Social Networks 55 (2018) 55–62

Contents lists available at ScienceDirect

Social Networks

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

Mutual assent or unilateral nomination? A performance comparison of intersection and union rules for integrating self-reports of social relationships☆

Francis Lee[a], Carter T. Butts[b,*]

[a] Department of Sociology, University of California, Irvine, United States
[b] Departments of Sociology, Statistics, and EECS, and Institute for Mathematical Behavioral Sciences, University of California, Irvine, United States

ARTICLE INFO

Article history:

Keywords:
Network inference
Measurement
Locally aggregated structures
Informant accuracy
Error

ABSTRACT

Data collection designs for social network studies frequently involve asking both parties to a potential relationship to report on the presence of absence of that relationship, resulting in two measurements per potential tie. When inferring the underlying network, is it better to estimate the tie as present only when both parties report it as present or do so when either reports it? Employing several data sets in which network structure can be well-determined from large numbers of informant reports, we examine the performance of these two simple rules. Our analysis shows better results for mutual assent across all data sets examined. A theoretical analysis of estimator performance shows that the best rule depends on both underlying error rates and the sparsity of the underlying network, with sparsity driving the superiority of mutual assent in typical social network settings.

Network inference is the problem of inferring an unknown graph from a set of error and/or missingness-prone observations. This problem is of fundamental importance in the study of social networks, where relationships between individuals, organizations, or other entities must typically be inferred from self or proxy reports, archival materials, or other imperfect source of information. Arguably, the most basic and familiar example of the network inference problem arises when attempting to integrate self-reports from subjects, each of whom is asked to identify all others with whom he or she has a particular relationship (or, in the case of a directed relationship, all others to/from whom he or she respectively sends and/or receives ties). Such data has been widely collected (see, e.g. Drabek et al., 1981; Reitz and White, 1989; Bernard et al., 1984; Killworth and Bernard, 1979; Pattison et al., 2000), and poses a basic challenge for the analyst: given two reports on the state of a given relationship, what is to be done when the subjects disagree? Krackhardt (1987) famously formalized two basic strategies for the analysis of such data (leading to respective estimators of the underlying network): regard an edge as present if either party reports it (the *union rule*); or regard an edge as present if and only if both parties report it (the *intersection rule*). While one or another rule has in some cases been argued to be preferred on substantive grounds, there has been little systematic investigation of how the rules perform on empirical data, and in particular on the relative performance of these rules in inferring network structure under realistic conditions.

This paper seeks to address this gap, employing interpersonal networks whose complete structures can be well-estimated through hierarchical Bayesian models (Butts, 2003) to assess the accuracy of these simpler (but more easily used) rules. Our findings demonstrate that the intersection rule ("mutual assent") generally outperforms the union rule ("unilateral nomination") for the networks studied here - a surprising result, given that our informants are not prone to making one particular type of error over another. We resolve this discrepancy by showing that the sparsity of the network is key to the performance of the two rules, with the intersection rule dominating the union rule for networks in which the opportunities for false positives greatly outweigh the opportunities for false negatives.

1. Background

Social network analysis is intrinsically and trivially dependent on the ability to accurately measure the structure of social relationships. Despite the rise of network measurement via online social networks, mobile devices, and other sources of observational data, collection of self-reports via sociometric surveys continues to be a popular method for network measurement in a wide range of set-

☆ This research was supported in part by ARO award W911NF-14-1-055.
* Corresponding author.

https://doi.org/10.1016/j.socnet.2018.05.005
0378-8733/

tings. At least since the seminal studies of Bernard, Killworth, and Sailer – who provocatively (if hyperbolically) concluded that "there is no evidence that people know who their network connections are" (Bernard et al., 1984) – error from informant observations has been known to be a major challenge in network data collection and subsequent inference. Though subsequent studies (e.g. Freeman et al., 1987; Romney and Faust, 1982) have tempered the extremity of Bernard et al.'s conclusion, it is clear that error rates are substantial enough to warrant concern for social network research. With self-report (i.e. informants reporting on their own ties) remaining a popular method of network data collection, there is an ongoing need for simple methods that can maximize the accuracy of networks inferred from this type of information.

Although highly accurate estimates of network structure can be obtained when many measures of each potential tie are available (e.g., from cognitive social structure data – see Butts, 2003), simple self-report designs allow only two observations per edge variable (one for each party involved in the potential edge). The question, then, is how best to integrate these reports to infer the underlying network. Although many techniques are possible, we here focus on simple, easily used methods of aggregation that (1) estimate the state of an edge variable as being consistent with informants' reports where they agree, and (2) resolve disagreements via a simple uniform rule. Such strategies lead to estimators[1] referred to by Krackhardt (1987) as *locally aggregated structures* (LAS), the union and intersection rules (U-LAS, I-LAS) being the special cases mentioned above. LAS estimators can be employed for both undirected relations (when both parties report on the presence/absence of a single undirected edge) and directed relations (when both parties report on their incoming/outgoing ties); indeed, many networks collected in the former manner are erroneously treated as directed, where a LAS or other estimator of an underlying directed relation should be employed. The present work is thus applicable to any situation in which we obtain edge observations associated with both potential endpoints.

