National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

08-19

Inference for Exponential-Family Random Graph Models from Egocentrically-Sampled Data with Alter–Alter Relations

Pavel N. Krivitsky, Martina Morris, and Michał Bojanowski

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998. Email: karink@uow.edu.au

Inference for Exponential-Family Random Graph Models from Egocentrically-Sampled Data with Alter–Alter Relations

Pavel N. Krivitsky, University of Wollongong
Martina Morris, University of Washington
Michał Bojanowski,[*] Kozminski University

2019-07-14

Abstract

Egocentric design is probably the most frequently used survey method for studying large offline social networks. There are several methods of estimating network descriptives from such data, but they lack a unified and rigorous statistical foundation for estimation and inference. Recent work has shown that data collected through an egocentric design can be used to estimate certain specifications of Exponential-family Random Graph Models (ERGM). In particular, data about the egos' immediate connections can be used to estimate models with nodal, degree, and assortative and disassortative mixing effects. In this work, we present methods of estimating triadic effects given that the data contains information about tie presence/absence between the alters. We also show that nodal effects of covariates can be estimated both with and without observing them on alters. Models are estimated through the pseudo-maximum-likelighood approach. Presented methodology is illustrated with an application to data on networks of discussing important matters collected in General Social Survey in 2004.

1 Introduction

Traditional methods used for analyzing networks required conducting a "sociocentric" network census: data on all nodes and links in the population of interest. This design arises naturally when the goal is to provide descriptive summaries of whole network properties like centrality/centralization, geodesics, paths and reachability, cliques/clusters and various forms of positional equivalence. But it imposes an onerous burden for empirical research, as the collection of census data is not feasible in many empirical settings. Estimation of network properties from sampled network data requires a statistical framework.

Two forms of network sampling have received some attention in recent years years: adaptive (link tracing) and egocentric sampling. Both designs require an initial sample of respondents, a "name generator" to enumerate their contacts, and a strategy for sampling these contacts. Adaptive designs start with a small sample of "seeds" from the population of interest, and employ a range of different link tracing strategies to recruit the contacts into the study. Variants include snowball samples, random walks, respondent driven samples (RDS) and a host of ad hoc designs. ERGM estimation based on networks observed with these designs has been developed by Handcock and Gile (2010), Koskinen, Robins, and Pattison (2010), Pattison et al. (2013), and others.

Egocentric designs employ standard survey sampling techniques to recruit respondents (egos), who then report on their contacts (alters); the alters are not recruited into the sample. These, too, have a range of design variations based on the enumeration and description of alters, and the possible

[*] MB thanks ICM UW for support through computational grant G74-3.

collection of alter–alter tie information. These designs reduce the burden of data collection, but also constrain the network information available in the sample, and thus the inferences that can be drawn from the data. In a previous paper, Krivitsky and Morris (2017) (which we abbreviate as KM throughout this work) presented a rigorous approach to estimation and inference of a subclass of Exponential-family Random Graph Models (ERGMs) from egocentrically sampled network data. In their framework, the primary condition for the subclass of models is that the model terms are recoverable from an egocentric census of the same design. The class of models developed in that paper were defined by the "minimal" egocentric sampling design—where the egos report on the number and attributes of their alters, but the alter–alter ties are not elicited. That sampling design makes it possible to estimate attribute-based degree heterogeneity and mixing, but not triadic effects. The paper showed that triadic closure terms should also be estimable provided that the data contain information about the relationships between the alters.

In this paper we present a framework for estimation and inference of triadic effects from egocentrically sampled network designs that include collection of the alter–alter matrix. The framework builds on the design-based approach presented by KM, with terms that are recoverable from an egocentric census with this additional information. We evaluate the properties of the estimates with a simulation study, and demonstrate their use in an application that analyzes data from the 2004 General Social Survey in the US, which included questions about friendship discussion networks.

1.1 Review of the General Framework

The inferential framework here is based on estimating the pseudo-MLE (PMLE) of the coefficients of an ERGM, using a design-based approach for estimating the sufficient statistics from the sampled data (KM). Below we briefly review the key components of this framework and introduce the notation needed.

1.1.1 Exponential-Family Random Graph Models (ERGMs)

ERGMs (Frank and Strauss 1986; Wasserman and Pattison 1996; Hunter and Handcock 2006, and many others), express the probability of an observed graph $\boldsymbol{y}$ as an exponential family:

$$\Pr_{\boldsymbol{g}}(\boldsymbol{Y}=\boldsymbol{y};\boldsymbol{x},\boldsymbol{\theta}) \equiv \exp\{\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})\}/\kappa_{\boldsymbol{g}}(\boldsymbol{\theta},\boldsymbol{x}),\ \boldsymbol{y}\in\mathcal{Y}. \tag{1}$$

It is specified by the sample space $\mathcal{Y}$ of possible networks (configurations of relationships) and a sufficient statistic vector $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})$, which is a function of the whole network $\boldsymbol{y}$ and possible covariates $\boldsymbol{x}$, and whose elements are selected to represent features of the network that are of substantive interest or believed relevant to the generative process of the relationships in the network; and it is parametrized by its vector of natural parameters $\boldsymbol{\theta}$. The normalizing constant $\kappa_{\boldsymbol{g}}(\boldsymbol{\theta},\boldsymbol{x}) \equiv \sum_{\boldsymbol{y}'\in\mathcal{Y}} \exp\{\boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y}',\boldsymbol{x})\}$ is usually intractable when the choice of $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})$ induces dependence among the relationship states.

Estimating $\boldsymbol{\theta}$ facilitates inference about the social forces that shape the network as well as principled simulation of complete networks whose features are similar, on average, to those of the network observed. In the case of sampled network data in particular, it would allow recovering possible full networks from which the sample may have been drawn. Therefore, $\boldsymbol{\theta}$ is our target of inference.

1.1.2 Survey sampling and sufficiency perspective

Exploiting the property of *sufficiency*—that one does not need to observe the whole network $\boldsymbol{y}$ to fit an ERGM, just its *sufficient statistic* vector $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})$—KM showed that an ERGM all of whose sufficient statistics $g_k(\cdot,\cdot)$ were expressible as the sum over all actors in the network of what each actor could report in an egocentric sample, i.e.,

$$g_k(\boldsymbol{y},\boldsymbol{x}) \equiv \sum_{i\in N} h_k(\boldsymbol{e}_i), \tag{2}$$

for some function $h_k(\cdot)$ of an ego's report, rigorous inference for $\boldsymbol{\theta}$ was possible. This was done in two steps:

1. Design-based inference is used to estimate $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})$ from by taking a sample mean of $\boldsymbol{h}(\boldsymbol{e}_i)$ values for $i \in S$ to get $\tilde{\boldsymbol{h}}(\boldsymbol{e}_S)$ and scaling it up to obtain $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S) \equiv |N|\tilde{\boldsymbol{h}}(\boldsymbol{e}_S)$.
2. An ERGM is fit to the estimated $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})$ to obtain $\tilde{\boldsymbol{\theta}}$. KM showed that the properties of the estimator $\tilde{\boldsymbol{g}}(\boldsymbol{e}_S)$ for $\boldsymbol{g}(\boldsymbol{y},\boldsymbol{x})$ can be transferred to $\tilde{\boldsymbol{\theta}}$.

ERGM statistics amenable to this type of analysis for data with no alter–alter relations included *dyadic-independent* Hunter et al. (2008) statistics that can be expressed in the general form of $g_k(\boldsymbol{y},\boldsymbol{x}) = \sum_{(i,j)\in\boldsymbol{y}} f_k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for some symmetric function $f_k(\cdot,\cdot)$ of two actors' attributes; and some *dyadic-dependent* statistics that can be expressed as $g_k(\boldsymbol{y},\boldsymbol{x}) = \sum_{i\in N} f_k\{\boldsymbol{x}_i,(\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}\}$ for some function $f_k(\cdot,\cdots)$ of the attributes of an actor and their network neighbors. A sampling of them is given in Table 1.

1.2 Notation for Egocentric Data

To facilitate generalizing the egocentric estimation to triadic configurations, we review the notation needed for egocentrically sampled data. We generally follow the notation of KM where possible, extending it to data containing alter–alter relations. To emphasize the difference in indexing of actors in the network as a whole and indexing of actors nominated by an ego, we use a prime (e.g., "i'") to identify the latter.

1.2.1 Population network

Let $N = \{1,\ldots,|N|\}$ be the set of actors whose relations are of interest. Each actor $i \in N$ has a vector of individual attributes $\boldsymbol{x}_i$, like age and socioeconomic status, and we use $\boldsymbol{x}_N$ (or just $\boldsymbol{x}$ when there is no ambiguity) to refer to the collection of such information about all actors in the population of interest. Further let $\mathbb{Y}(N) \equiv \{\{i,j\} : (i,j) \in N\times N \wedge i \neq j\}$ (distinct unordered pairs of actors) be the set of *dyads* (potential ties) in an undirected network of these actors. Then, let $\mathcal{Y}(N,\boldsymbol{x}) \subseteq 2^{\mathbb{Y}(N)}$ (set of subsets of potential ties) be the set of networks (sets of ties) of interest. $\mathcal{Y}(\cdot,\cdot)$ may incorporate exogenous constraints. For a network $\boldsymbol{y} \in \mathcal{Y}(N,\boldsymbol{x})$, let $y_{i,j} \equiv y_{j,i}$ be an indicator function of whether a tie between i and j is present in $\boldsymbol{y}$ and $\boldsymbol{y}_i = \{j \in N : y_{i,j} = 1\}$, the set of i's network neighbors, with $|\boldsymbol{y}_i|$ being their number,

Throughout, $\boldsymbol{y}$ will refer to what we will call the *population network*: a fixed but unknown network of relationships of interest.

