Received: 30 May 2022 | Revised: 17 July 2022 | Accepted: 29 July 2022

DOI: 10.1002/sta4.491

ORIGINAL ARTICLE

WILEY

# Fallacy of data-selective inference in modelling networks

**Stefan Stein** | **Chenlei Leng**

Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK

**Correspondence**
Chenlei Leng, Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.
Email: c.leng@warwick.ac.uk

Recent years have seen a growing array of activities in developing statistical models for modelling real-life networks. Since many of these networks are sparse, an all too often practice in the literature is to apply a developed model to a subnetwork typically by discarding nodes due to their lack of connectivity. In this note, we provide the first result highlighting issues with this practice which we call the fallacy of data-selective inference. We demonstrate this fallacy by examining the estimation bias in the Erdős–Rényi model theoretically and in the stochastic block model empirically.

KEYWORDS
data-selective inference, Erdős–Rényi, sparse networks, stochastic block model

## 1 | INTRODUCTION

The emergence of complex systems often calls for models for modelling interactions among their individual components. Towards this, statistical analysis of networks offers a useful approach that can depict their stochastic natures while providing tools for uncertainty quantification and prediction. As a result, we have seen a growing array of activities in developing statistical network models (Fienberg, 2012; Kolaczyk, 2009; Newman, 2018). These include the stochastic block model (Holland et al., 1983) and the $\beta$-model (Chatterjee et al., 2011) and their variants, among many others.

Despite the rapid deployment of these models, there are many open questions and unresolved challenges surrounding statistical analysis of networks, some of which are discussed in Kolaczyk (2017). In the present paper, we discuss in detail another widely employed yet highly questionable practice which is somewhat overlooked in the literature and highlight its fallacy. We name this fallacy *data-selective inference*. It often arises when a developed network model is applied to real-life networks, typically by focusing on selected nodes while leaving out a nonnegligible portion. The nodes that are retained for modelling are usually in a giant component and/or better connected than those discarded ones. This practice is widely applied in statistics. Below, we present several prominent examples including two statisticians' favorite datasets, with a snapshot list of references.

- Political blog data. This is a dataset recorded during the 2004 U.S. Presidential Election in the form of a directed network of hyperlinks between 1, 494 political blogs (Adamic & Glance, 2005). Depending on their political views, these blogs can be liberal or conservative. Often converted to an undirected graph for analysis, this dataset has become a testbed for many network models for community detection, especially for the stochastic block model and its generalizations. In practice, however, most papers chose to focus on 1,222 blogs that appear in a giant component or 1, 224 nodes which have at least one connection. In the latter case, all isolated nodes are removed for further analysis. Either way, this amounts to removing about 18% of the nodes in this network. See Amini et al. (2013), Olhede and Wolfe (2014), Jin (2015), Cai and Li (2015), Caron and Fox (2017), Chen and Lei (2018), Huang and Feng (2018) and Ma, Ma & Yuan (2020) for various analysis of this dataset, among many others.
- Statistician citation network. This dataset, curated by Ji and Jin (2016), contains rich citation information about all papers published between 2003 and 2012 in four statistics journals. The original dataset has 3,607 authors or nodes based on which various networks have been constructed, but almost all attempts to use this data have chosen to examine subnetworks with fewer than 3,607 nodes. Ji and Jin (2016) applied various community detection methods to three networks constructed from this dataset. The first one is a co-authorship network with

236 nodes (7% of all nodes), in which a link is formed between two authors if they wrote at least two papers together. See also Jin et al. (2021). The second one is another co-authorship network with 2,236 nodes (63% of all nodes), in which a link is formed between two authors if they wrote at least one paper together. The third one is a directed citation network with 2,654 authors (74% of all nodes). See also Zhang et al. (2021). Other attempts to use this dataset include Li et al. (2020) in which a network with 706 authors (20% of all nodes) was formed by repeatedly deleting nodes with less than 15 mutual citations and their corresponding edges. Jin et al. (2021) examined a citee network with 1,790 (50% of all nodes) constructed by tying an edge between two authors if they have been cited at least once by the same author other than themselves.

In addition to the datasets above, there is a large growing body of works opting for data-selective inference that remove nodes before analysis. Among many others, see Chen et al. (2018) and Ma, Ma & Yuan (2020) for the Simmons College and Caltech data, two datasets on friendship networks in universities, Sengupta and Chen (2018) for the British MPs network (where 329 out of 360 MPs belonging to the giant component were analyzed), Ma, Su & Zhang (2020) for Pokec social network (for which only those nodes with no fewer than 10 links were retained for analysis), and Yan et al. (2019) for using a directed $\beta$-model for analysing the Lawyer dataset (in which 8 out of 71 nodes were removed). A notable feature of the analyses in these papers is that nonnegligible portions of the nodes are excluded.

