Social Networks xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

The next steps in the study of missing individuals in networks: a comment on Smith et al. (2017)

Matthew J. Silk

Environment and Sustainability Institute, University of Exeter, Penryn, Cornwall, UK

ARTICLE INFO

Article history:
Received 12 October 2016
Accepted 4 May 2017
Available online xxx

Keywords:
network sampling
precision
bias
accuracy
statistical modelling

ABSTRACT

Social network analysis is now used widely to study social behaviour in humans and non-human animals, and missing individuals can represent a problem for network studies. This problem is becoming especially frequent in studies using bio-logging to collect interaction data, which is an approach used particularly frequently in the construction of animal networks. This therefore represents an important audience for Smith et al. (2017) who investigate how sub-sampling from networks impacts the outcome of subsequent analysis. Here I take advantage of the progress made by this paper to outline key issues that still require addressing to understand the effect of missing individuals on social network analysis.

1. Introduction

As a consequence of being relational data, the sampling of networks might inherently be expected to result in greater bias than other types of data (Alba, 1982; Silk et al., 2015; Smith et al., 2017; Smith and Moody, 2013). This could apply both to missing individuals (nodes) and missing relationships (edges), but the former is easier to quantify and address. A number of studies have explored the impact of missing individuals on network analysis, the most recent of which is Smith et al. (2017). The authors made substantial progress on a number of key issues, in particular in: a) assessing how non-random missingness of individuals might change the effect of sub-sampling from a network, and b) in providing a tool to allow researchers to determine the likely impact of missing individuals in a range of network structures and sizes.

Smith et al. (2017) stated that "*By looking at a wide range of networks, measures and types of missing data, we can offer recommendations and best practices for applied network practitioners*". A major application of network analysis away from the social sciences is in the study of animal behaviour (Croft et al., 2008; Krause et al., 2014). Missing individuals are a frequent problem in animal network studies, when it is often necessary to capture and mark individuals to gather data. However, as the use of bio-logging technology becomes more widespread to study human social networks (e.g. Isella et al., 2011; Kiti et al., 2016; Mastrandrea et al., 2015), similar problems often arise. I will provide a perspective as an applied network practitioner working on animal social behaviour as to the utility of their new findings, and then build on this to highlight important outstanding questions relating to missing individuals in networks. Finally, I will present some R code designed to test how missing individuals affect the calculation of network metrics in animal social networks that I hope will complement the java applet provided in that paper.

2. An applied perspective on the implications of Smith et al. (2017)

Smith et al. (2017) built upon previous work by the same authors (Smith and Moody, 2013) in exploring the consequences of missing individuals on the calculation of a range of network metrics. Together, especially when taken alongside complementary findings in other fields (e.g. Silk et al., 2015), the results of this work have revealed a set of important considerations when analysing networks with missing data, of which the general rules are already very useful to applied network practitioners. In particular, knowledge of how network structure and size influence the impact of sub-sampling from a network is paramount, as is an understanding that more global metrics (such as Betweenness) are less resilient to the presence of missing individuals. Finally, the exploration of biases in missing individuals addressed by Smith et al. (2017) is especially valuable from the perspective of animal behaviour research. This is partly as an aid in determining which individuals to collect data on when resources are limited, but also because methods of capture to make individual animals identifiable for network studies

E-mail address: matthewsilk@outlook.com

http://dx.doi.org/10.1016/j.socnet.2017.05.002
0378-8733/

may place implicit biases on which individuals are most likely to remain unsampled. However, one opportunity that is missed here is the opportunity to discuss the importance of different types of "bias" caused by missing network data. I would like to highlight here the nomenclature/approach used by Silk et al. (2015) which looked at the effect of sub-sampling on three distinct properties of network metrics. In this paper the authors looked at *precision*, *accuracy* and *bias* of metric values in sub-sampled networks. As defined by these authors, *precision* is the correlation between values calculated in the partial (observed) network and those in the equivalent true network, *accuracy* is the value of the metric obtained from the observed partial network relative to its true value, and *bias* is systematic variation in the precision of metric values in the partial (observed) network. The impact of sub-sampling on each of these properties can depend on network structure and the type of network metric being investigated. Considering each independently is important in wider applications of network analysis, as these different properties can be used to address different questions. For example, in animal network studies it is the *precision* of metric values in sub-sampled networks that is important if a researcher asks a question about whether network position and a personality trait are linked, but it is *accuracy* that is important if a researcher seeks to compare the true values of network metrics between different contexts.