1.2.2 Egocentric data

Let "$i{:}j'$" for $i \in N$ and $j' \in \{1,\ldots,|\boldsymbol{y}_i|\}$ be the index in N of j'th alter nominated by i.

Then, following KM, let $\boldsymbol{e}_i$ be the "egocentric" view of network $\boldsymbol{y}$ from the point of view of actor i ("ego"). It comprises $\boldsymbol{e}_i^{\mathrm{e}} \equiv \boldsymbol{x}_i$: i's own attributes, and $\boldsymbol{e}_i^{\mathrm{a}} \equiv (\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}$: a list of attribute vectors of i's immediate neighbors ("alters"), but usually their identities (indices in N). Then, j'th element of $\boldsymbol{e}_i^{\mathrm{a}}$, $\boldsymbol{e}_{i,j'}^{\mathrm{a}} \equiv \boldsymbol{x}_{i:j'}$. Then, the kth attribute/covariate observed on ego i and its alters as $e_{i,k}^{\mathrm{e}} \equiv x_{i,k}$

and $e^{\mathrm{a}}_{i,j',k} \equiv x_{i:j',k}$. In a mild abuse of notation, we will use "$|\boldsymbol{e}^{\mathrm{a}}_i|$" to refer to the number of alters nominated by i.

This was the framework described by KM for the information retained in a single egocentric observaton. We describe two extensions here.

1.2.3 Alter–alter information

In many egocentric studies, including the US General Social Survey (GSS) (Burt 1984), an ego is also asked to identify which of the alters does he or she believe to have the relationship of interest. If such data are collected, we define $\boldsymbol{e}^{\mathrm{aa}}_{i:j',l'} \equiv y_{i:j',i:l'}$. Note that, as before, j' and l' are local to i, as they are not globally identified for most surveys (including the GSS).

We may also include observations of alter–alter tie or non-tie attributes ($\boldsymbol{e}^{\mathrm{aa}}_{i:j',l',k}$), but we leave its development for future work.

1.2.4 Attributes observed only on egos or only on alters

Typically, in large surveys (e.g., Laumann et al. 1992; Burt 1984; National Survey of Family Growth Staff, n.d.) the information collected about egos is quite extensive, while the network section of the questionnaire is relatively brief. This means that the data collected on the attributes of alters will be a (small) subset of the attributes collected for egos.

We do not introduce special notation for this, except to specify that $e^{\mathrm{e}}_{i,k}$ and $e^{\mathrm{a}}_{i,j,k}$ for the same k will refer to the same attribute observed on both.

1.2.5 Egocentric census and sample

Given the above notation, we define, $(\boldsymbol{e}_i)_{i\in N}$ ($\boldsymbol{e}_N$ for short) to be the *egocentric census*, the information retained in an egocentric sample, and information about $\boldsymbol{y}$ contained in an *egocentric sample* of actors $S \subseteq N$ can then be represented as $\boldsymbol{e}_S \equiv (\boldsymbol{e}_i)_{i\in S}$.

2 Types of egocentric statistics

The egocentric statistics are reproduced from that of KM and augmented by the developments here are given in Table 1.

2.1 Actor-level attribute effects

Actor attributes inform a range of dyad-independent terms for degree heterogeneity and assortative mixing.

2.1.1 Covariates observed on both egos and alters

This case was discussed by KM. Data of this type gives rise to statistics that can be expressed in the general form of $g_k(\boldsymbol{y},\boldsymbol{x}) = \sum_{(i,j)\in\boldsymbol{y}} f_k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ for some symmetric function $f_k(\cdot,\cdot)$ of two actors' attributes; and some *dyadic-dependent* statistics that can be expressed as $g_k(\boldsymbol{y},\boldsymbol{x}) = \sum_{i\in N} f_k\{\boldsymbol{x}_i,(\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}\}$ for some function $f_k(\cdot,\cdots)$ of the attributes of an actor and their network neighbors.

Table 1: Examples of egocentric statistics for undirected networks. $x_{i,k}$ may be a dummy variable indicating i's membership in a particular exogenously defined group. $h_k(\boldsymbol{e}_i)$ that sum over ties may be halved because each tie is observed egocentrically twice: once at each end.

Statistic	Source	$g_k(\boldsymbol{y}, \boldsymbol{x})$	$h_k(\boldsymbol{e}_i)$
General sum over ties	KM	$\sum_{(i,j)\in\boldsymbol{y}} f_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$	$\frac{1}{2}\sum_{j'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert} f_k\left(\boldsymbol{e}_i^{\mathrm{e}}, \boldsymbol{e}_{i,j'}^{\mathrm{a}}\right)$
Number of ties in the network	KM	$\lvert\boldsymbol{y}\rvert \equiv \sum_{(i,j)\in\boldsymbol{y}} 1$	$\frac{1}{2}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert$
weighted by actor covariate $x_{i,k}$	KM	$\sum_{(i,j)\in\boldsymbol{y}}(x_{i,k} + x_{j,k})$	
covariate observed on alter	KM		$\frac{1}{2}(e_{i,k}^{\mathrm{e}}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert + \sum_{j'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert} e_{i,j',k}^{\mathrm{a}})$
covariate not observed on alter	Sec. 2.1.2		$e_{i,k}^{\mathrm{e}}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert$
weighted by difference in $x_{i,k}$	KM	$\sum_{(i,j)\in\boldsymbol{y}}\lvert x_{i,k} - x_{j,k}\rvert$	$\frac{1}{2}\sum_{j'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert}\lvert e_{i,k}^{\mathrm{e}} - e_{i,j',k}^{\mathrm{a}}\rvert$
within groups identified by $x_{i,k}$	KM	$\sum_{(i,j)\in\boldsymbol{y}}\mathbb{I}(x_{i,k} = x_{j,k})$	$\frac{1}{2}\sum_{j'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert}\mathbb{I}(e_{i,k}^{\mathrm{e}} = \boldsymbol{e}_{i,j',k}^{\mathrm{a}})$
General sum over actors	KM	$\sum_{i\in N} f_k\left\{\boldsymbol{x}_i, (\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}\right\}$	$f_k\left(\boldsymbol{e}_i^{\mathrm{e}}, \boldsymbol{e}_i^{\mathrm{a}}\right)$
Number of actors with d neighbors	KM	$\sum_{i\in N}\mathbb{I}(\lvert\boldsymbol{y}_i\rvert = d)$	$\mathbb{I}(\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert = d)$
weighted by actor covariate $x_{i,k}$	KM	$\sum_{i\in N} x_{i,k}\,\mathbb{I}(\lvert\boldsymbol{y}_i\rvert = d)$	$e_{i,k}^{\mathrm{e}}\,\mathbb{I}(\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert = d)$
"Transitive ties"	KM	$\sum_{(i,j)\in\boldsymbol{y}}\max_{l\in N} y_{i,l}y_{l,j}$	$\frac{1}{2}\sum_{j'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert}\max_{l'\in\{1,\ldots,\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert\}}\boldsymbol{e}_{i:j',l'}^{\mathrm{aa}}$
Number of ties with d shared partners	Sec. 2.2.1	$\sum_{(i,j)\in\boldsymbol{y}}\mathbb{I}(\sum_{l\in N} y_{i,l}y_{l,j} = d)$	$\frac{1}{2}\sum_{j'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert}\mathbb{I}(\sum_{l'=1}^{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert}\boldsymbol{e}_{i:j',l'}^{\mathrm{aa}} = d)$

2.1.2 Covariates observed only on egos or only on alters

As mentioned in Section 1.2.4, it is often the case that not all of the information collected about the egos is also collected about the alters. We discuss the implications of this here.

The effects that can be estimated are limited, as always, by what is observed in the data. For example, estimating effects of homophily for income would require the income of both ego and alter to be observed.

However, actor-level effects are still estimable. Consider the effect of actor covariate $x_{i,k}$, i.e., $g_k(\boldsymbol{y}, \boldsymbol{x}) = \sum_{(i,j)\in\boldsymbol{y}} (x_{i,k} + x_{j,k})$. We require an $h^{\mathrm{e}}_k(\boldsymbol{e}_i)$ having the property (2): that if we sum its value over all of the egos in the network, we recover the original statistic.

If covariate $x_{i,k}$ is observed on the egos only, $h^{\mathrm{e}}_k(\boldsymbol{e}_i) \equiv e^{\mathrm{e}}_{i,k}|\boldsymbol{e}^{\mathrm{a}}_i|$ has this property. Similarly, if such a covariate were observed on the alters but not on the egos, it could be recovered using $h^{\mathrm{a}}_k(\boldsymbol{e}_i) \equiv \sum_{j'=1}^{|\boldsymbol{e}^{\mathrm{a}}_i|} e^{\mathrm{a}}_{i,j',k}$. Unlike the $h_k(\cdot)$ from KM reproduced in Table 1, these are not divided by two, since each dyad is only counted once. The disadvantage of this approach is that the variance of the resulting estimates will almost certainly be higher.

2.2 Effects based on alter–alter ties

Observing alter–alter ties expand the scope of effects that can be modelled, but they also have limitations. One is respondent accuracy: an ego is more likely to err when reporting relations among alters than their own relations. Another is a form of censoring: whereas an ego i can observe all of their own ties, i cannot observe ties between i's neighbors and i's non-neighbors. For this reason, alter–alter information with unidentified alters is actually more limited than it appears at first glance.