Statistical models are inherently motivated by data in real life. Real-life networks are known to be typically sparse in the sense that the total number of observed edges does not scale proportionally to the total number of possibly edges. As a result, there often exists a nonnegligible fraction of nodes either disconnected from the largest connected component or not having enough links, prompting the use of data-selective inference. We have seen several examples discussed already. There are various reasons why this inference is employed. Chief among them is probably due to convenience but more significantly due to the nature of a developed model and its associated algorithms not working in case of disconnected networks or very sparse networks. For example, the degree-corrected stochastic block model (Karrer & Newman, 2011) and the $\beta$-model (Chatterjee et al., 2011) cannot handle nodes with zero degree for obvious reasons. On the other hand, it is frequently argued that those nodes and edges in the giant component, or those nodes that are better connected, are the only nodes and edges that matter.

However, this data-selective inference selects nodes via what is known as nonrandom sampling in statistics, since nodes in a giant component or well-connected nodes are systematically favored over other nodes. This discussion raises immediately the following fundamental question:

*Does data-selective inference provide valid inference?*

Before we provide an answer, we note that an argument to avoid the question above would be to simply assume that the intended statistical model only works for the nodes in a giant component or those nodes that are well connected. While this argument is acceptable for mathematical convenience, it is not logically coherent or correct from a statistical or practical point of view. The selection of nodes is based entirely on the links (the response variable in a network model) and thus is nonrandom. Intuitively, this type of biased sampling may produce artificial signal that does not exist at all or mitigate existing signals or both, leading to problematic or even completely wrong findings. The fact that a nonnegligible fraction of nonrandom nodes is removed from analysis suggests that systematic bias will occur as a result.

The practice of ignoring selected nodes for modelling a network appears to originate from physics and computer science communities where the intention was to find meaningful clusters of nodes and hence is not statistical model-based (Girvan & Newman, 2002; Newman, 2006). Later, statisticians injected rigour into this line of research, notably by introducing likelihood-based estimators for statistical network models (Bickel & Chen, 2009). On the one hand, statistical modelling is extremely attractive because it provides a proper probabilistic framework for statistical inference and allows easy generalization of a model to more complex situations. On the other hand, however, issues such as sampling and asymptotics, including consistency and limiting distributions as the size of a network grows, inevitably arise. In particular, it is no longer appropriate for a statistical framework to ignore the nonrandom sampling issue in data-selective inference that removes nodes based on their links.

In the next section, we highlight the problem of data-selective inference by quantifying the bias of the parameter estimates for the simplest network model theoretically and that for the stochastic block model via simulation. Since consistency of parameter estimation is a basic requirement for any statistical model, our rational for examining estimation consistency is obvious. In particular, any systematic bias incurred in this data-selective inference will have knock-on effects on all aspects of any downstream analysis, including goodness-of-fit measures of a model, hypothesis testing and model selection; see, for example, Bickel and Sarkar (2016), Lei (2016), Wang and Bickel (2017), and Hu et al. (2020), for additional use of data-selective inference for data analysis.

## 2 | DATA-SELECTIVE INFERENCE PRODUCES BIAS

We now quantify the bias caused by data-selective inference in some simple network models when only those nodes in a giant component are used to fit a model. If data-selective inference poses problems for these simple models, it will not work for more complex models. In what follows, we assume that an observed network is the realization of some statistical network model $f \in \mathcal{F}$, with $\mathcal{F}$ a family of candidate models. Crucially, we will assume that $f$ would have produced the whole network, including any isolated vertices or small components. Given a realized network from the model, we want to quantify the bias of the estimator of the unknown parameter(s) in $f$, if we only use those nodes in the giant component.

Motivated by the several widely used datasets in the literature as discussed in Section 1, we will specify the parameter(s) in $f$ such that a fixed proportion of nodes is not in the giant component.

We will derive theoretically the bias of data-selective inference in the Erdős–Rényi model (Erdős & Rényi, 1959, 1960) and use simulation to study the bias in estimating the parameters in a simple stochastic block model. The Erdős–Rényi model is interesting because the methods and approaches developed for understanding this model and the insight it brings provide the foundation for the study of more general graphs. On the other hand, the stochastic block model is one of the most popular network models that is widely applied, studied and extended in the literature.

## 2.1 | The Erdős–Rényi model

As a reminder, the Erdős–Rényi model deals with undirected graphs in which edges are independently formed with the same probability $p$. A sufficient condition for a realized network from this model to have a giant component and smaller components with probability tending to one is to take $p = p(n) = \lambda/n$ for some fixed $\lambda > 1$ (Chapter 4, e.g., van der Hofstad, 2016). We will denote this family of models by $\mathrm{ER}(\lambda/n)$, and we are interested in estimating $p$ given a network generated from this model. Notice that $\lambda$ is very close to the expected degree of each node, which is given by $\lambda \cdot (n-1)/n$. Consequently, $\mathrm{ER}(\lambda/n)$ produces sparse networks with the expected total degree scaling linearly in $n$.

