge errors in real-world and synthetic networks generated by an augmented Barabasi–Albert model. This may suggest that, at least for degree-based rankings, network properties (like clustering) are less important for the representation of minorities in rankings under systematic noise. These aspects should be investigated in future work.

Clearly, our study is only a first step towards better understanding of the effects of systematic edge errors on network analysis. For instance, effects that arise from the addition of edges instead of deletion are left to be investigated in future work. However, our error model provides a platform to account for a wide range of systematic errors that are in need of better understanding, and will thus hopefully trigger further research in this direction. We emphasize that our model is not restricted to binary node attributes, it allows for continuous attributes such as age to simulate e.g. age discrimination.

Moreover, the ideas presented here can be easily extended to directed and/or weighted networks. In this case we could also consider asymmetric biases, e.g. of groups reporting differently on each other (in the context of directed graphs), or edge-weight thresh-holding effects on ties (in the context of weighted graphs). Such problems are for example present in Wikipedia Clickstream data, where the number of occurrences of each pairs of page requests and referers is only counted if exceeding 10 requests (Rodi et al. 2017). Eventually, we may consider correlated edge errors, which can arise, for instance, if nodes selectively do not report a whole set of connections of a particular type, or replicate the reporting behavior of other nodes. More broadly, we may also want to include effects that arise from sampling nodes, or having uncertain or incomplete node (attribute) information.

In our experiments we considered synthetic and empirical networks and removed edges based on different systematic errors. We thus treated the initial networks

effectively as an accurate system representation, even though we can only assert this for the synthetically generated networks. However, the empirical networks used are potentially not error free. Yet, such network can often serve as a (more or less accurate) proxy for real network structure. In this context, it should also be acknowledged that multiple types of biases may lead to the same kind of observation errors. For instance, in a network, whose structure may be due to homophilic interactions between socially alike groups, reporting bias correlated with the group structure (modelled by node attribute-based edge noise) may lead to similar effects on the network than noise arising due to an availability bias correlated with a strong neighborhood overlap (modelled by a form of structure-based noise). While the resulting edge noise may thus be the same, the interpretation of its cause can be very different. This underlines that without additional information it is impossible to infer the reason for a specific edge structure from a single network observation, and we can only use models to investigate potential implications.

In this work we study the impact of systematic edge errors motivated by social biases. Although we provide detailed motivation for the semantics of such edge error in form of certain biases, we cannot ascertain that an empirically observed network has been subject to a particular systematic edge error (without any additional information). This is somewhat reminiscent of the correlation versus causation debate when considering homophily and contagion in observational social network studies, as investigated in Shalizi and Thomas (2011).

Finally, we have focused on degree-based rankings within the scope of this paper. However, this is obviously not the only network analysis task that can be affected by noise. In future work, we plan to investigate the effects of systematic errors on other types of centrality such as eigenvector centrality, or entirely different analysis methods such as community detection.

Conclusion

We introduced a general framework for simulating systematic edge errors in attributed networks. Our framework discriminates between node attribute-based errors, such as label congruence-based and label specific errors, as well as network structure-based errors, such as neighborhood overlap-based errors and centrality based-errors. We applied this simulation framework to investigate the representation of minorities in rankings based on the degree centrality of nodes with binary labels, representing two groups. In our numerical simulations on synthetic networks we find that the effect of the systematic bias is dependant on the network topology: In heterophilic networks, majority/minority representations in rankings are sensitive to the type of edge error present. In contrast, in homophilic networks we find that minorities are at a disadvantage regardless of the type of error present. We also performed error simulations on real world networks, which led to similar effects as the synthetic networks.

Our results emphasize that systematic errors can heavily influence results of network analyses and the nature of the effect depends on the specific data set. This emphasises the necessity of a flexible framework such as ours to enable researchers to account for systematic errors in their social network analysis.