In particular, alter–alter ties alone cannot be used to recover such features as network density. For an illustration, consider a network with 4 actors and 4 ties in a four-cycle: 1–2–3–4–1. From Table 1, $h_{\mathrm{edges}}(\boldsymbol{e}_i) = \frac{1}{2}|\boldsymbol{e}^{\mathrm{a}}_i|$, and each ego would report 2 alters, which yiels a contribution of $2/2 = 1$, and therefore (2) will recover the the network edge count. On the other hand, each ego would report *no* alter–alter ties, same as for an empty network. Thus, it is not possible to construct $h^{\mathrm{aa}}_{\mathrm{edges}}(\boldsymbol{e}^{\mathrm{aa}}_i)$ that fulfills (2). This does not mean that there is no information about network density in alter–alter ties, but it does mean that it cannot be used directly.

2.2.1 Triadic effects

2.2.1.1 Edgewise Shared Partner statistics

For egocentric surveys that solicit alter–alter information, it is straightforward to recover the alternating k-triangle (Snijders et al. 2006) or their reparametrization, the Geometrically-Weighted Edgewise Shared Partner (GWESP) (Hunter 2007) statistics. (Smith (2012) also points this out, and Krivitsky and Kolaczyk (2015) and KM provide a special case.) Their sufficient statistic is

$$g_{\mathrm{ESP}(k)}(\boldsymbol{y}) = \sum_{(i,j)\in\boldsymbol{y}} \mathbb{I}\left(\sum_{l\in N\backslash\{i,j\}} y_{i,l}y_{l,j} = k\right) \quad \text{for } k = 1, \ldots, |N| - 2.$$

This statistic can be recovered by observing that in an egocentric census, every tie will appear twice as an ego–alter report. Using it as the base of the k-triangle, the ties to other alters would form the

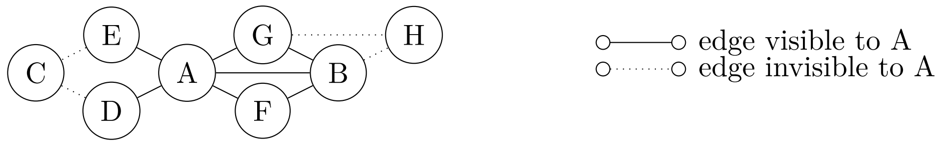

Figure 1: Edgewise and dyadwise shared partners. Observe that actor A is able to report the number of shared partners on edges A–B (2), A–F (1), and A–G (1), and is also able to report with confidence that no other shared partners are present on these edges, because if they were, they would also be partners of A and therefore visible to A. However, as C is not a neighbour of A, A cannot know the number of shared partners across the A–C dyad. In fact, no actor in this network is in the position to count them: C is in the same situation as A, and D is not aware of E and vice versa. Also worth noting is that although A can see the G–B edge, A has no way of knowing whether there are any additional shared partners on that edge (e.g., H) and therefore cannot report the shared partner count on it. However, either B or G could.

first segoment of the 2-path and the alter–alter ties the second, yielding the following expression:

$$h_{\mathrm{ESP}(k)}(\boldsymbol{e}_i) = \frac{1}{2} \sum_{j'=1}^{|\boldsymbol{e}_i^{\mathrm{a}}|} \mathbb{I}\left(\sum_{l'=1}^{|\boldsymbol{e}_i^{\mathrm{a}}|} \boldsymbol{e}_{i:j',l'}^{\mathrm{aa}} = k \right).$$

Figure 1 illustrates this.

Aside: Triangle count Although generally not recommended for use in ERG modelling (Schweinberger 2011; Snijders et al. 2006), the triangle count can be recovered from alter–alter data by observing that $g_{\mathrm{triangle}}(\boldsymbol{y}) = \frac{1}{3}\sum_{k=1}^{n-2} g_{\mathrm{ESP}(k)}(\boldsymbol{y}) \times k$.

2.2.1.2 Dyadwise Shared Partner statistics

On the other hand, Dyadwise Shared Partners (DSP) or k-two-path statistics

$$g_{\mathrm{DSP}(k)}(\boldsymbol{y}) = \sum_{(i,j)\in\mathbb{Y}} \mathbb{I}\left(\sum_{l\in N\setminus\{i,j\}} y_{i,l} y_{l,j} = k \right)$$

or Nonedge-wise Shared Partners (NSP)

$$g_{\mathrm{NSP}(k)}(\boldsymbol{y}) = \sum_{(i,j)\in\mathbb{Y}\setminus\boldsymbol{y}} \mathbb{I}\left(\sum_{l\in N\setminus\{i,j\}} y_{i,l} y_{l,j} = k \right)$$

statistics are far more challenging, even with alter–alter data.

Whereas ESP statistics are computed using ego–alter nominations as bases, the DSP and NSP statistics must also consider shared partners on non-tie bases, but even using the non-ties in the alter–alter matrix would miss shared partners which are not directly connected to the ego. Figure 1 provides an example of this as well. Approaches to approximating these statistics are subject for future research.

3 Estimation

In some cases, the population network size is very large or even unknown. For example, the population for the US General Social Survey (Burt 1984) is in the hundreds of millions. In these circumstances, it is desirable to estimate a model that is *invariant* to network size: whose substantive results do not depend on how large the population may be.

3.1 Network size effects for dyadic-independent terms

To facilitate estimation when $|N|$ is very large, KM used the network size adjustment of Krivitsky, Handcock, and Morris (2011): an ERGM with an "offset" term having the form $\Pr_{\boldsymbol{g}}(\boldsymbol{Y} = \boldsymbol{y}; \boldsymbol{x}, \boldsymbol{\theta}) \propto \exp\{-\log(|N|)|\boldsymbol{y}| + \boldsymbol{\theta}^\top \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{x})\}$ preserves degree distribution and observed covariate effects across a variety of network sizes (where without the offset, the density is preserved). They thus constructed the "microcosm" N' of the target population, then adjusted the estimate to be applicable. This adjustment is invariant to network size only for dyad-independent models.

3.2 Network size effects for dyadic-dependent terms

For higher-order statistics, the simple adjustment of $-\log(|N|)|\boldsymbol{y}|$ fails. In particular, Krivitsky and Kolaczyk (2015) showed that for mutuality, using this adjustment stabilizes mean degree as desired, but mutuality, which requires a "coincidence" of two relations, vanishes. (An adjustment was proposed to correct this.)

For triadic effects, Krivitsky and Kolaczyk (2015) observed that the $-\log(|N|)|\boldsymbol{y}|$ adjustment is effectively a penalty, reflecting limited opportunity of a given pair of actors (e.g., i and j) to meet in an increasingly large network, and that such a penalty should not apply if a third actor, say, l, has ties to both of them and can therefore "introduce" them.

The statistic that operationalizes this effect is the *transitive ties* statistic $g_{\text{tr. ties}}(\boldsymbol{y}) = \sum_{(i,j)\in\boldsymbol{y}} \max_{l\in N\setminus\{i,j\}} y_{i,l}y_{l,j}$, evaluating the number of ties in the network for which at least one shared partner (i.e., the "introducer") exists. (We use this name for historical reasons—undirected network do not distinguish between transitivity and cyclicality of triads.) Because the closing tie creates not one but three transitive ties, they suggested that the coefficient on this term be $+1/3$, resulting in the following offset: $\log(|N|)\{-|\boldsymbol{y}| + g_{\text{tr. ties}}(\boldsymbol{y})\}$. In a simulation study, they demonstrated that for large, sparse networks of a variety of sizes but with constant level of transitive closure (from the point of view of each ego), this relationship held.

4 Parameter recovery simulation study

In this section we demonstrate that our methodology can accurately estimate triadic effects introduced in Section 2 and quanitfy their uncertainty. The simulations use the "Faux Magnolia High" dataset available in package `ergm` (Handcock et al. 2019; Hunter et al. 2008) having 1461 nodes and 974 edges. Network nodes are characterized with the following attributes: Grade (from 7 to 12), Race (Asian, Black, Hispanic, Native American, White, or Other), and Sex.

In order to test whether the estimates based on an egocentric sample are accurate we will first fit an ERGM to the complete network. Second, we will draw 1000 independent egocentric samples of size 730, which is half of the size of the complete network. Third, to each egocentric sample we will fit an egocentric ERGM having the same terms as the complete network model. Finally, fourth, we

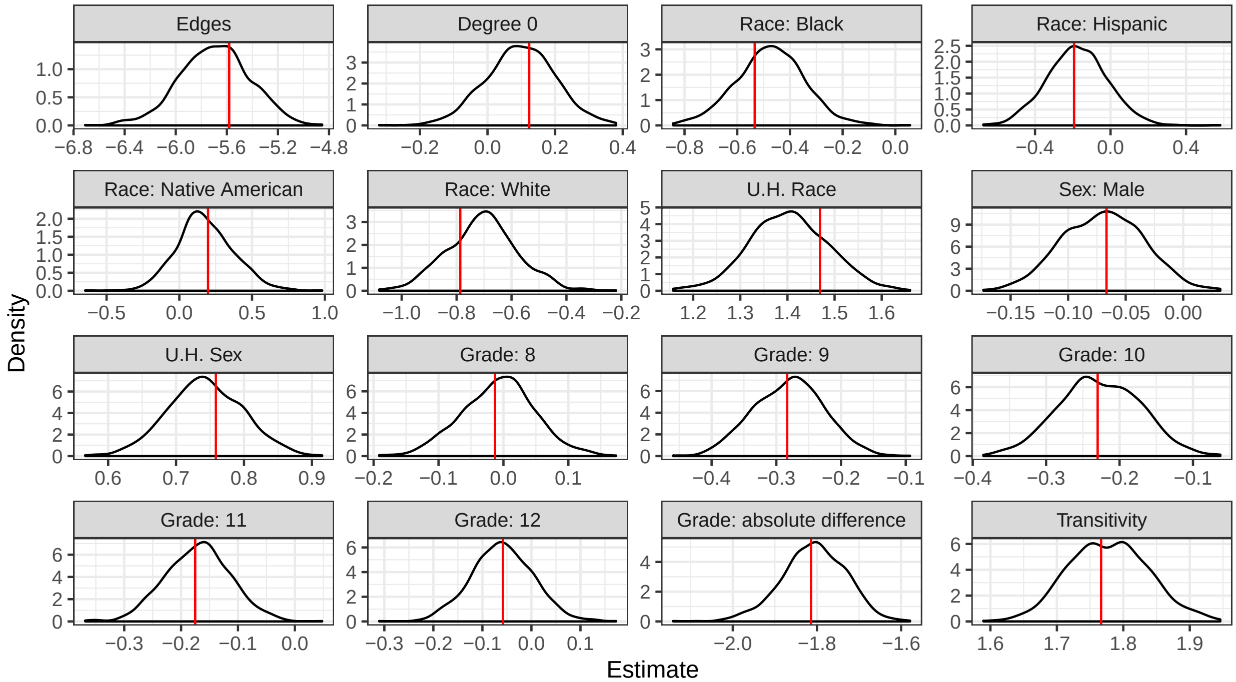

Figure 2: Density estimates for distributions of egocentrically estimated coefficients (black curves) and the 'true values from the complete network model (vertical red lines).

will compare the distributions of the paramaters of the egocentric models to the "true" values of the parameters of the complete network model.

