Outliers and Influential Observations in Exponential Random Graph Models

Johan Koskinen [*†], Peng Wang [‡], Garry Robins [§], and Philippa Pattison [¶]

April 29, 2018

[*]E-mail: johan.koskinen@manchester.ac.uk. The author would like to acknowledge financial support from the Leverhulme Trust Grant RPG-2013-140

[†]The Mitchell Centre for Social Network Analysis and the Department of Social Statistics, School of Social Sciences, University of Manchester, Manchester, M139PL UK; Melbourne School of Psychological Sciences, The University of Melbourne, Australia; Institute of Analytical Sociology, University of Linköping.

[‡]Centre for Transformative Innovation, Faculty of Business and Law, Swinburne University of Technology, Australia

[§]Melbourne School of Psychological Sciences, The University of Melbourne, Australia

[¶]The University of Sydney, Australia

Abstract

We discuss measuring and detecting influential observations and outliers in the context of exponential family random graph (ERG) models for social networks. We focus on the level of the nodes of the network and consider those nodes whose removal result in changes to the model as extreme or "central" with respect to the structural features that "matter". We construe removal in terms of two case deletion strategies: the tie-variables of an actor are assumed to be unobserved or the node is removed resulting in the induced subgraph. We define the difference in inferred model resulting from case deletion from the perspective of information theory and difference in estimates, both in the natural and mean value parameterisation, representing varying degrees of approximation. We arrive at several measures of influence, and propose the use of two that do not require refitting of the model and lend themselves to routine application in the ERGM fitting procedure. MCMC p-values are obtained for testing how extreme each node is with respect to the network structure. The influence measures are applied to two well-known data sets to illustrate the information they provide. From a network perspective, the proposed statistics offer an indication of which actors are most distinctive in the network structure, in terms of not abiding by the structural norms present across other actors.

1. Introduction

It is in the nature of statistical models that parameter estimates change with the addition or removal of observations. If, however, an observation substantially alters the overall inference we might suspect that this observation has a major influence on our model. It could also be that an observation does not alter our overall conclusions but that it is highly unusual given the other information we have. Consequently considerable attention in the statistical literature has been devoted to developing diagnostics tools that pick out influential observation and outliers (see e.g. Chatterjee and Hadi, 1986, in the case of linear regression and Pregibon, 1981; Williams, 1987; Lesaffre and Albert, 1989; and Hines, Lawless, and Carter,1992, for extensions to

varying forms of generalized linear models).

For social network data (Wasserman and Faust, 1994), the class of exponential random graph models (ERGM)(Holland and Leinhardt, 1981; Frank and Strauss, 1986; Wasserman & Pattison, 1996; Pattison & Wasserman, 1999; Snijders et al., 2006; Hunter and Handcock, 2006) has become an important approach for capturing the complex dependencies giving rise to observable tie variables in social networks (Robins and Morris, 2007). ERGM are a class of log-linear models for the tie-variables of nodes in a network. Using the Hammersley-Clifford theorem (Besag 1974), Frank and Strauss (1986) derived a set of sufficient statistics for ERGM from assumptions about how the ties of nodes may depend on each other. These statistics are interactions between tie-variables that correspond to different order stars as well as triangles. Snijders et al. (2006) elaborated on these dependence assumptions to derive an extended class of network statistics.

The statistical literature on influence has largely drawn on linear regression and therefore has been concerned with defining analogies to residuals that may be used to study how, for example, case deletion changes the deviance. This means that you can define residuals and case-deletion even when you do not have independently defined error terms. This approach, which works well in logistic regression (Pregibon, 1981) and GLM (Williams, 1987), relies to a large extent on the assumption of independence of observations. For ERGMs, while we may still consider changes in deviance, the intrinsic assumption of interdependent observations prevents us from adopting the standard approach of expressing this in terms of residuals. The analysis of outliers in contingency tables is closely related to the case of ERGM but has the advantage of being able to rely on distributional assumptions (e.g. Kuhnt, 2004) that do not apply for ERGM.