(a) h=0.25 (heterophilic)   **(b) h=0.75 (homophilic)**

Fig. 7 Sensitivity of our results with regard to the choice of k. We visualize the fraction of the minority in the top k ranked nodes as a function of k, for different retain rates for majority associated noise. The solid line is the mean while the enveloping curve represent the range of values obtained in 10 iterations. Ties are resolved randomly. Due to large noise for the top ranked position we only start visualising for k > 10. We see that the effects displayed in Fig. 3 for k = 100 are representative for a wider range of k values

(a) h=0.25 (heterophilic)   **(b) h=0.75 (homophilic)**

Fig. 8 Complementary cumulative distribution by class for synthetic networks and majority associated noise. The statistics for minority/majority are represented by dashed/solid lines respectively. Different colours correspond to different values of the retain parameter ρ. To increase clarity, standard deviations are not shown. For fixed ρ, comparatively higher values of the solid line indicate an over representation of the majority in degree rankings, equal values proportional representation. As expected the majority noise has a strong effect on the majority but also a minor effect on the minority. In heterophilic regimes the minority remains disproportionately represented for any ranking. In homophilic regimes for $\rho = 0.1$ we can have a ranking with proportionate representation if we consider rankings above the top 0.1% of nodes

Appendix

Sensitivity analysis of *k*

In the main article, we investigated the representation of the minority in the top $k = 100$ ranked nodes. We now examine the dependency of these results on *k*. Figure 7 shows the fraction of minority in the top *k* ranked nodes in a heterophilic ($h = 0.25$) and a homophilic ($h = 0.75$) regime for different strengths of majority-noise. We observe that in the heterophilic regime, the over-representation of the minority in the top-k nodes is essentially independent of the parameter *k* used. By contrast, in the homophilic regime we see a slight increase in representation with larger values of *k* for values of $\rho > 0.1$. For $\rho = 0.1$ the over-representation is present almost throughout the range of *k* values.

An alternative way to access the impact on node rankings across different values of *k* is shown in Fig. 8. There we see the impact of majority noise on the group specific complementary cumulative degree distribution. While these plots do not allow to access a ranking directly, we can easily extract rankings for nodes with degree greater *x*, where *x* is chosen on the abscissa. Thus, differences between the solid and the dashed line at a specific value on the x-axis indicate an under- or over-representation of the minority.

Table 3 Overview of Copenhagen Network Study networks (Sapiezynski et al. 2019)

Name	Nodes	Edges	Q_{mod}	A	f (%)
Facebook	744	5865	0.043	0.11	22
Sms	429	580	0.093	0.23	27
Calls	322	440	0.081	0.22	25
Bluetooth (face-to-face)	658	74,808	0.01	0.03	22

As in Table 2, we use Q_{mod}/A to denote the modularity/assortativity of the partition based on minority and majority label. Both metrics may be interpreted as a proxy for the homophily in the dataset. The fraction of the minority in the dataset is denoted by f
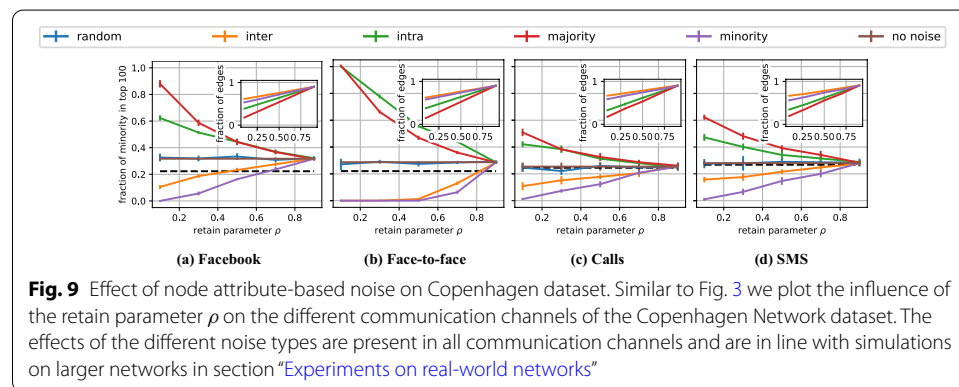
Fig. 9 Effect of node attribute-based noise on Copenhagen dataset. Similar to Fig. 3 we plot the influence of the retain parameter ρ on the different communication channels of the Copenhagen Network dataset. The effects of the different noise types are present in all communication channels and are in line with simulations on larger networks in section "Experiments on real-world networks"
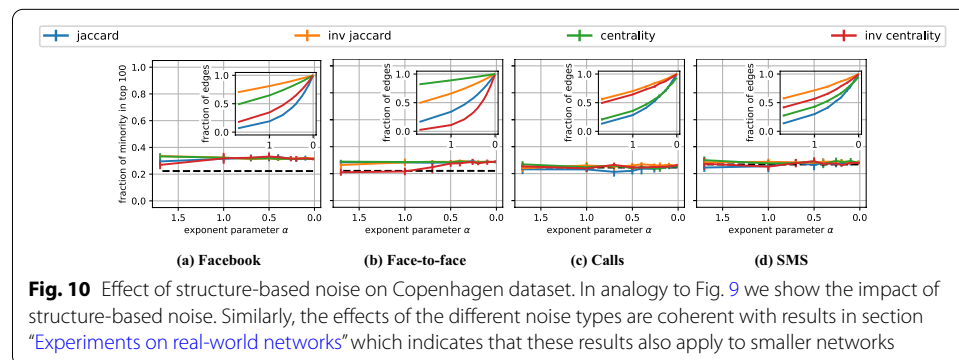
Fig. 10 Effect of structure-based noise on Copenhagen dataset. In analogy to Fig. 9 we show the impact of structure-based noise. Similarly, the effects of the different noise types are coherent with results in section "Experiments on real-world networks" which indicates that these results also apply to smaller networks