The specification that we use includes the main effects of all the nodal attributes, uniform homophily effects for Race and Sex, absolute difference in Grade, number of isolates, and triadic closure effect restricted to the first edgewise-shared partner. Figure 2 shows density estimates of the distributions of coefficient values across 1000 ego samples (in black) together with the "true" values of these coefficients from the model fit to the complete network (in red). We can see noticable bias for edges term, main effects of Race: Black, and Race: White. Point estimate bias and confidence interval coverage are summarized in Figures 3 and 4 respectively.

Largest biases in point estimates are associated with the terms modeling the effect of Race. The most biased is the uniform homophily effect of Race. It is equal to almost 80% of the simulated standard deviation of $\tilde{\theta}$, which corresponds to about 64% of mean squared error of the estimator. Main effects for White (vs Asian) and Black (vs Asian) are also relatively large.

Simulated standard errors overestimate true standard errors by no more than 25% for the majority of parameters. Larger biases are for the number of isolates, homophily parameter for sex, and triadic closure.

5 Application to General Social Survey Data

Respondents of the General Social Survey conducted in 2004 were asked with whom do you discuss important matters (Burt 1984)? We use answers to this question along with additional variables describing the respondents (egos), their nominated peers (alters), and reports on ties between alters

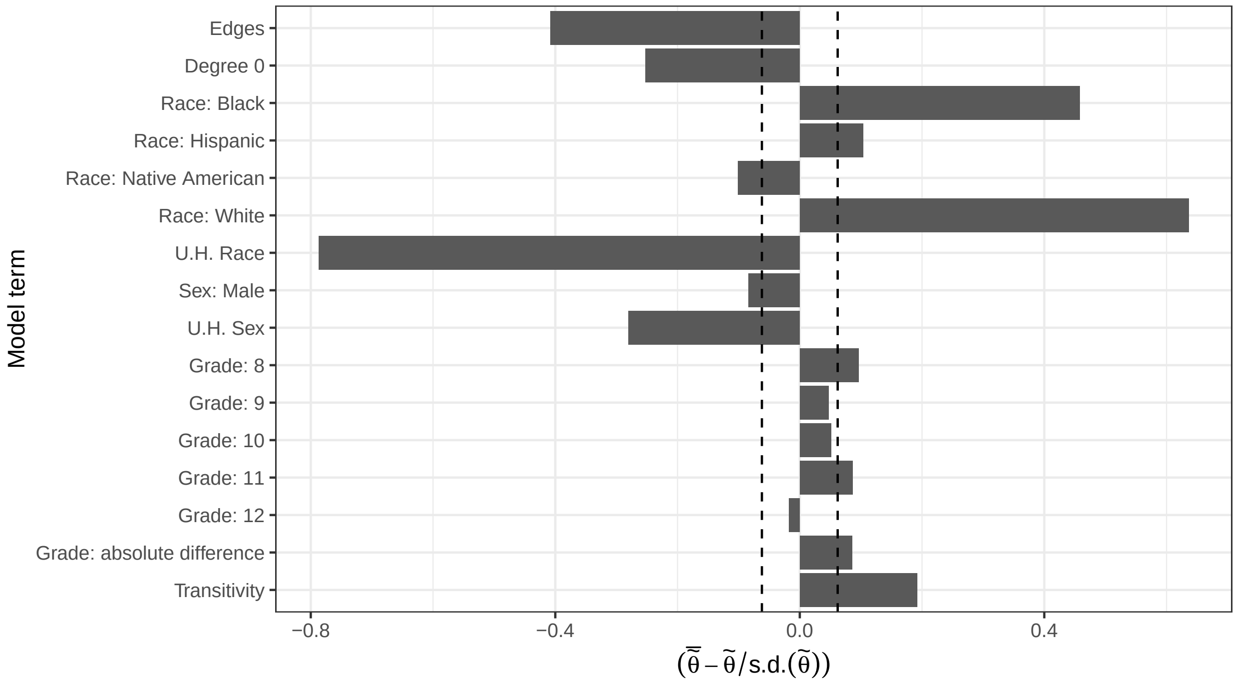

Figure 3: Simulated bias in the point estimates relative to the simulated standard deviation. Dashed lines are 5% critical values: 95% of observations should fall between them if the estimator is unbiased.

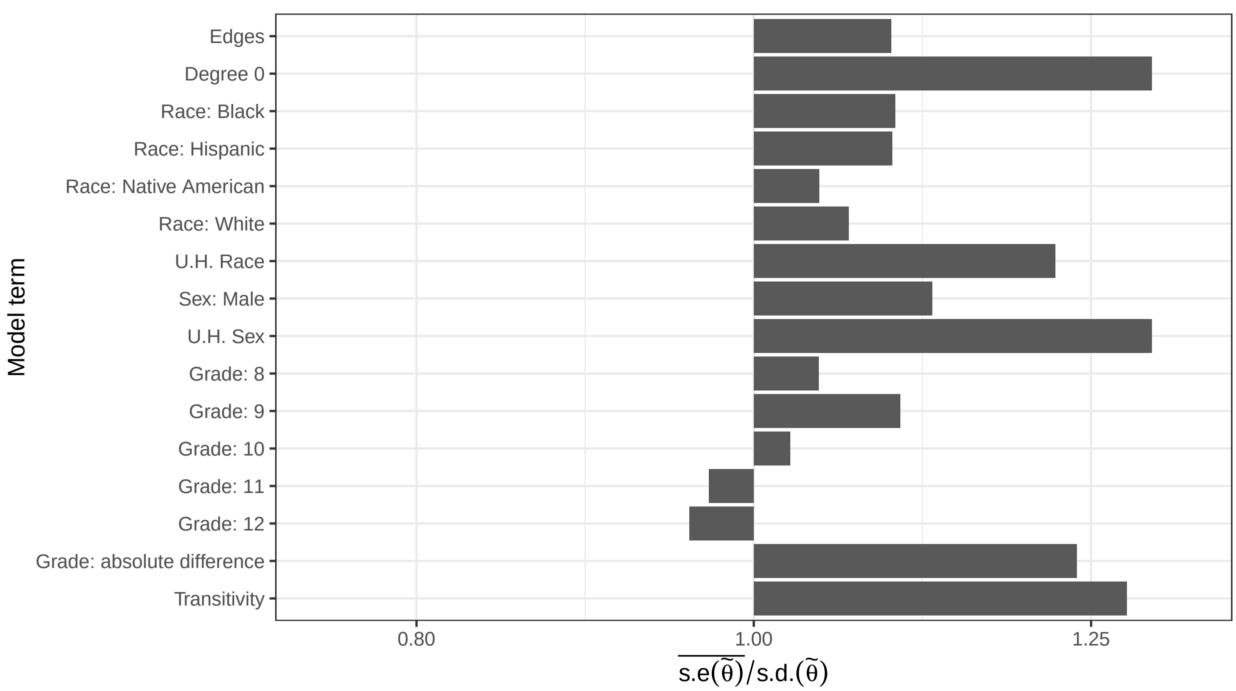

Figure 4: Simulated 95% confidence interval coverage relative to the miss probability 5%. Dashed lines are 5% critical values: 95% of values should fall between them if the estimator has nominal coverage. The scale of the horizontal axis is logged.

to illustrate proposed models.

5.1 Data

The GSS 2004 data provide information on 1472 respondents. Each respondent could name at most 5 alters as peers in discussing important matters. There are in total 3075 alters among which 2838 are valid valid nominations and 237 are non-responses, i.e. respondent nominated an alter but did not report any attributes or any alter–alter ties involving that alter. These alters are dropped from our analysis.

For every pair of nominated alters the respondent reported perceived closeness of the two as one of "total strangers", "know each other", "especially close". This brings us to an important issue of tie definition. The model we propose assumes that all ties in the data, whether ego–alter ties or alter–alter ties, are elements of the same population network (see Section 1.2.1). In contrast, the data are collected through two different survey questions: the name generator on discussing important matters (ego–alter ties) and perceived closeness (alter–alter ties). To make the data compatible with the model, we must make a decision about which type of alter–alter ties would be comparable to an ego–alter tie. Here, we define an alter–alter as present when ego considers them as "especially close", and absent otherwise. This results in a total of 1417 alter–alter ties, which consititutes 42% of all the alter–alter dyads.