Denote $\eta_\lambda$ as the unique solution for which $\eta_\lambda < 1$ of the equation

$$\eta_\lambda = e^{\lambda\cdot(\eta_\lambda - 1)}, \tag{1}$$

which exists if and only if $\lambda > 1$. It is known that in this regime, $\mathrm{ER}(\lambda/n)$ will produce giant components with size tightly concentrating around $(1-\eta_\lambda)\cdot n$ (Theorem 4.8, e.g., van der Hofstad, 2016). We can interpret $1-\eta_\lambda$ as the survival probability of $\mathcal{P}(\lambda)$, the Poisson branching process with mean offspring $\lambda$, whose behaviour is closely linked to the connectivity behaviour of $\mathrm{ER}(\lambda/n)$ (Chapters 3 and 4, van der Hofstad, 2016). Consider the estimation of $p$ by only focusing on the induced subgraph $G_{\max}$ of $\mathrm{ER}(\lambda/n)$ consisting of the giant component, that is, $G_{\max} = (\mathcal{C}_{\max}, E(\mathcal{C}_{\max}))$, where $\mathcal{C}_{\max}$ denotes the giant component and $E(\mathcal{C}_{\max})$ contains only those edges between the nodes contained in $\mathcal{C}_{\max}$. Denote by $\hat{p}_{\max}$ the maximum likelihood estimator of $p$ based on $G_{\max}$, that is,

$$\hat{p}_{\max} = \frac{|E(\mathcal{C}_{\max})|}{\binom{|\mathcal{C}_{\max}|}{2}} \tag{2}$$

is the maximum likelihood estimator using data-selective inference.

**Proposition 1.** Fix any $\lambda > 1$ and consider the models $\mathrm{ER}(\lambda/n)$. Let $G_{\max} = (\mathcal{C}_{\max}, E(\mathcal{C}_{\max}))$ be the induced subgraph of $\mathrm{ER}(\lambda/n)$ consisting only of the giant component. Let $\hat{p}_{\max}$ be as in (2), $p = p(n) = \lambda/n$ and $\eta_\lambda$ as in (1). Then,

$$\frac{\hat{p}_{\max}}{p} \xrightarrow{P} \frac{1+\eta_\lambda}{1-\eta_\lambda}.$$

A few remarks are in order. Firstly, the estimator $\hat{p}_{\max}$ is unbiased asymptotically only when $(1+\eta_\lambda)/(1-\eta_\lambda) = 1$, which never holds for fixed $\lambda > 1$. Indeed, $(1+\eta_\lambda)/(1-\eta_\lambda) > 1$, for any fixed $\lambda > 1$. Thus, the incurred bias, that is, the asymptotic factor by which we are overestimating $p$, will not disappear as $n$ grows large. Figure 1 shows how the asymptotic bias $(1+\eta_\lambda)/(1-\eta_\lambda)$ deviates from one as a function of $\lambda$. Secondly, we can see that the bias increases when $\lambda$ decreases. Indeed, the larger $\lambda$, the more nodes in the giant component and the smaller the probability $\eta_\lambda$ and thus the smaller the bias. On the other hand, as $\lambda \to 1$, it is easy to see that $\eta_\lambda \to 1$, making the bias in Proposition 1 approach $+$. For the limit case $\lambda = 1$, $\mathcal{C}_{\max}$ will have size of order $n^{2/3}$ (Chapter 5 van der Hofstad, 2016), which in light of the proposition means that we must abandon all hope of recovering $p$ if we only focus on the giant component.

While Proposition 1 follows readily from results in the random network literature, to the best of our knowledge, it has not been stated in this form in the literature before. In particular, the results we draw upon for its proof are mostly rooted in probability theory and as far as we know have not been used before to explicitly quantify the biases incurred by statistical procedures.

We have chosen to focus on the regime $p \sim \lambda/n$ with a fixed $\lambda > 1$, since this regime seems most appropriate for modelling many networks observed in practice; see, for example, the discussion in Section 2.3. Nonetheless, it can be instructive to consider other regions of the parameter space for the Erdős–Rényi model, especially the scenarios in which $pn = \lambda = \lambda_n \to$ . When $\lambda > (1+\epsilon)\log(n)$, for an arbitrary $\epsilon > 0$, the Erdős–Rényi random graph will be connected with high probability (Theorem 5.8, van der Hofstad, 2016). As long as the only requirement for the intended model is connectivity of the observed network, data-selective inference will not be an issue in this regime. On the other hand, if $1 \ll \lambda < (1-\varepsilon)\log(n)$ for any $\epsilon > 0$, not only will the network be disconnected with high probability, it will also contain isolated nodes, making this