Proportional representation in degree-based rankings is achieved whenever both lines intersect.

We have highlighted one such point which shows that rankings, which include fewer than the top 0.1% of nodes have a proportionate representation.

Analysis of network size effects

Our analysis of both synthetic and real-world networks emphasise that the robustness of network analysis results with regard to a specific bias depend on the particular network topology. In section "Experiments on real-world networks" we mainly applied our model

to large real world networks. However, these insights are valuable to political scientists, sociologists, and economists who might be interested in networks of smaller sizes. In this section, we also carried out numerical experiments on networks with smaller node sets in order to know if the effects of our error simulations depend on the network size.

In particular, we use the different interaction networks which were collected via smartphones as part of the Copenhagen Network Study (Sapiezynski et al. 2019). We include the network of physical proximity among the participants (estimated via Bluetooth signal strength), the network of phone calls, the network of text messages, and the network of Facebook friendships. A summary of the network statistics can be found in Table 3.

We provide a complete overview of simulations for all node attribute-based and structure-based biases in Figs. 9 and 10. The results are in line with the observations for large-size networks of similar modularity (as a proxy for the degree of homophily in the system) in section "Experiments on real-world networks" which implies that the qualitative effects of the different systematic errors is mainly impacted by the degree of homophily in the system but not the network size. This is due to the fact that we only consider node-degree based rankings in this work which is a local node property not influenced by networks size.

Declarations

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

Author details
[1]RWTH Aachen University, Aachen, Germany. [2]University of Passau, Passau, Germany. [3]GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany.

Neuhäuser *et al. Appl Netw Sci*   (2021) 6:86

Page 21 of 22

References

Adiga A, Vullikanti AKS (2013) How robust is the core of a network? In: Blockeel H, Kersting K, Nijssen S, Železný F (eds) Machine learning and knowledge discovery in databases. Springer, pp 541–556

Almquist ZW (2012) Random errors in egocentric networks. Soc Netw 34(4):493–505. https://doi.org/10.1016/j.socnet.2012.03.002

Avella-Medina M, Parise F, Schaub MT, Segarra S (2020) Centrality measures for graphons: accounting for uncertainty in networks. IEEE Trans Netw Sci Eng 7(1):520–537. https://doi.org/10.1109/TNSE.2018.2884235

Bell DC, Belli-McQueen B, Haider A (2007) Partner naming and forgetting: recall of network members. Soc Netw 29(2):279–299. https://doi.org/10.1016/j.socnet.2006.12.004

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Soc Netw 28(2):124–136. https://doi.org/10.1016/j.socnet.2005.05.001

Braithwaite I, Callender T, Bullock M, Aldridge RW (2020) Automated and partly automated contact tracing: a systematic review to inform the control of COVID-19. Lancet Digit Health 2(11):e607–e621. https://doi.org/10.1016/S2589-7500(20)30184-9

Brewer DD (2000) Forgetting in the recall-based elicitation of personal and social networks. Soc Netw 22(1):29–43. https://doi.org/10.1016/S0378-8733(99)00017-9