Another major step forward in Smith et al. (2017) is the development of a tool (a java applet) that can provide an idea of the impact of missing individuals (incorporating any biases in their centrality) according to network size for a range of network structures. For an applied network researcher that has worked on study systems with substantial proportions of missing individuals this is an exciting development, and has the potential to be useful as a guide to researchers designing network studies. However, from this perspective I also feel that it is essential to see a tool designed in this way just as a starting point. The use of network analysis in animals is now highly question-driven and such a fixed tool has only restricted possibilities for use. It would be great to see a more modular set of functions that were able to use pilot empirical data or researcher knowledge to simulate a realistic network structure, and then sample from this structure, before determining how it might affect the outcome of a range of network analytical perspective. Such a package of functions would be best developed as a community, preferably with researchers from a range of fields, so that the generation of networks, and the types of metrics to calculate (or statistical models to assess) was relevant to as wide a range of studies as possible. The advantages of taking this approach is that as a user assesses a new sampling framework and/or question, the code they used can be added to the system and shared with researchers who might be faced with a similar problem in another field.

3. Three outstanding missing network data problems

In this section I will highlight some important gaps in our understanding on the impact of sub-sampling from social networks that simulation-modelling could easily test and greatly aid the design of empirical studies. While these ideas come from a background of employing network analysis to study animal behaviour, I feel all are widely applicable in the use of social networks more generally.

3.1. What is the best way of sub-sampling social networks?

In many animal network studies time, cost and effort is required to capture individuals and make them individually identifiable for social network studies. This trade-off is now becoming more frequent for all types of network study, including in humans, as the use of bio-logging approaches to produce reality mining data on social behaviour is increasing (Barrat and Cattuto, 2015; Isella et al., 2011). Often these approaches are costly, and it is possible to use only relatively small sample sizes. Therefore, deciding how best to deploy bio-logging devices for network studies remains an open question. For example, is it best to intensively sample a small part of a network or sample a larger part of the network more sparsely? Or similarly, how would studying replicate networks in multiple populations trade-off against the intensity of tagging individuals within each population? It is likely that the type of question being asked is important to making this decision. For example, work focussing on fine-scale behavioural interactions, such as dominance behaviour in animals (e.g. Dey et al., 2015), is likely to benefit from intensively sampling particular groups. In contrast, for the study of population-level processes such as disease transmission, a more even distribution of identifiable individuals throughout a population may be beneficial, especially when attempting to record infrequent interactions.

In order to assist empirical researchers making these decisions, it will be important to move simulation models of non-random sampling beyond the missingness-centrality correlation investigated by Smith et al. (2017) to assess the impacts of the clustering of identifiable and unidentifiable individuals within sub-sampled networks. Exploring this in a range of social network structures will be an important step forward in aiding the study design of network studies, especially for studies using bio-logging approaches.

3.2. How do missing individuals affect individual-level hypothesis testing in networks?

The relationship between social network position and other individual traits has been a major cross-disciplinary research focus (e.g. Aplin et al., 2013; Bollen et al., 2011; Goodreau et al., 2009; Rosenquist et al., 2011). The most popular methods to test these hypotheses has been different, for example the use of exponential random graph models (ERGMs; Lusher et al., 2013) and stochastic actor-oriented models (SAOMs; Snijders et al., 2010) in the social sciences, versus the development of randomisation-based generalised linear mixed model approaches in animal behaviour (Croft et al., 2011; Farine and Whitehead, 2015). Regardless, the impact of missing individuals (and edges) on statistical inferences made using these all of these approaches remains an open and important question.

Smith et al. (2017) made a step towards addressing this, by looking at the consequences of missing individuals for tests of behavioural homophily within a network (finding that it was possible to detect patterns of behavioural homophily when there was both a high proportion of missing individuals and a bias in which individuals were missing). However, a notion of the preponderance of type I and type II errors when different modelling frameworks are used to analyse networks with missing individuals would represent a considerable step forward in our understanding of the consequences of sub-sampling social networks. For example, while Shalizi and Rinaldo (2013) have suggested that ERGMs estimated on a sampled network are unlikely to reflect population-level parameters (*accuracy* in the framework outlined previously), this may not affect their ability to test hypotheses related to individual differences.

It would seem fairly simple to build on previous simulation-modelling work to examine how hypothesis testing using any of the statistical models mentioned above might be affected by the sub-sampling of networks. For example, the addition of a response variable that depended on network structure to the R function outlined in the next section would enable the impact on inference from generalised linear models to be addressed. In the case of models relating individual traits to individual-level network models, such as those suggested above, there are two main considerations to

Please cite this article in press as: Silk, M.J., The next steps in the study of missing individuals in networks: a comment on Smith et al. (2017). Soc. Netw. (2017), http://dx.doi.org/10.1016/j.socnet.2017.05.002