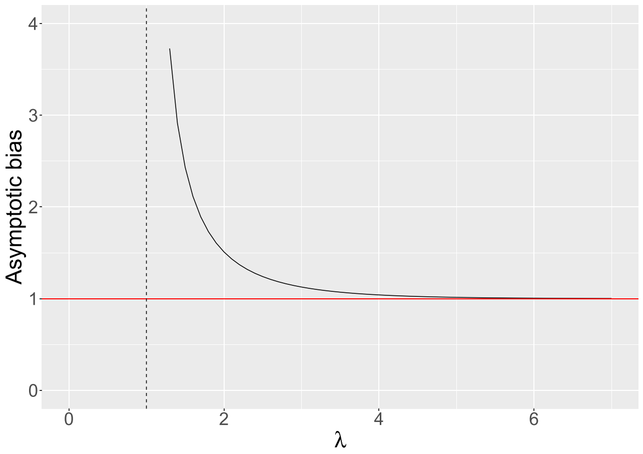

**FIGURE 1** Asymptotic bias $(1+\eta_\lambda)/(1-\eta_\lambda)$ of $\hat{p}_{\max}$ as a function of $\lambda$. For better visibility, we only display values of $\lambda$ ranging from 1.3 to 7 since the bias diverges to + when $\lambda$ approaches 1

regime susceptible to data-selective inference[1] (Theorem 5.8, van der Hofstad, 2016). More precisely, for $\lambda = \alpha\log(n), \alpha > 0$, the expected number of isolated vertices is roughly $n^{1-\alpha}$ (Proposition 5.9, van der Hofstad, 2016). An interesting behaviour can be observed in the critical regime between these two cases, when $\lambda = \log(n)+\beta$, for some fixed $\beta \in \mathbb{R}$. In this case, the number of isolated vertices converges in distribution to a Poisson random variable with mean $e^{-\beta}$, meaning asymptotically, we expect to observe $e^{-\beta}$ isolated vertices and we will observe no isolated vertices with probability $e^{-e^{-\beta}} + o(1)$ (equation (5.3.29), van der Hofstad, 2016). To give a concrete example, suppose a scaling of $\lambda = log(n)$ thus $\beta = 0$. Then, there would only be a chance of roughly $e^{-1} \approx 0.37$ that the observed network is connected and the number of isolated nodes that would be discarded by data-selective inference would be of constant order.

## 2.2 | The stochastic block model

This model postulates that nodes in a network can be grouped into communities where the probability of any pair of nodes making connections depends only on their community membership. We here focus on what is called the symmetric stochastic block model with two communities, for which the probability matrix of making connections is

$$P = \frac{1}{n}\begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

where $a,b > 0$ are constants. For this model, a pair of nodes link with probability $a/n$ within the same community and with probability $b/n$ between communities. The scaling $1/n$ ensures that a resulting network from this model will have a giant component with a fixed, smaller-than-one proportion of the nodes with high probability. See Figure 2 for the proportion of the nodes in the giant components produced under this parametrization. The symmetric stochastic block model is widely studied and relatively well understood (Abbe, 2018). Because of the sparsity of any resulting network, many clustering methods for community detection including spectral methods based on the adjacency matrix or the graph Laplacian, as well as their semidefinite relaxations, do not work well under this parametrization. Indeed, Zhang and Zhou (2016) showed that under our scaling, no consistent algorithm exists that achieves vanishingly small misclassification. In view of this, we take an *oracle* approach by assuming that the community membership of each node is known *a priori* and focus on what happens if we estimate $P$ when nodes not in the giant component are removed as in data-selective inference.

In our simulation, we fix the number of nodes to be $n = 10{,}000$ and set the size of each community as $n/2$. We consider a fine grid of values $(a,b)$ by taking their values from 0.05 to 8.05 in steps of 0.05, resulting in $25,921$ distinct combinations of $(a,b)$. For each such pair $(a,b)$, we sample a network from the symmetric stochastic block model and calculate the maximum likelihood estimate for $a$ and $b$ using only the nodes in the giant component. We repeat this process $M = 1{,}000$ times for every pair $(a,b)$.

Denote the resulting estimators as $\hat{a}$, $\hat{b}$ and $\hat{P}$. We measure the accuracy of the estimator by calculating the ratio of the spectral norm of the estimated probability matrix and that of the true probability matrix defined as

$$\rho = \|\hat{P}\|_2 / \|P\|_2,$$

[1] Indeed, for $n \geq \lambda > \alpha\log(n), \alpha > 1/2$, connectivity is essentially equivalent to the absence of isolated vertices, in the sense that, as $n \rightarrow$ , $\mathbb{P}(\mathrm{ER}(\lambda/n)\ \text{connected}) = \mathbb{P}(\text{there are no isolated vertices}) + o(1)$ (Proposition 5.10 van der Hofstad, 2016).