Calloway M, Morrissey JP, Paulson RI (1993) Accuracy and reliability of self-reported data in interorganizational networks. Soc Netw 15(4):377–398. https://doi.org/10.1016/0378-8733(93)90013-B

Calvó-Armengol A, Jackson MO (2004) The effects of social networks on employment and inequality. Am Econ Rev 94(3):426–454. https://doi.org/10.1257/0002828041464542

Clauset A, Moore C, Newman MEJ (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453(7191):98–101. https://doi.org/10.1038/nature06830

Dasaratha K (2020) Distributions of centrality on networks. Games Econ Behav 2020:27. https://doi.org/10.1016/j.geb.2020.03.008

Easley D., Kleinberg J (2010) Networks, crowds, and markets: reasoning about a highly connected World. Cambridge University Press

DuBois T, Eubank S, Srinivasan A (2012) The effect of random edge removal on network degree sequence. Electron J Comb 19(1):v19i1p51

Everett JAC, Faber NS, Crockett M (2015) Preferences and beliefs in ingroup favoritism. Front Behav Neurosci 9:15. https://doi.org/10.3389/fnbeh.2015.00015

Feld SL, Carter WC (2002) Detecting measurement bias in respondent reports of personal networks. Soc Netw 2002:19. https://doi.org/10.1016/S0378-8733(02)00013-8

Frantz TL, Cataldo M, Carley KM (2009) Robustness of centrality measures under uncertainty: examining the role of network topology. Comput Math Organ Theory 15(4):303–328. https://doi.org/10.1007/s10588-009-9063-5

González-Bailón S (2014) Assessing the bias in samples of large online networks. Soc Netw 2014:12

Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the reconstruction of complex networks. Proc Natl Acad Sci 106(52):22073–22078. https://doi.org/10.1073/pnas.0908366106

Hannák A, Wagner C, Garcia D, Mislove A, Strohmaier M, Wilson C (2017) Bias in online freelance marketplaces: evidence from TaskRabbit and fiverr. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing. ACM, Portland Oregon USA, 1914–1933. https://doi.org/10.1145/2998181.2998327

Holland PW, Leinhardt S (1973) The structural implications of measurement error in sociometry. J Math Sociol 3(1):85–111. https://doi.org/10.1080/0022250X.1973.9989825

Holme P, Edling CR, Liljeros F (2004) Structure and time evolution of an Internet dating community. Soc Netw 26(2):155–174. https://doi.org/10.1016/j.socnet.2004.01.007

Karimi F (2019) Github repository for Github and APS dataset. https://github.com/frbkrm/NtwPerceptionBias

Karimi F, Génois M, Wagner C, Singer P, Strohmaier M (2018) Homophily influences ranking of minorities in social networks. Sci Rep 8(1):11077. https://doi.org/10.1038/s41598-018-29405-7

Karimi F, Wagner C, Lemmerich F, Jadidi M, Strohmaier M (2016) Inferring gender from names on the web: a comparative evaluation of gender detection methods. In: Proceedings of the 25th international conference companion on World Wide Web, pp 53–54

Kossinets G (2006) Effects of missing data in social networks. Soc Netw 28(3):247–268. https://doi.org/10.1016/j.socnet.2005.07.002

Lee SH, Kim P-J, Jeong H (2006) Statistical properties of sampled networks. Phys Rev E 73(1):016102. https://doi.org/10.1103/PhysRevE.73.016102

Lerman K, Yan X, Wu X-Z (2016) The "majority illusion" in social networks. PLoS ONE 2016:13. https://doi.org/10.1371/journal.pone.0147617

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inform Sci Technol 58(7):1019–1031. https://doi.org/10.1002/asi.20591

Lü L, Pan L, Zhou T, Zhang Y-C, Stanley HE (2015) Toward link predictability of complex networks. Proc Natl Acad Sci 112(8):2325–2330. https://doi.org/10.1073/pnas.1424644112

Marsden PV (1990) Network data and measurement. Annu Rev Sociol 16(1):435–463. https://doi.org/10.1146/annurev.so.16.080190.002251

Marsden PV (2003) Interviewer effects in measuring network size using a single name generator. Soc Netw 25(1):1–16

Martin S, Carr RD, Faulon J-L (2006) Random removal of edges from scale free graphs. Phys A Stat Mech Appl 371(2):870–876