ERGMs generally cater for a degree of heterogeneity with respect to the observables among the actors. Even if a model asserts that actors are stochastically equivalent (in the sense for example that the model is permutation invariant with respect to permutations of the node labels), for the actual realisation we might have big differ-

ences between the interactional patterns of individuals. Some actors may for example have many ties whereas others may have no ties at all. In a manner of speaking, for some models you may even say that it is expected that some actors are unexpectedly different. Robins, Pattison and Woolcock (2005) demonstrate exactly this behaviour in their thorough investigation of ERGM specifications. Naturally you have a similar situation in standard statistical models where the deviation from the general tendency has "long tails" - a regression model with errors distributed according to a Cauchy distribution may have extreme outliers - but in the case of ERGMs this phenomena is subtly different in that the observations pertaining to one actor affects the interpretation of observations pertaining to other actors (Wang et al., 2013). For example, for models for repeated measures, observations are dependent within individuals but measurement occasions are nested within individuals. Residuals can thus be defined on the individual level (Waternaux, Laird, and Ware, 1989; Weiss and Lazaro, 1992). In a network each tie-variable is however cross-classified by its constituent nodes.

The rest of the paper is structured as follows. We begin by defining the ERGM framework, accompanied by some notation necessary for the purpose of the proposed methodology, and present the main arguments for the particular type of "case deletion" chosen here. We proceed by presenting two approaches to removing an actor and the associated case-deletion estimators, which is followed by a derivation of measures that weigh together the shifts in estimates as compared to the complete data analysis and a series of approximations. We present a Monte Carlo-based test of the statistics that can be used to gather further insight into the extent to which nodes are extreme. The measures are then applied to two well-known data sets with a thoroughly researched set of model-specifications. The approximate measures are shown to be good and the most compelling one has been implemented in MPNet. While one of the measures have the intuitive appeal of comparing the estimated model to a model with the actor entirely removed, the preferred measure uses a missing data approach which does not require that we can interpret ERGMs for subsets.

2. The model

In the following we assume that we are interested in modelling a graph of order n, with fixed vertex set V, but stochastic edge set $E \subseteq \mathcal{E} = \binom{V}{2}$. We assume that the model is defined for graphs with adjacency matrices $y \in \mathcal{Y}$, and that given a set of fixed covariates x it has the form

$$p_{\theta,x}(y) \equiv \Pr(Y = y|\theta, x) = \exp\{\theta^{\mathrm{T}} z(y; x) - \psi_{\mathcal{Y}}(\theta; x)\},$$

where θ is a $p \times 1$ vector of parameters, $\theta \in \Theta \subseteq \mathbb{R}^p$, $z(y; x)$ is a vector valued function of y for each x, and $\psi_{\mathcal{Y}}(\theta; x) = \log \sum_{y \in \mathcal{Y}} \exp\{\theta^{\mathrm{T}} z(y; x)\}$ is a normalising constant. For the simplest case where ties are assumed independent and no covariates are used, $z(y; x) = z(y)$ is just a count of the number of ties in the network, i.e. $z(y) = \sum_{1 \le i \le n} y_{i+}/2$, where $y_{i+} = \sum_{j \neq i} y_{ij}$. Frank and Strauss (1986) proposed a Markov dependence assumption for ERGMs, whereby the tie-variables Y_{ij} and $Y_{k\ell}$ are conditionally independent, conditional on the rest of the graph, unless $\{i, j\} \cap \{k, \ell\} \neq \emptyset$. This implies a model that in addition to the edge-statistic, has higher-order interaction terms such as the number of k-stars $S_k = \sum_{1 \le i \le n} \binom{y_{i+}}{k}$, $k = 2, \ldots, n-1$, and the number of triangles $\sum_{1 \le i < j < k \le n} \sum_{k \neq i,j} y_{ij} y_{ik} y_{jk}$. While defining a parsimonious class of models for complex dependencies, these type of models have long been known to be badly specified (Strauss, 1986; Jonasson, 1999; Handcock, 2003). Snijders et al. (2006) proposed a modified set of statistics that have proved to lead to better behaved models and that have since been successfully employed in empirical analysis (Lusher et al., 2013). These models replace k-star statistics with an alternating star statistic

$$u^{(s)}_{\lambda_s}(y) = \sum_{k=2}^{n-1} (-1)^k \frac{S_k}{\lambda_s^{k-2}}, \tag{1}$$

and an alternating triangle statistic

$$u^{(t)}_{\lambda_t}(y) = \lambda_t \sum_{1\leq i<j\leq n} y_{ij} \left\{ 1 - \left(1 - \frac{1}{\lambda_t} \right)^{L_{ij}} \right\}, \tag{2}$$

where $L_{ij} = \sum_{h\neq i,j} y_{ih} y_{hj}$ is a count of the number of two-paths connecting i and j. For the statistics (1) and (2), λ_s and λ_t are either considered user-defined smoothing constants or parameters to be estimated (Hunter and Handcock, 2006; Koskinen et al., 2010). Schweinberger (2011) analyses Markov graphs and the models defined by the new specifications of Snijders et al. (2006) in great detail and concludes that the latter are more stable than the former. In particular, a model with statistics $\sum_{1\leq i<j\leq n} y_{ij}$ and (2), is stable for $\lambda_t \geq 0.5$ (Schweinberger, 2011).