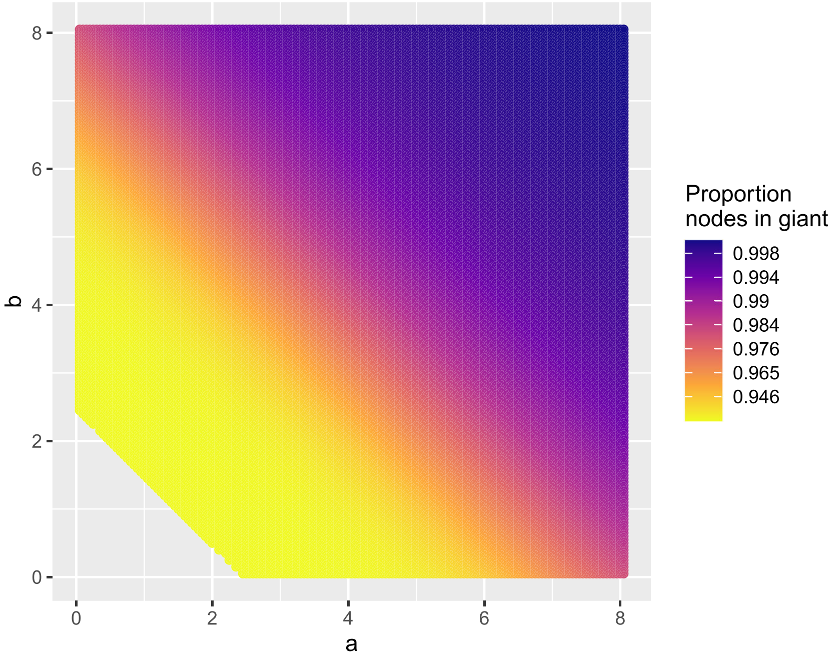

**FIGURE 2** Mean proportion of nodes in the giant component for each pair (a,b), averaged over $M=1{,}000$ repetitions. On the border $a+b=2.5$, the proportion of nodes in the giant is 37%. For better visibility, we have truncated by only including points for which $a+b\geq 2.5$. Also note the use of an exponential colour scaling for better visibility

where $\|P\|_2=(a+b)/n$ for the $2\times 2$ matrix. Note that we have purposefully chosen a large $n$ and $M$ so that the resulting averages of the estimates will be close to their true limit. Figure 2 shows the average proportion of nodes in the giant component for each pair $(a,b)$ and Figure 3 the average value of $\rho$. When the proportion is one, the giant component contains all the nodes. The closer the average value of $\rho$ is to one, the less biased the estimates are. In addition to the figures, Table 1 provides the average values of the size of the giant component and $\rho$ and their standard deviations, for selected values of $\|P\|_2$.

The simulations show that the incurred bias and the size of the giant component behave similarly when $a+b$ is a constant. We highlight the bias of the parameter estimates more closely. When $a+b=2.5$, the giant component on average contains 37% of all nodes and we overestimate $\|P\|_2$ by a factor of 4.4. This may not exactly come as a surprise: If we discard a large proportion of nodes, it can be expected that estimates are inaccurate. A more interesting and more critical behavior is observed as we make our way from the bottom left to the top right corner of the plots in Figures 2 and 3.

For $a+b=4$ (bottom black line in Figure 3), the giant component on average contains around 80% of all nodes, with an average bias $\rho=1.5$. For $a+b=6$ (second black line in Figure 3), the giant component contains on average 94% of all nodes, while the average estimated $\rho$ is 1.13, which is still significantly larger than one. Even for $a+b=8$ (middle black line in Figure 3), when the giant component contains on average 98% of all nodes, we overestimate $\|P\|_2$ by a factor of 1.04. Only once $a+b\geq 10.70$, where the giant component contains 99.5% of all nodes is the incurred bias smaller than 1%.

The above results illustrate that even in the idealistic scenario where the statistician has perfect knowledge of the underlying communities, and even if only a seemingly insignificant fraction of say, 0.5% of the nodes is removed, parameter estimation based solely on the giant component will be biased regardless of the network size. This casts severe doubt on the suitability of the stochastic block model and its variants for fitting many popular datasets if only the nodes in giant components are retained. As we have argued, having biased estimators will have consequences in all aspects of any downstream statistical inference including consistency, model selection, hypothesis testing and so on.

## 2.3 | A quick illustration of the fallacy

We now illustrate the fallacy of data-selective inference with the stochastic block model when it is applied to detecting communities in the political blog data network, to highlight a wider problem in statistical modelling of networks. If one assumes that this model generates all the 1,494 nodes, the mere existence of more than 200 nodes not in the giant component suggests imposing a scaling of $1/n$ on the connectivity probability matrix of the data generating process. This is similarly done previously to ensure that the resulting giant component contains a positive fraction of