Moore C, Ghoshal G, Newman MEJ (2006) Exact solutions for models of evolving networks with addition and deletion of nodes. Phys Rev E 74(3):036121. https://doi.org/10.1103/PhysRevE.74.036121

Murai S, Yoshida Y (2019) Estimating walk-based similarities using random walk. In: The World Wide Web conference on—WWW '19. ACM Press, San Francisco, CA, USA, 1321–1331. https://doi.org/10.1145/3308558.3313421

Newman M (2018a) Networks. Oxford University Press

Newman MEJ (2018b) Network structure from rich but noisy data. Nat Phys 14:5

Newman MEJ, Clauset A (2016) Structure and inference in annotated networks. Nat Commun 7(1):11863. https://doi.org/10.1038/ncomms11863

Nilizadeh S, Groggel A, Lista P, Das S, Ahn Y-Y, Kapadia A, Rojas F (2016) Twitter's glass ceiling: the effect of perceived gender on online visibility, p 10 (2016). https://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13003

Peel L, Larremore DB, Clauset A (2017) The ground truth about metadata and community detection in networks. Sci Adv 3(5):e1602548. https://doi.org/10.1126/sciadv.1602548

Peixoto TP (2018) Reconstructing networks with unknown and heterogeneous errors. Phys Rev X 8(4):041011. https://doi.org/10.1103/PhysRevX.8.041011

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. https://doi.org/10.1073/pnas.0914080107

Rodi GC, Loreto V, Tria F (2017) Search strategies of Wikipedia readers. PLoS ONE 12(2):e0170746. https://doi.org/10.1371/journal.pone.0170746

Sapiezynski P, Stopczynski A, Lassen DD, Lehmann S (2019) Interaction data from the Copenhagen Networks Study. Sci Data 6(1):315. https://doi.org/10.1038/s41597-019-0325-x

Sen I, Floeck F, Weller K, Weiss B, Wagner C (2019) A total error framework for digital traces of humans. arXiv preprint arXiv:1907.08228

Shalizi Cosma Rohilla, Thomas Andrew C (2011) Homophily and Contagion Are Generically Confounded in Observational Social Network Studies:. *Sociological Methods & Research*. https://doi.org/10.1177/0049124111404820 Publisher: SAGE PublicationsSage CA: Los Angeles, CA

Smieszek T, Burri EU, Scherzinger R, Scholz RW (2012) Collecting close-contact social mixing data with contact diaries: reporting errors and biases. Epidemiol Infect 140(4):744–752. https://doi.org/10.1017/S0950268811001130

Smith A, Duggan M (2013) Online dating & relationships. https://www.pewresearch.org/internet/2013/10/21/online-dating-relationships/

Strogatz SH (2001) Exploring complex networks. Nature 410(6825):268–276. https://doi.org/10.1038/35065725

van Tilburg T (1998) Interviewer effects in the measurement of personal network size: a nonexperimental study. Sociol Methods Res 26:300–328. https://doi.org/10.1177/0049124198026003002

Wagner C, Singer P, Karimi F (2017) Sampling from social networks with attributes, pp 1181–1190. https://doi.org/10.1145/3038912.3052665

Wang DJ, Shi X, McFarland DA, Leskovec J (2012) Measurement error in network data: a re-classification. Soc Netw 34(4):396–409. https://doi.org/10.1016/j.socnet.2012.01.003

Wiese J, Min J-K, Hong JI, Zimmerman J (2015) "You never call, you never write": call and SMS logs do not always indicate tie strength. In: Proceedings of the 18th ACM conference on computer supported cooperative work and social computing(CSCW '15). Association for Computing Machinery, New York, NY, USA, pp 765–774. https://doi.org/10.1145/2675133.2675143

Yang J, McAuley J, Leskovec J (2013) Community detection in networks with node attributes. In: 2013 IEEE 13th international conference on data mining, pp 1151–1156. https://doi.org/10.1109/ICDM.2013.167

Yang J, Ribeiro B, Neville J (2017) Should we be confident in peer effects estimated from partial crawls of social networks?

Young J-G, Cantwell GT, Newman MEJ (2020) Robust Bayesian inference of network structure from unreliable data. arXiv:2008.03334 [physics, stat]

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.