ERGMs admit dependence of tie-variables on exogenous nodal (and dyadic) covariates (Robins, Elliott, and Pattison, 2001; Robins, Pattison, and Elliott, 2001). For a monadic binary covariate $x = (x_i)_{i\in V}$, we may define the main effect of this covariate on the probability of a tie through the statistic $\sum_{1\leq i<j\leq n} y_{ij}(x_i + x_j)$. If the corresponding parameter is positive, $x_i = 1$ is associated with i being incident to more edges. Homophily, the tendency for nodes with similar attributes to be more likely to be directly connected than dissimilar nodes (McPherson et al., 2001), can be modelled through the inclusion of the statistic $\sum_{1\leq i<j\leq n} y_{ij} \mathbf{1}\{x_i = x_j\}$. Similar statistics may be defined for categorical and continuous attributes (Robins and Daraganova, 2013).

The entries of y are typically not independent and if Y_A and Y_B are the collection of variables corresponding to disjoint subsets $A, B \subseteq \mathcal{E}$ we generally do *not* have that $\Pr(Y_A = u, Y_B = v|\theta, x) = \Pr(Y_A = u|\theta, x) \Pr(Y_B = v|\theta, x)$. The "smallest" observational unit is the dyad and we could consider the Y_{ij}'s ($ij \in \mathcal{E}$) to constitute our observations. For linear models with independently defined error terms the residuals are straightforwardly defined as the difference between the observed and fitted value. For binary response we may similarly define residuals as $e_{ij} = y_{ij} - \hat{\pi}_{ij}$, for a dyad, where $\hat{\pi}_{ij}$ is the predicted tie-probability (for GLMs other forms may be considered; see, for example, Williams, 1984, and Pierce and Schafer,

1986). For independent observations, $\hat{\pi}_{ij}$ is unambiguously defined as the marginal probability but for ERGM the marginal probabilities $E_\theta\{Y_{ij}\} = \sum_{y\in\mathcal{Y}} y_{ij} p_{\theta,x}(y)$ are intractable. It is tempting to, as in Wasserman and Pattison (1996), use the conditional $\hat{\pi}_{ij|-ij} = \Pr(Y_{ij} = 1|\hat{\theta}, Y_{\mathcal{E}\setminus\{i,j\}} = y_{\mathcal{E}\setminus\{i,j\}})$ rather than the marginal probabilities. Conditional probabilities $\hat{\pi}_{ij|-ij}$ are however a poor choice for assessing fit as you need to condition on observed data.

Marginal probabilities can be approximated numerically using the Monte Carlo estimate of $E_\theta\{Y_{ij}\}$. A homogenous ERGM is however permutation invariant (Frank and Strauss, 1986; Schweinberger et al., 2017), meaning that residuals will not be sensitive to model specification. For example, consider on the one hand a Bernoulli model with the sufficient statistic $\sum_{1\leq i<j\leq n} y_{ij}$, and a model that in addition has the sufficient statistic defined by (2) on the other. For any pair $\{i, j\}$ the predicted tie-probability will be the same under the two models and consequently the residuals will all be the same for the two models (Block et al., in press). The interpretation of this is that the added dependencies of a Markov model, or higher order dependencies such as (2), fit, or account for, interactions of tie-variables, not the marginal probabilities of tie-variables. It might for example be the case that e_{ij} and $e_{k\ell}$ considered separately may appear to be small but that when they are considered jointly they are large. A simple example is when we are considering the variables y_{ij} and y_{ji} in a directed graph, the marginal tie-probabilities may be low but a reciprocated dyad $y_{ij} = y_{ji} = 1$ may be much more likely than an asymmetric one $y_{ij} \neq y_{ji}$. So while residuals of individual tie-variables are not informative, defining residuals for all possible interaction effects is not feasible.