We also use the following nodal attributes. They were created by harmonizing ego-level and alter-level variables as follows:

Religion Original variable for egos has 13 categories, while the variable for alters has only 5. We therefore recoded the egos' religion using the categorization used for alters consisting of categories: Protestant, Catholic, Jewish, None, and Other.

Race Alters' race was measured using 5 categories, while race of the egos was measured using only 3 categories (White, Black, and Other). Alter's race was recoded to the ego's categories.

Age Age was measured in years both for egos and alters.

Education Education of the egos was measured with years of schooling, while alters' education used 7 categories. Both measured were brought to a common scale of 5 categories: "Less than high school degree", "High school degree", "Some college", "BA/BS degree", and "Postgraduate".

Location of residence Type of area of residence was measured on the egos only, with two categories: less than 2,000 inhabitants or more than 2,000 inhabitants.

Married Measured only on egos, as married or not married.

A few cases of missing responses were imputed once using methods proposed by King, Honaker, and Scheve (2001) and implemented in software package `Amelia` (Honaker, King, and Blackwell 2011).

General Social Survey data is collected using a complex probability sample design which evolved somewhat through the years. The sampling complexities of GSS 2004 require taking into account both the stratified clustered nature of the sample of households as well as the weights that adjust the data due to variable number of adults in each household and sub-sampling of non-respondents (Smith et al. 2019, App. A). In Section 1.1.2 we have explained that in our approach the sufficient statistics are first estimated using design-based inference. This is a clear advantage as we can include all the necessary sampling-related adjustments, namely stratification and weighting, in our analyses below.

5.2 Models

We are fitting the following four specifications to the GSS data:

Model 1 Main effects of nodal attributes: Race (reference category: "White"), Sex (reference category: "Male"), Education (reference category: "Less than High School"), Religion (reference category: "Protestant"), Age (in 10s of years), Type of area of residence (reference category: "less than 2000 inhabitants"), and Married (reference category: "not married").

Model 2 It is Model 1 extended with uniform homophily effects for all the attributes apart from Size of residence and Married as they are observed only for egos. Homophily on Age is modeled as an effect of absolute difference in Age.

Model 3 It is Model 1 extended with differential homophily effects for all attributes apart from Size of residence and Married. Homophily effect related to Age is measured in the same way as in Model 2.

Model 4 It is Model 3 extended with geometrically-weighted edgewise shared parnters term (GWESP) with decay value fixed at 0.1.

All specifications include the network size adjustment offset (Krivitsky, Handcock, and Morris 2011). Estimation was performed using pseudo-population size 10,000. This ensures that respondents with the smallest weights are represented by, on average, 3 to 4 nodes in the pseudo-population network.

5.3 Results

The coefficients of the four models described above are presented jointly with 95% confidence intervals in Figure 5. Numerical values are presented in Table 2 in the Appendix.

Main effects estimated in Model 1 describe mean degree heterogeneities. We observe no statistically significant differences in the number of discussants between the categories of race, sex, religion, age, marital status or size of residence. With respect to education, people with less than high school education have on average less friends than the remaining four groups of higher educational level.

Model 2 estimates uniform homophily effects. We observe strong homophily effects for race and religion. For race the odds ratio of discussing important matters with a person of the same race is more than 7 times greater than with a person of a different race. It is approximately 5 for religion. Discussing important matters is also an interaction which is more likely between people of the same sex and of similar education and age.

In Model 3 the assumption of uniform homophily is relaxed and homophily in each group according to a nodal attribute is characterized with a separate parameter. With respect to race we observe the strongest homophily for blacks. The pattern of values of homophily coefficients seems common to ordinal variables—strongest homophily is observed for groups of people with lowest and highest education. Among the religious groups Jews are characterized with the strongest homophily. Interestingly the non-religious group is characterized with the lowest value of homophily, even lower than the mixed group "Other religions".

Finally, Model 4 adds a triadic effect GWESP. It is positive and significant which suggests that people sharring common discussion patners are, net of other effects described above, more likely to discuss important matters with each other. It is worthwhile to observe that adding the triadic term to the model reduced the values of the homophily parameters—they are lower in Model 4 than in Models 2 and 3. This is a consequence of the fact that network density in general, both global, i.e. graph-level, as well as within-group, i.e. homophily, creates triadic closure as a byproduct.

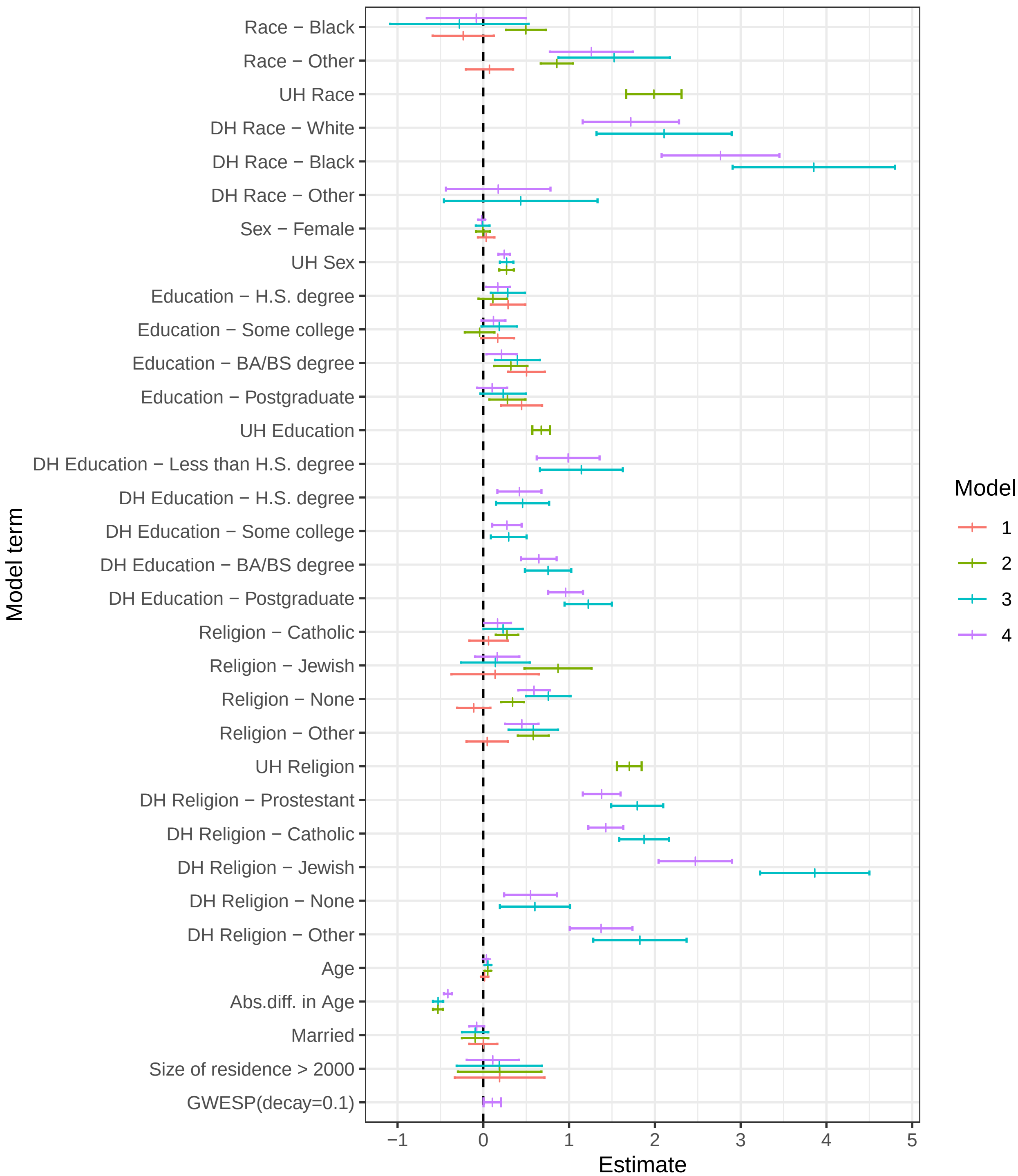

Figure 5: Coefficients of four fitted models: (1) Main effects of nodal attributes; (2) Main effects and uniform homophily effects of nodal attributes; (3) Main effects and differential homophily effects of nodal attributes; (4) Model 3 with geometrically weighted edgewise shared partners term with decay set to 0.1. UH—Uniform homophily parameter. DH—Differential homphily parameter. See Table 2 in the Appendix for tabular information.

Adding GWESP as a separate parameter in Model 4 shows that indeed there is a genuine tenendcy for triadic closure next to the homophily effects of the measured nodal attributes.

Goodness of fit can be partially assessed comparing model-simulated and observed degree distributions. This is presented in Figures 6, 7, 8, and 9. Models 1 through 3 underestimate the number of isolates and overestimate the number of people with degree 1 and 3. Model 4 captures the shape of the distribution pretty well. The largest discrepancy is for people with degree 5. This is the result of questionnaire design as the egos could list at most 5 alters. In other words, the degree distribution is right-censored at 5 due to data collection. Modeling degree censoring is the subject of the ongoing research.

6 Discussion

We have extended the framework of Krivitsky and Morris (2017) to two important use cases. On the one hand, when alter–alter relations are observed, we describe a way to recover the popular GWESP statistic and explained why GWDSP and GWNSP statistics cannot be exactly recovered from this information. On the other hand, when certain actor attributes are observed on the egos but not on the alters, we describe how it is still possible to estimate the main effects, albeit likely with less accuracy. We have conducted a simulation to demonstrate the technique and applied it to the US General Social Survey dataset.

We now discuss some limitations and future directions of this work.

6.1 Limitations of the framework

Our framework admits many, perhaps most commonly used ERGM statistics. Here, we consider some that present a particular challenge.