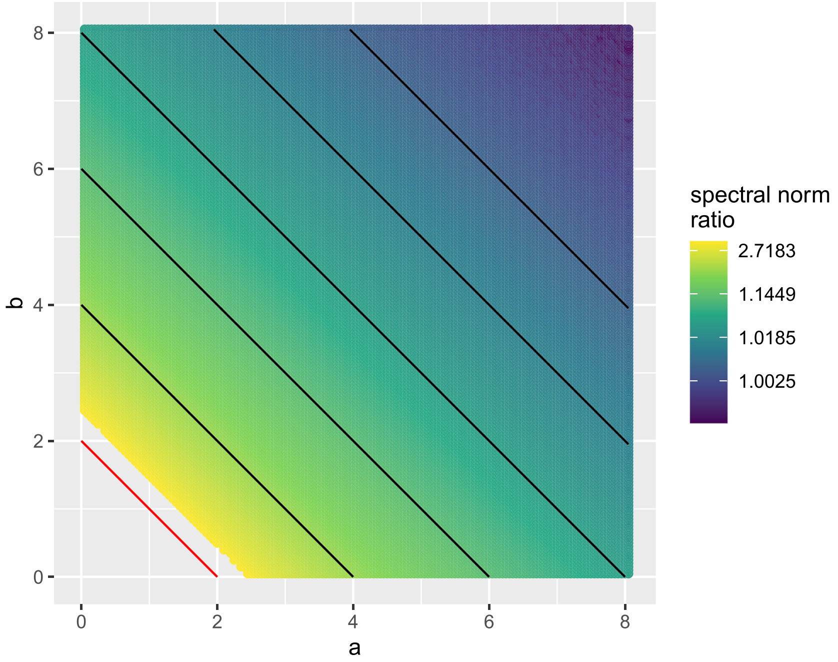

**FIGURE 3** Mean spectral ratio $\rho$ when estimates are based on the giant component only, averaged over $M = 1{,}000$ repetitions. The red line corresponds to $a+b=2$ and an average bias factor of 60.57. The black lines correspond to $a+b=4,6,8,10,12$ with bias factors $\rho = 1.51, 1.13, 1.04, 1.014, 1.005$, respectively. For better visibility, we have truncated by only including points for which $a+b \geq 2.5$. Also note the use of an exponential colour scaling for better visibility

**TABLE 1** Average size of the giant component and biases $\rho$ for different values of $\|P\|_2$

| $\|P\|_2$ | Size of the giant component | $\rho$ |
|---|---|---|
| 1 | 0.0018 (0e+00) | 2208 (163.9) |
| 2 | 0.0433 (1e-03) | 60.57 (2.1281) |
| 3 | 0.5824 (4e-04) | 2.4345 (0.0015) |
| 4 | 0.7968 (2e-04) | 1.5102 (0.0004) |
| 5 | 0.8927 (1e-04) | 1.2406 (0.0003) |
| 6 | 0.9405 (1e-04) | 1.1265 (0.0002) |
| 7 | 0.9660 (1e-04) | 1.0704 (0.0002) |
| 8 | 0.9802 (0e+00) | 1.0404 (0.0002) |
| 9 | 0.9883 (0e+00) | 1.0236 (0.0002) |
| 10 | 0.9930 (0e+00) | 1.0140 (0.0002) |
| 11 | 0.9958 (0e+00) | 1.0084 (0.0002) |
| 12 | 0.9975 (0e+00) | 1.0050 (0.0002) |

*Note*. The respective standard deviations, rounded to four significant digits, are given in parenthesis.

the vertices. Under this regime, however, there is no way to separate all the vertices and thus no algorithm can provide consistent community detection or parameter estimation (Abbe, 2018). If instead one focuses on the giant component and assumes consistent community estimation for the nodes in the giant component, the connectivity probability matrix will be overestimated with bias, as we have illustrated. By focusing on a nonrandom sample, we have no idea whether an intended model truly reflects the data generating process or rather is merely an artefact of biased sampling. The estimation bias of course is just the tip of a larger problem in biased sampling, as this bias will propagate through all stages of modelling process including model selection and inference.

Thus, for this dataset and many other widely used network datasets, our arguments lead to a fundamental choice that we statisticians need to make. If we assume that the stochastic block model generates the whole network including all the nodes, consistent community detection and parameter estimation will be impossible. On the other hand, if we make the unrealistic assumption that the model only generates a subnetwork consisting only of those nodes in the giant component, we face the problem of data-selective inference.

## 3 | CONCLUSION

We have discussed and highlighted for the first time the fallacy of data-selective inference, an all too often practice to discard nodes in a network due to their lack of connectivity. Although this fallacy seems ubiquitous in statistical applications of many network models, we had not been aware of any systematic study of its effect before our work.

The paper has focused on revealing the bias for estimating the probability of connectivity which is a function of edges. Intuitively, similar bias issues will arise for estimating other network related quantities as a result of data-selective inference. For example, by focusing on better connected nodes, one will have a distorted degree distribution, optimistic measures of various centralities of nodes, and an inflated transitivity index that measures the probability adjacent nodes of a network being connected.

We note that the practice of data-selective inference arises mainly in the network literature for community detection. For applying methods in this line of research to their datasets, we advocate that practitioners only focus on those networks that have few or no isolated nodes so that there is no need to employ bias sampling before analysis. On the other hand, if fitting stochastic block models to the whole dataset poses issues, an approach to alleviate these is to employ a Bayesian approach with a prior, perhaps informative enough, to discourage degenerate solutions. We remark that as far as community estimation is concerned, excluding isolated nodes or not will not cause much practical issue. Issues arise when it comes to estimating parameters in a model and conducting statistical inference.