A natural way of grouping the variables in Y is by the nodes. This is also a natural approach in, for example, repeated measures models, where residuals can be defined on the individual level (Waternaux, Laird, and Ware, 1989) or individual by occasion for each individual (Weiss and Lazaro, 1992). Here we let $Y_{(i)}$ and $y_{(i)}$ denote the adjacency matrix of the subgraph of order $n-1$ induced by removing node i, for $i \in V$. Analogously we let $x_{(i)}$ be the collection of covariates that do not include those

of node i, and let $z(y_{(i)}; x_{(i)})$ be the vector of statistics evaluated for $y_{(i)}$ and $x_{(i)}$. Note the departure from repeated measures where measures are nested within individuals. For graphs, as ties are not nested within individuals, removing observations for i also means removing observations y_{ij} for all $j \neq i$. We assume that z may unambiguously be defined on a graph of order $n-1$, and we do not make any notational distinctions beyond that which is implied by the arguments of z.

In general, removing i leads to different sufficient statistics than setting $y_{ij} = 0$ (for $j \neq i$). In other words, $z(y_{(i)}; x_{(i)})$ is not the same as evaluating $z(y^*; x)$ for an adjacency matrix y^* with elements $y^*_{ij} = 0$ but $y^*_{k,\ell} = y_{k,\ell}$ for all $\{k, \ell\}$ such that $\{k, \ell\} \cap \{i\} = \emptyset$ (Snijders, 2010, elaborate on statistics defined for subsets of nodes under different conditions). For instance, if a count of the number of isolates is part of z, then these two statistics are different. In order to make explicit the link between $y_{(i)}$ and $x_{(i)}$, we denote the range space of $Y_{(i)}$ by $\mathcal{Y}_{(i)}$.

In the following, the collection of tie-variables that involves i are denoted by $y_{i\bullet}$, and the corresponding attribute vector $x_{i\bullet}$.

3. Estimation and case deletion

Since the model $p_{\theta,x}(y)$ is an exponential family distribution (Barndorff- Nielsen 1978; Lehmann, 1983), the maximum likelihood estimate (MLE), $\hat{\theta}$, given an observation is such that it satisfies

$$\mu_{\mathcal{Y}}(\hat{\theta}; x) = z(y; x), \tag{3}$$

where $\mu_{\mathcal{Y}}(\hat{\theta}; x) = E_{\hat{\theta}}\{z(Y; x)|x\}$ is the expected value. Furthermore the Fisher information matrix and the negative Hessian are both equal to $I(\hat{\theta}) = Cov_{\hat{\theta}}\{z(Y; x)|x\}$. The moment equation (3) may be solved numerically for the MLE and once an estimate is obtained $I(\hat{\theta})$ may be approximated by the corresponding MCMC quantity (Corander, Dahmström, and Dahmström, 1998, 2002; Crouch, Wasserman, and Trachtenberg, 1998; Snijders, 2002; Handcock, 2003; Hunter and Handcock, 2006).

Handcock (2003) showed that an alternative parametrisation, the mean value parametrisation (MVP), of the ERGM could provide additional insight into the model. More specifically the MVP of the ERGM on $\mathcal{Y}$ and x, is a mapping $\mu_{\mathcal{Y}} : \Theta \to C$, where C is the relative interior of the convex hull on $\{t \in \mathbb{R} : z(y;x) = t$, for some $y \in \mathcal{Y}\}$, and as defined above $\mu_{\mathcal{Y}}(\theta;x) = E_\theta\{z(Y;x)|x\}$. We may note a particularly useful property of the MVP, namely that the MLE is given by $\hat{\mu} = z(y;x)$. The Fisher information matrix is given by $I(\mu_{\mathcal{Y}}^{-1}(\hat{\mu}))^{-1} = Cov_{\mu_{\mathcal{Y}}^{-1}(\hat{\mu})}\{z(Y;x)|x\}^{-1}$, where $\mu_{\mathcal{Y}}^{-1}$ denotes the inverse function, $\mu_{\mathcal{Y}}^{-1}(A) = \{\theta \in \Theta : \mu_{\mathcal{Y}}(\theta;x) \in A\}$.