6.1.1 Dyadic covariates

Although estimation is possible from very limited, sampled data, ERGM estimation fundamentally requires that any dyadic covariate be observed for every dyad, whether or not it has a tie. Formally, if $g_k(\boldsymbol{y}; \boldsymbol{x}_{\cdot,\cdot,k}) = \sum_{(i,j)\in\boldsymbol{y}} x_{i,j,k}$ for some $|N| \times |N|$ covariate matrix $\boldsymbol{x}_{\cdot,\cdot,k}$, evaluation of $g_k(\boldsymbol{y}; \boldsymbol{x}_{\cdot,\cdot,k})$ only requires knowing $x_{i,j,k}$ for $(i, j) \in \boldsymbol{y}$. However, *estimating* the ERGM requires being able to evaluate $g_k(\boldsymbol{y}'; \boldsymbol{x}_{\cdot,\cdot,k})$ for all all possible networks $\boldsymbol{y}' \in \mathcal{Y}$, and hence every element $x_{i,j,k}$ for $(i, j) \in \mathbb{Y}$.

Fortunately, most attributes of dyads used as covariates in practice can be expressed as functions of actor attributes. That is, a function $f_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of actor attributes exists such that $x_{i,j,k} \equiv f_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. For example geographic distance for the dyad (i, j) can be computed as a function of those actors' locations.

However, some dyadic attributes that we may wish to use as covariates may be inherently dyadic. For example, suppose that we wished to model the tendency to have conversations on important subjects as a function of kinship (however defined). Kinship is a function of ancestry, marriage, and adoption, and while a complete genealogy of all individuals in the network could be used to derive kinship between any pair of individuals, genealogical data are dyadic in the first place, and so while a network census might observe it, an egocentric sample would not.

Even more frequently, the actor attributes needed to calculate the dyadic covariate may be unobserved, even if the dyadic covariate itself is. For example, a social survey might ask about an alter X, "Did

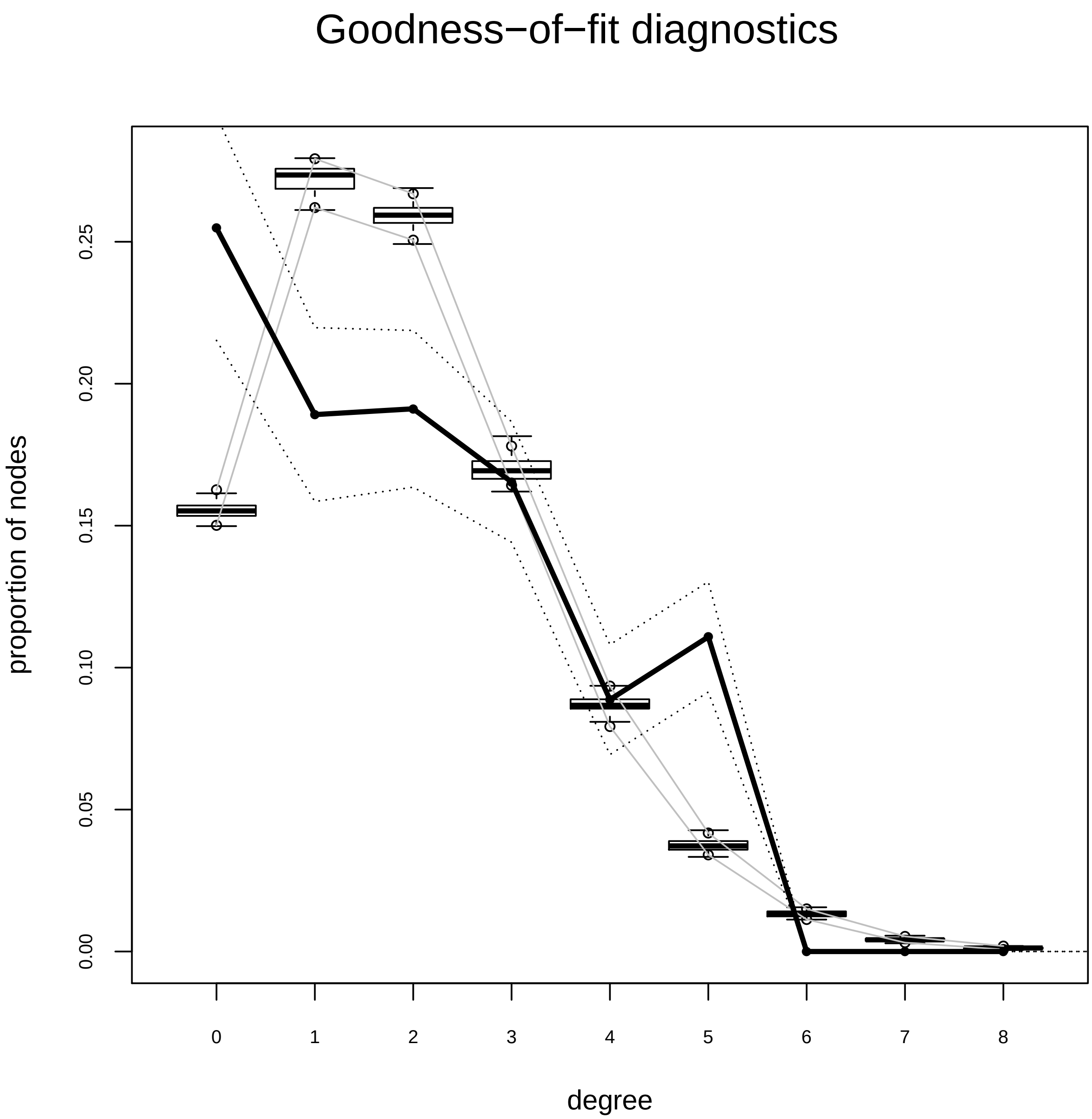

Figure 6: Degree Goodness-of-fit of Model 1. Model-simulated (boxplots) and observed (black lines) degree distributions.

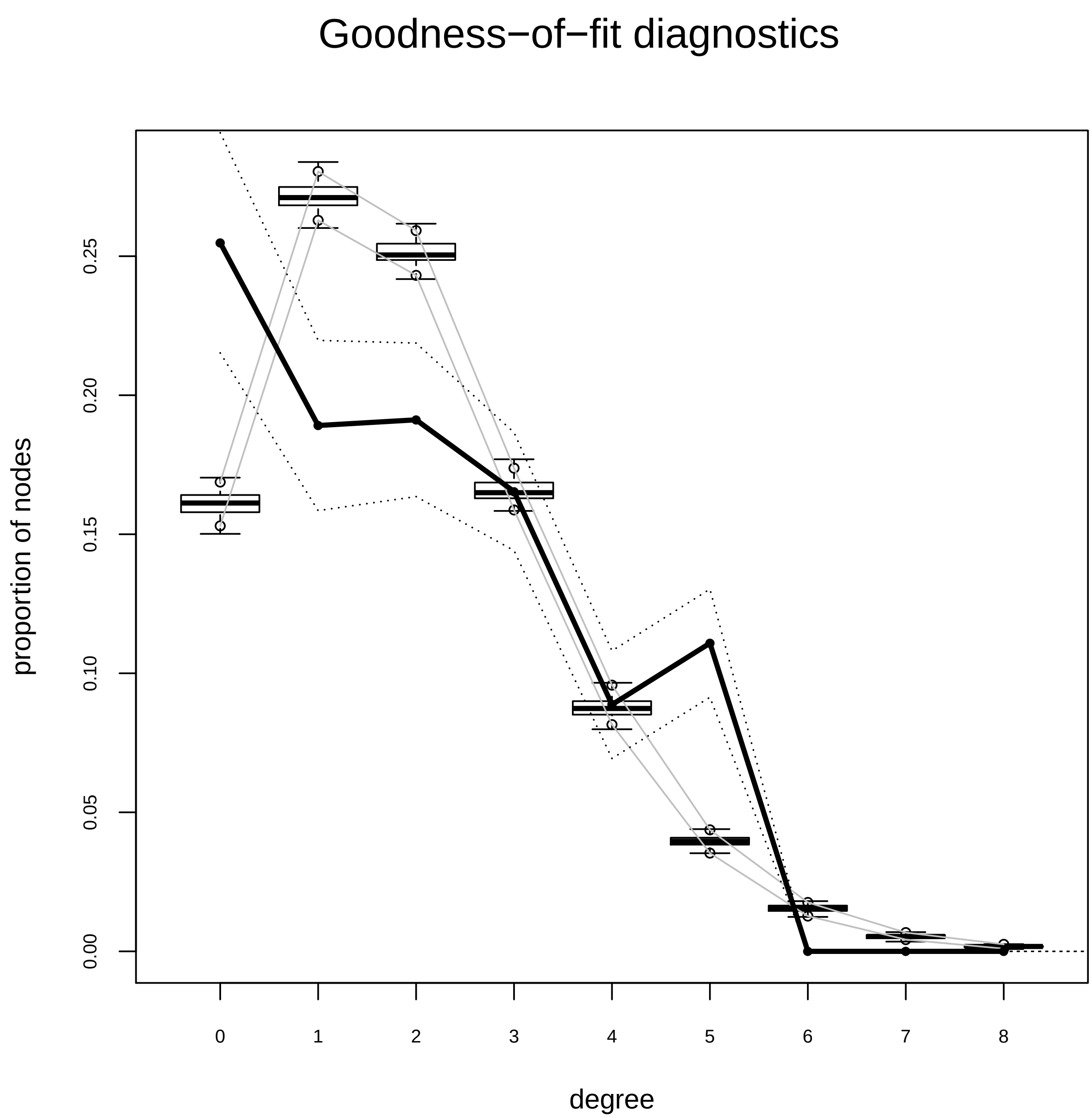

Figure 7: Degree Goodness-of-fit of Model 2. Model-simulated (boxplots) and observed (black lines) degree distributions.