While our paper may have painted a somewhat discouraging picture, we note that there are models that can model all the nodes. The Erdős–Rényi model offers a welcome start although it does not offer the needed complication for real-life networks. Another useful model is the exponential random graph model (Robins et al., 2007) having network motifs as sufficient statistics. In the Bayesian literature, the additive multiplicative model (Hoff, 2021) is another choice, though the theoretical properties of the estimators for this model are not clear.

By accounting for degree heterogeneity that characterizes the differential tendency of nodes to form links, the sparse $\beta$-model (Chen et al., 2021) handles all the nodes in a network. The model can be further generalized by adding covariates, as in Stein and Leng (2020) for undirected networks and in Stein and Leng (2021) for directed networks.

## ACKNOWLEDGEMENTS

We thank the Editor in Chief, an Associate Editor and two reviewers for their constructive comments that have led to an improved paper.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

*Chenlei Leng* https://orcid.org/0000-0001-5703-9617

## REFERENCES

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, *18*, 1–86.

Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 us election: Divided they blog. In *Proceedings of the 3rd international workshop on link discovery*, pp. 36–43.

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Annals of Statistics*, *41*(4), 2097–2122.

Bickel, P. J., & Chen, J. (2009). A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Science*, *106*, 21068–21073.

Bickel, P. J., & Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, *78*, 253–273.

Cai, T. T., & Li, X. (2015). Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Annals of Statistics*, *43*(3), 1027–1059.

Caron, F., & Fox, E. (2017). Sparse graphs using exchangeable random measures (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *79*, 1295–1366.

Chatterjee, S., Diaconis, P., & Sly, A. (2011). Random graphs with a given degree sequence. *Annals of Applied Probability*, *21*(4), 1400–1435.

Chen, K., & Lei, J. (2018). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, *113*(521), 241–251.

Chen, M., Kato, K., & Leng, C. (2021). Analysis of networks via the sparse beta model. *Journal of the Royal Statistical Society, Series B*, *83*, 887-910.

Chen, Y., Li, X., & Xu, J. (2018). Convexified modularity maximization for degree-corrected stochastic block models. *Annals of Statistics*, *46*(4), 1573–1602.

Erdős, P., & Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae (Debrecen)*, *6*, 290–297.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*, 17–60.

Fienberg, S. E. (2012). A brief history of statistical models for network analysis and open challenges. *Journal of Computational and Graphical Statistics*, *21*, 825–839.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, *99*(12), 7821–7826.

Hoff, P. (2021). Additive and multiplicative effects network models. *Statistical Science*, *36*(1), 34–50.

Holland, P. W., Laskey, K., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*, 109–137.

Hu, J., Qin, H., Yan, T., & Zhao, Y. (2020). Corrected Bayesian information criterion for stochastic block models. *Journal of the American Statistical Association*, *115*(532), 1771–1783.

Huang, S., & Feng, Y. (2018). Pairwise covariates-adjusted block model for community detection. arXiv:1807.03469.

Ji, P., & Jin, J. (2016). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, *10*(4), 1779–1812.

Jin, J. (2015). Fast community detection by score. *Annals of Statistics*, *43*(1), 57–89.

Jin, J., Ke, Z. T., & Luo, S. (2021). Estimating network memberships by simplex vertex hunting. *Annals of Statistics*.

Karrer, B., & Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical review E*, *83*(1), 016107.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and models*: Springer.

Kolaczyk, E. D. (2017). *Topics at the frontier of statistics and network analysis: (re)visiting the foundations*: Cambridge University Press.

Lei, J. (2016). A goodness-of-fit test for stochastic block models. *The Annals of Statistics*, *44*(1), 401–424.

Li, T., Levina, E., & Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, *107*(2), 257–276.

Ma, S., Su, L., & Zhang, Y. (2020). Detecting latent communities in network formation models. arXiv preprint arXiv:2005.03226.

Ma, Z., Ma, Z., & Yuan, H. (2020). Universal latent space model fitting for large networks with edge covariates. *Journal of Machine Learning Research*, *21*(4), 1–67.

Newman, M. (2018). *Networks* (2nd ed.). Oxford University Press.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, *103*(23), 8577–8582.

Olhede, S.C., & Wolfe, P.J. (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, *111*(41), 14722–14727.

Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph models for social networks. *Social Networks*, *29*, 173–191.

Sengupta, S., & Chen, Y. (2018). A block model for node popularity in networks with community structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(2), 365–386.

Stein, S., & Leng, C. (2020). A sparse $\beta$-model with covariates for networks. arXiv 2010.13604.

Stein, S., & Leng, C. (2021). A sparse random graph model for directed networks.

van der Hofstad, R. (2016). *Random graphs and complex networks*, Vol. 1: Cambridge University Press.

van der Hofstad, R. (2021). *Random graphs and complex networks*, Vol. 2: Preprint.