1.1. Formal framework

To formalize the above, our network inference problem may be summarized as follows. Let $G=(V, E)$ represent an unknown network of interest, with fixed and known vertex set V and unknown edge set E. Without loss of generality, we will represent G via its adjacency matrix, Θ; where G is undirected, Θ is constrained to be symmetric. Here, we assume the vertex set to be fixed and the edge set unknown. Informant reports are represented via an informant by sender by receiver adjacency array, Y, such that $Y_{ijk}=1$ if i reports that the edge from j to k is present (with 0 otherwise). In our setting, we assume that informants report only on their own ties, and hence only Y_{iij} and Y_{jij} (and their reciprocating edge variables) are employed. The locally aggregated structure (LAS) introduced by Krackhardt (1987) has been a popular network inference tool for aggregating an ego's and alter's judgments from such data. While this has traditionally been explored through cognitive social structures (which collect an informant's perception of the entire social structure), the only responses *needed* from an informant are the ties they report sending out and the ties they perceive others send to them. In terms of the above, the union and intersection LAS estimators are defined as follows:

$$\hat{\Theta}^U_{ij} = 1-(1-Y_{iij})(1-Y_{jij}) \tag{1}$$

$$\hat{\Theta}^I_{ij} = Y_{iij}Y_{jij} \tag{2}$$

[1] Krackhardt (1987) does not explicitly treat the LAS as a family of estimators per se, but employs them in a manner consistent with this interpretation.

As noted above, $\hat{\Theta}^U$ estimates an edge as being present when either party reports it, while $\hat{\Theta}^I$ does so only when both parties agree. Both rules are simple and easily understood, but may lead to very different estimates of network structure. If one must employ either $\hat{\Theta}^U$ or $\hat{\Theta}^I$, which should one use? To determine this, we consider the accuracy of each estimator under realistic conditions.

2. LAS accuracy: some basic theory

It is not immediately clear which LAS method would provide a more accurate estimate of the unknown graph. If informants uniformly make more false positive errors (reporting that an edge exists when it does not) relative to their false negative rate (reporting that an edge does not exist when it does) in their edge observations, then it would seem that the Intersection LAS would obtain an estimated graph close to the unknown graph. Conversely, if informants make a greater number of false negative errors relative to the false positive rate, then the Union LAS would be expected to provide a closer estimate to the unknown graph. This intuition follows from the response of the respective rules to informant error rates on a per-edge basis. Define the false positive and false negative error rates for an arbitrary informant i by

$$e^+_i = Pr(Y_{iij}=1|\Theta_{ij}=0) = Pr(Y_{iji}=1|\Theta_{ji}=0)$$
$$e^-_i = Pr(Y_{iij}=0|\Theta_{ij}=1) = Pr(Y_{iji}=0|\Theta_{ji}=1),$$

with $e=(e^+, e^-)$ being the full set of error rates. Assuming that errors occur independently, it then immediately follows that the per-edge error rates for $\hat{\Theta}^I$ and $\hat{\Theta}^U$ are given by

$$Pr(\hat{\Theta}^I_{ij} \neq \Theta_{ij}|e) = \begin{cases} e^+_i e^+_j & \Theta_{ij}=0 \\ 1-(1-e^-_i)(1-e^-_j) & \Theta_{ij}=1 \end{cases}$$
$$Pr(\hat{\Theta}^U_{ij} \neq \Theta_{ij}|e) = \begin{cases} 1-(1-e^+_i)(1-e^+_j) & \Theta_{ij}=0 \\ e^-_i e^-_j & \Theta_{ij}=1 \end{cases}.$$

where error rates are approximately equal across informants, $\hat{\Theta}^I$ false positive rates scale with the square of the individual false positive rates, with the same holding mutatis mutandis for $\hat{\Theta}^U$ and false negative rates; while this can result in substantial suppression for these types of errors, errors of the opposite type (false negatives for $\hat{\Theta}^I$, false positives for $\hat{\Theta}^U$) are correspondingly magnified. This is illustrated graphically in Fig. 1, which shows the "worst case" probabilities of a correct inference as a function of informant accuracy (with both informants assumed to have the same error rates). In the plausible setting for which $e^- > e^+$ – i.e., omission of true ties, due e.g. to forgetting, is more common than fabrication or confabulation of nonexistent ties – this analysis suggests that $\hat{\Theta}^U$ should be more accurate than $\hat{\Theta}^I$ (perhaps by a large margin).

There is, however, another aspect to this problem. Consider the expected total (Hamming) errors for $\hat{\Theta}^I$ and $\hat{\Theta}^U$ given the true graph state:

$$\mathbf{E}\sum_{(i,j)\in\mathcal{D}} \left|\hat{\Theta}^I - \Theta\right| = \sum_{(i,j)\in\mathcal{D}} \left[\Theta_{ij}\left(e^-_i + e^-_j - e^-_i e^-_j\right) + \left(1-\Theta_{ij}\right) e^+_i e^+_j\right] \tag{3}$$

$$\mathbf{E}\sum_{(i,j)\in\mathcal{D}} \left|\hat{\Theta}^U - \Theta\right| = \sum_{(i,j)\in\mathcal{D}} \left[\Theta_{ij} e^-_i e^-_j + \left(1-\Theta_{ij}\right)\left(e^+_i + e^+_j - e^+_i e^+_j\right)\right] \tag{4}$$

where $\mathcal{D}$ is the set of potential edges. As a simplifying assumption, let us take all informants to have the same error rates (hence