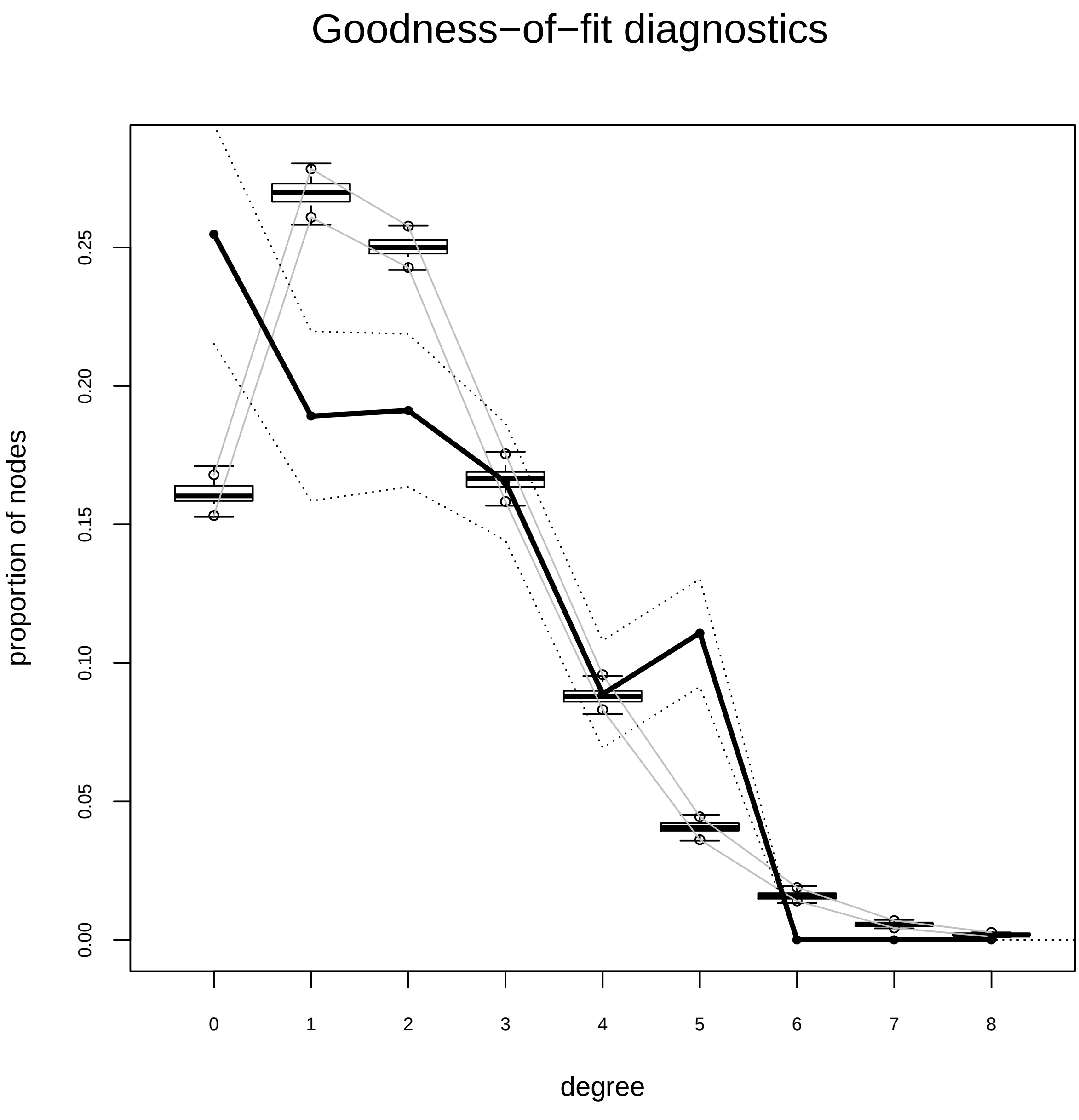

Figure 8: Degree Goodness-of-fit of Model 3. Model-simulated (boxplots) and observed (black lines) degree distributions.

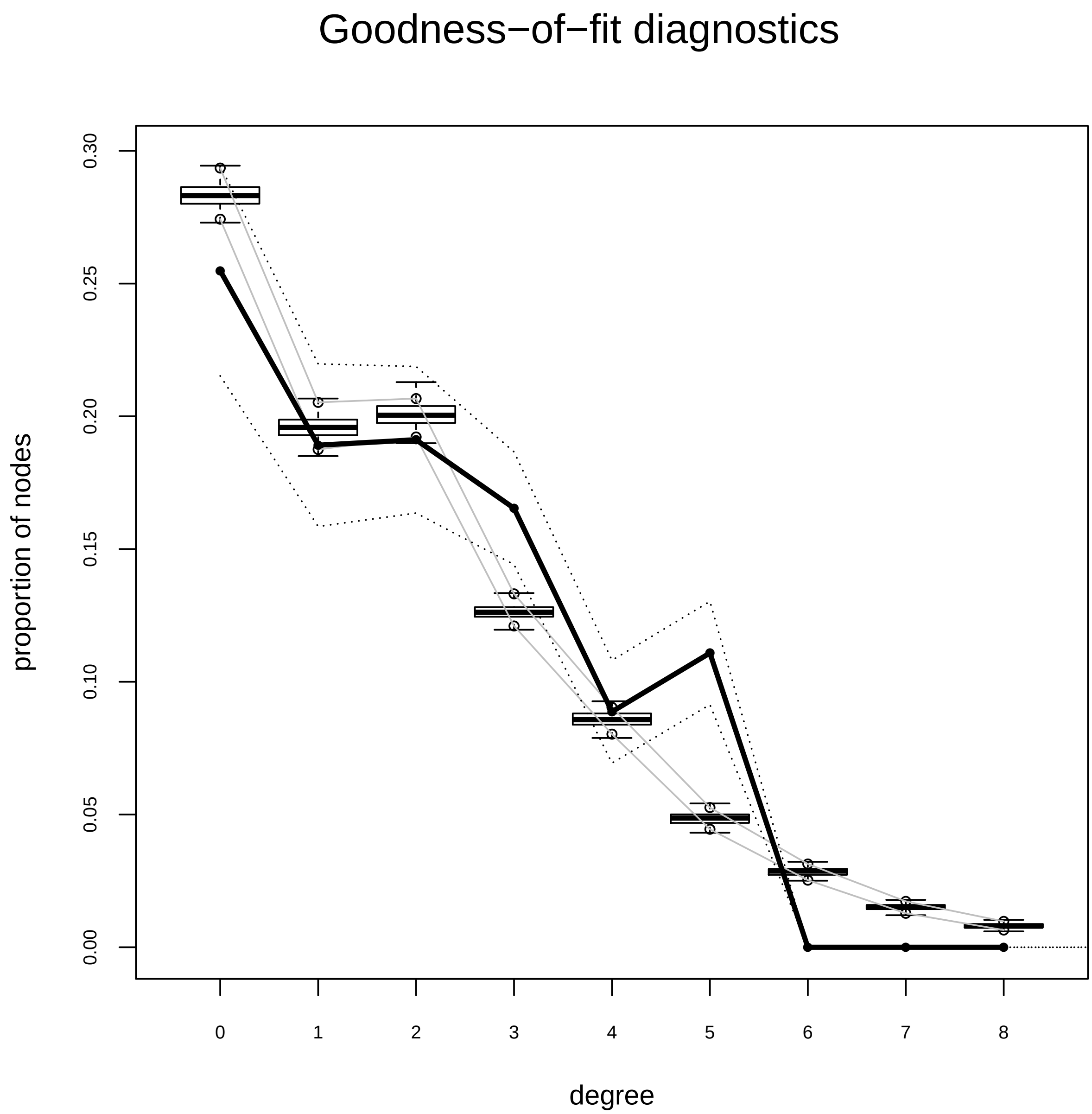

Figure 9: Degree Goodness-of-fit of Model 4. Model-simulated (boxplots) and observed (black lines) degree distributions.

you go to the same school as X?", without asking to *which* schools did the respondent and X go. Without this node-level information, is is not straightforward to reconstruct the dyadic covariates for *all* potential ties in the network, as model estimation requires.

This situation must be distinguished from *tie* attributes such as mutually exclusive (polytomous) type of relationship (short-term, long-term, or marital sexual partnership), frequency of interaction, or multiple non-mutually-exclusive relationship types (friendship and collaboration). Because these are meaningful only for ties and not for all potential relations, they can be handled in a valued or multiplex network framework described below.

6.1.2 Degree censoring

As mentioned in Section 5.3, each respondent was able to nominate at most 5 alters. This form of censoring is common in egocentric and dyad-census designs alike (Udry 2003, for example), and a number of methods have been proposed for addressing it in the dyad-census case (Hoff et al. 2013; Ott et al. 2017). We did not adjust for it in our application, and how to do so is subject for future research.

6.2 Future directions

6.2.1 Directed networks

We have focused on modeling undirected relations. Relations collected in a social survey may be directed as well, however, with questions like "Did you lend money to X?" as opposed to "Did X lend money to you?" and similarly for alter–alter ties.

Directed relations present additional "degrees of freedom" for data collection design, such as whether out-alters, in-alters, or both are observed; and different dyadic structures (e.g., mutuality vs. asymmetry) and triadic structures (e.g., transitivity vs. cyclicality) become estimable, depending on the available data. We leave this for future work.