Wang, Y. X. R., & Bickel, P.J. (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics*, *45*(2), 500–528.

Yan, T., Jiang, B., Fienberg, S.E., & Leng, C. (2019). Statistical inference in a directed network model with covariates. *Journal of the American Statistical Association*, *114*(526), 857–868.

Zhang, A. Y., & Zhou, H. H. (2016). Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, *44*(5), 2252–2280.

Zhang, J., He, X., & Wang, J. (2021). Directed community detection with network embedding. *Journal of the American Statistical Association*, 1–11.

**How to cite this article:** Stein, S., & Leng, C. (2022). Fallacy of data-selective inference in modelling networks. *Stat*, *11*(1), e491. https://doi.org/10.1002/sta4.491

## APPENDIX A: PROOFS

Denote the Poisson branching process with mean offspring $\lambda$ by $\mathcal{P}(\lambda)$. That is, each vertex in $\mathcal{P}(\lambda)$ has offspring distribution *Poi*$(\lambda)$. It is well-known that the behaviour of ER$(\lambda/n)$ is intimately linked to the properties of $\mathcal{P}(\lambda)$. The following results on Poisson branching processes can be found, for example, in van der Hofstad (2016), chapter 3, and the links between $\mathcal{P}(\lambda)$ and $ER(\lambda/n)$ can be found in chapter 4 of van der Hofstad (2016).

We know that $\mathcal{P}(\lambda)$ with $\lambda > 1$ has a positive probability of surviving forever. More precisely, denote by $\eta_\lambda$ the probability that $\mathcal{P}(\lambda)$ dies out. Recall from Equation (1) that we may calculate $\eta_\lambda$ explicitly by finding the unique solution that is smaller than one to the equation (equation 3.6.2 van der Hofstad, 2016).

$$\eta_\lambda = e^{\lambda\cdot(\eta_\lambda - 1)}.$$

Notice that a solution smaller than one to (1) exists if and only if $\lambda > 1$. Conversely, we define the probability that $\mathcal{P}(\lambda)$ survives forever as

$$\zeta_\lambda = 1 - \eta_\lambda. \tag{A1}$$

Denote by $\mathcal{C}_{\max}$ the largest connected component of $ER(\lambda/n)$, where we suppress the dependence of $\mathcal{C}_{\max}$ on $\lambda$ and $n$ in our notation as it will be clear from the context. Then, for $\lambda > 1$, the size $|\mathcal{C}_{\max}|$ of the giant component will concentrate closely around $\zeta_\lambda n$ (Theorem 4.8 van der Hofstad, 2016).

We now prove Proposition 1. Proposition 1 follows from the following deep result on the phase transition of ER$(\lambda/n)$, which can be found in van der Hofstad (2021).

**Theorem 1 Phase transition in Erdős–Rényi random graphs, Theorem 2.33 in van der Hofstad (2021), abbreviated.** Fix $\lambda > 0$ and let $\mathcal{C}_{\max}$ be the largest connected component of the Erdős–Rényi graph ER$(\lambda/n)$. Then,

$$\frac{|\mathcal{C}_{\max}|}{n} \xrightarrow{P} \zeta_\lambda, \tag{A2}$$

where $\zeta_\lambda$ is the survival probability of a Poisson branching process with mean offspring $\lambda$. In particular, $\zeta_\lambda > 0$ precisely when $\lambda > 1$. Further, for $\lambda > 0$, with $\eta_\lambda = 1 - \zeta_\lambda$,

$$\frac{|E(\mathcal{C}_{\max})|}{n} \xrightarrow{P} \frac{1}{2}\lambda(1-\eta_\lambda^2). \tag{A3}$$

*Proof of Proposition* 1. Proposition 1 follows from Theorem 1 by repeated application of Slutzky's theorem. First, by (A2) and Slutzky,

$$\frac{|\mathcal{C}_{\max}|-1}{n} \xrightarrow{P} \zeta_\lambda.$$

Thus, by Slutzky,

$$\binom{|\mathcal{C}_{\max}|}{2} \cdot \frac{1}{n^2} \xrightarrow{P} \frac{1}{2} \cdot \zeta_\lambda^2.$$

Now, a final application of Slutzky's theorem together with (A2) and (A3) yields

$$\frac{\hat{p}_{\max}}{p} = \frac{|E(\mathcal{C}_{\max})|}{\binom{|\mathcal{C}_{\max}|}{2}} \cdot \frac{n}{\lambda} = \frac{|E(\mathcal{C}_{\max})|}{n} \cdot \frac{n^2}{\binom{|\mathcal{C}_{\max}|}{2}} \cdot \frac{1}{\lambda} \xrightarrow{P} \frac{(1-\eta_\lambda^2)}{\zeta_\lambda^2} = \frac{1+\eta_\lambda}{1-\eta_\lambda},$$

where we used the definition of $\zeta_\lambda$ in the last step.