6.2.2 Valued networks

The inferential framework of KM revolves around recovering sufficient statistics of the population network, and would therefore extend directly to valued ties like counts, provided all alters with non-zero relationship values are observed; and similarly for polytomous observations.

6.2.3 Multiplex networks

Joint modeling of multiple relationship types over the same set of actors is a relatively recent area for ERGMs, with signed networks receiving particular attention (Huitsing et al. 2012). The framework described should extend to such data, again provied the sufficient statistics can be expressed in the form (2).

6.2.4 Repeated measures

Some social surveys conducted in multiple waves deliberately overlap the samples from successive waves, so as to obtain longitudinal data about the actors. Estimation of dynamic network models based on such data is subject for ongoing research.

Appendix

Table 2 shows estimates and standard errors (in parentheses) of coefficients of the four models presented graphically in Figure 5.

References

Burt, Ronald. S. 1984. "Network Items and the General Social Survey." *Social Networks* 6 (4): 293–339. https://doi.org/10.1016/0378-8733(84)90007-8.

Frank, Ove, and David Strauss. 1986. "Markov Graphs." *J. Am. Stat. Assoc.* 81 (395): 832–42.

Handcock, Mark S., and Krista J. Gile. 2010. "Modeling Social Networks from Sampled Data." *Ann. Appl. Stat.* 4 (1): 5–25. https://doi.org/10.1214/08-AOAS221.

Handcock, Mark S., David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. 2019. *Ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks.* The Statnet Project (http://www.statnet.org). CRAN.R-project.org/package=ergm.

Hoff, Peter, Bailey Fosdick, Alex Volfovsky, and Katherine Stovel. 2013. "Likelihoods for Fixed Rank Nomination Networks." *Network Science* 1 (3): 253–77. https://doi.org/10.1017/nws.2013.17.

Honaker, James, Gary King, and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45 (7): 1–47. http://www.jstatsoft.org/v45/i07/.

Huitsing, Gijs, Marijtje A.J. van Duijn, Tom A.B. Snijders, Peng Wang, Miia Sainio, Christina Salmivalli, and René Veenstra. 2012. "Univariate and Multivariate Models of Positive and Negative Networks: Liking, Disliking, and Bully–Victim Relationships." *Social Networks* 34 (4): 645–57. https://doi.org/10.1016/j.socnet.2012.08.001.

Hunter, David R. 2007. "Curved Exponential Family Models for Social Networks." *Social Networks* 29: 216–30. https://doi.org/10.1016/j.socnet.2006.08.005.

Hunter, David R., and Mark S. Handcock. 2006. "Inference in Curved Exponential Family Models for Networks." *J. Comput. Graph. Stat.* 15 (3). American Statistical Association: 565–83.

Hunter, David R., Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. 2008. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks." *J. Stat. Softw.* 24 (3): 1–29. http://www.jstatsoft.org/v24/i03.

King, Gary, James Honaker, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." *American Political Science Review* 95 (1): 49–69.

Koskinen, Johan H., Garry L. Robins, and Philippa E. Pattison. 2010. "Analysing Exponential Random Graph (P-Star) Models with Missing Data Using Bayesian Data Augmentation." *Stat. Methodol.* 7 (3): 366–84. https://doi.org/10.1016/j.stamet.2009.09.007.

Krivitsky, Pavel N., Mark S. Handcock, and Martina Morris. 2011. "Adjusting for Network Size and Composition Effects in Exponential-Family Random Graph Models." *Stat. Methodol.* 8 (4): 319–39. https://doi.org/10.1016/j.stamet.2011.01.005.

Krivitsky, Pavel N., and Eric D. Kolaczyk. 2015. "On the Question of Effective Sample Size in Network Modeling: An Asymptotic Inquiry." *Statistical Science* 30 (2). The Institute of

Mathematical Statistics: 184–98. https://doi.org/10.1214/14-STS502.

Krivitsky, Pavel N., and Martina Morris. 2017. "Inference for Social Network Models from Egocentrically-Sampled Data, with Application to Understanding Persistent Racial Disparities in HIV Prevalence in the US." *Annals of Applied Statistics* 11 (1): 427–55. https://doi.org/10.1214/16-AOAS1010.

Laumann, Edward O., John H. Gagnon, Robert T. Michael, and Stuart Michaels. 1992. "National Health and Social Life Survey." Chicago, IL, USA: University of Chicago and National Opinion Research Center [producer], 1995. Ann Arbor, MI, USA: Inter-university Consortium for Political and Social Research [distributor], 2008-04-17. https://doi.org/10.3886/ICPSR06647.

National Survey of Family Growth Staff. n.d. "National Survey of Family Growth (Nsfg)." Division of Vital Statistics, National Center for Health Statistics. http://www.cdc.gov/nchs/nsfg.htm.

Ott, Miles Q., Matthew T. Harrison, Krista J. Gile, Nancy P. Barnett, and Joseph W. Hogan. 2017. "Fixed Choice Design and Augmented Fixed Choice Design for Network Data with Missing Observations." *Biostatistics* 20 (1): 97–110. https://doi.org/10.1093/biostatistics/kxx066.

Pattison, Philippa E., Garry L. Robins, Tom A.B. Snijders, and Peng Wang. 2013. "Conditional Estimation of Exponential Random Graph Models from Snowball Sampling Designs." *J. Math. Psychol.* 57: 284–96.

Schweinberger, Michael. 2011. "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families." *Journal of the American Statistical Association* 106 (496): 1361–70. https://doi.org/10.1198/jasa.2011.tm10747.

Smith, Jeffrey A. 2012. "Macrostructure from Microstructure: Generating Whole Systems from Ego Networks." *Sociol. Methodol.* 42 (1): 155–205. https://doi.org/10.1177/0081175012455628.

Smith, Tom W., Michael Davern, Jeremy Freese, and Stephen L. Morgan. 2019. *General Social Surveys, 1972-2018: Cumulative Codebook*. Chicago: National Opinion Research Center (NORC).

Snijders, Tom A. B., Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. 2006. "New Specifications for Exponential Random Graph Models." *Sociological Methodology* 36 (1). Blackwell Synergy: 99–153.

Udry, J. Richard. 2003. "The National Longitudinal Study of Adolescent Health (Add Health), Waves I & II, 1994–1996; Wave III, 2001–2002." Carolina Population Center, University of North Carolina at Chapel Hill.

Wasserman, Stanley S., and Philippa Pattison. 1996. "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*." *Psychometrika* 61 (3). Springer: 401–25.

Table 2: Model coefficients. Standard errors in parentheses.

Term	Model 1	Model 2	Model 3	Model 4
Race - Black	-0.235 (0.182)	0.496 (0.118)	-0.279 (0.412)	-0.082 (0.294)
Race - Other	0.071 (0.141)	0.857 (0.096)	1.525 (0.332)	1.26 (0.247)
UH Race	–	1.989 (0.165)	–	–
DH Race - White	–	–	2.107 (0.402)	1.72 (0.287)
DH Race - Black	–	–	3.853 (0.483)	2.765 (0.35)
DH Race - Other	–	–	0.436 (0.457)	0.173 (0.311)
Sex - Female	0.033 (0.049)	-0.005 (0.041)	-0.007 (0.041)	-0.019 (0.021)
UH Sex	–	0.271 (0.043)	0.272 (0.04)	0.243 (0.034)
Education - H.S. degree	0.288 (0.103)	0.111 (0.085)	0.285 (0.101)	0.169 (0.071)
Education - Some college	0.168 (0.098)	-0.043 (0.088)	0.186 (0.106)	0.118 (0.071)
Education - BA/BS degree	0.504 (0.109)	0.321 (0.099)	0.397 (0.134)	0.212 (0.09)
Education - Postgraduate	0.446 (0.122)	0.281 (0.107)	0.231 (0.135)	0.103 (0.089)
UH Education	–	0.675 (0.053)	–	–
DH Education - Less than H.S. degree	–	–	1.143 (0.247)	0.989 (0.187)
DH Education - H.S. degree	–	–	0.458 (0.158)	0.421 (0.131)
DH Education - Some college	–	–	0.296 (0.106)	0.275 (0.087)
DH Education - BA/BS degree	–	–	0.755 (0.138)	0.648 (0.106)
DH Education - Postgraduate	–	–	1.223 (0.141)	0.959 (0.103)
Religion - Catholic	0.061 (0.114)	0.276 (0.067)	0.23 (0.117)	0.166 (0.08)
Religion - Jewish	0.137 (0.26)	0.871 (0.2)	0.14 (0.204)	0.162 (0.132)
Religion - None	-0.111 (0.099)	0.341 (0.068)	0.757 (0.133)	0.591 (0.093)
Religion - Other	0.045 (0.123)	0.582 (0.092)	0.583 (0.147)	0.449 (0.1)
UH Religion	–	1.702 (0.073)	–	–
DH Religion - Prostestant	–	–	1.794 (0.155)	1.38 (0.112)
DH Religion - Catholic	–	–	1.874 (0.148)	1.428 (0.104)
DH Religion - Jewish	–	–	3.864 (0.325)	2.471 (0.219)
DH Religion - None	–	–	0.601 (0.209)	0.55 (0.157)
DH Religion - Other	–	–	1.826 (0.278)	1.373 (0.187)
Age	0.015 (0.022)	0.052 (0.017)	0.053 (0.018)	0.036 (0.01)
Abs.diff. in Age	–	-0.529 (0.03)	-0.528 (0.03)	-0.413 (0.024)
Married	0 (0.083)	-0.095 (0.078)	-0.095 (0.078)	-0.077 (0.044)
Size of residence > 2000	0.189 (0.267)	0.191 (0.248)	0.186 (0.253)	0.11 (0.155)
GWESP(decay=0.1)	–	–	–	0.105 (0.052)