For the purposes of investigating how large an influence the observations pertaining to an actor has on the estimate $\hat{\theta}$, how do we conceptualise fitting the model with that actor removed? Here we propose two alternative and complementary interpretations. The first is to remove the information about the part of y that pertains to i. The second is to remove the part of y that pertains to i altogether. By the first approach we mean something that might be expressed as "what would our estimates be had we not known the values of y_{ij} for any of the j's"? We shall refer to this approach as the "missing data (MD) approach" and the estimate we obtain when i is removed according to the MD approach is denoted by $\hat{\theta}_{(i)}$, the missing data MLE (MDMLE). While we assume that information on y_{ij} is missing, for $j \in V\backslash\{i\}$, the values on all the covariates are considered known. The analogy to analysis of ERGMs with missing data is that the MDMLE would be the MLE for the network had y_{ij} been missing for all j (what Huisman, 2009, refers to as *item non-response* in the case of social network data) and observations missing at random in the sense of Rubin (1976) as demonstrated in Handcock and Gile (2010). Hence the name "MD" approach.

In the second approach node i is removed entirely from the network as are its covariate values, so that instead of having the observations y and x, we have the observations $y_{(i)}$ and $x_{(i)}$. Since this is analogous to fitting a model to the part of the network that is known, using only the available case when there is missing information for i, this approach is called the available case (AC) approach (c.p. "available-case" analysis, Little and Rubin, 1987). The corresponding estimate is denoted by $\widetilde{\theta}_{(i)}$, the

available case MLE (ACMLE).

3.1 Estimation

For AC the estimation is done using the same procedure as for the completely observed network, and $\widetilde{\theta}_{(i)}$ satisfies

$$\mu_{\mathcal{Y}_{(i)}}(\widetilde{\theta}_{(i)}; x_{(i)}) = z(y_{(i)}; x_{(i)}),$$

but where $\mu_{\mathcal{Y}_{(i)}}(\widetilde{\theta}_{(i)}; x_{(i)}) = E_{\widetilde{\theta}_{(i)}}\{z(Y_{(i)}; x_{(i)})|x_{(i)}\}$, and the Fisher information matrix is $I(\widetilde{\theta}_{(i)}) = Cov_{\widetilde{\theta}_{(i)}}\{z(Y_{(i)}; x_{(i)})|x_{(i)}\}$. Note that expectations are now taken with respect to a distribution on $\mathcal{Y}_{(i)}$ rather than $\mathcal{Y}$. Obtaining the MDMLE is a more involved but similarly entails finding the estimate $\hat{\theta}_{(i)}$ that satistfies

$$\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) = \mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x), \tag{4}$$

where $\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x)$ is defined as before but

$$\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x) = E_{\hat{\theta}_{(i)}}\{z(Y; x)|x, Y_{(i)} = y_{(i)}\},$$

i.e., with respect to the conditional distribution restricted to $\mathcal{Y}^i(y_{(i)}) = \{u \in \mathcal{Y} : u_{(i)} = y_{(i)}\}$. This follows from simply setting to zero the differentiated log likelihood $\frac{\partial}{\partial\theta}\log\sum_{u\in\mathcal{Y}^i(y_{(i)})} p_{\theta,x}(u)$, and solving for θ. Handcock and Gile (2010) proposed a maximum likelihood-based scheme for fitting the ERGM with missing data. Here we will use stochastic approximation to solve the equation (4) (Koskinen and Snijders, 2013). The negative of the Hessian is straightforward to obtain as $Cov_{\hat{\theta}_{(i)}}\{z(Y; x)|x\} - Cov_{\hat{\theta}_{(i)}}\{z(Y; x)|x, Y_{(i)} = y_{(i)}\}$. We do not pursue a Bayesian data-augmentation scheme (Koskinen, Robins, Pattison, 2010) as the proposed measures ultimately do not require estimation with missing values.

AC is in a sense straightforward to interpret as it is the direct equivalent of the standard case-deletion approach (Cook, 1977). However, while models for indepen-

dent cases scale up and are well-defined on subsets of data, this is not necessarily true for networks (Anderson, Butts, and Carley, 1999). For ERGM it is a know fact that they do not marginalise, something which follows from definition of the dependence graph (Koskinen, Robins, and Pattison, 2010). Snijders (2010) points out that if the graph on V follows an ERGM, then the subgraph induced by node set $V^* \subset V$ only follows an ERGM for trivial models. He goes on to identify the conditions under which the graph on V^* follows an ERGM with the same parameters as that for V conditionally. For example, if V^* are the nodes of a saturated snowball sample, the graph on V^* follows an ERGM with same parameters as the ERGM for V restricted to the space of connected graphs on V^*. Schweinberger et al. (2017) study properties of ERGMs defined on subsets under a large number of different conditions and assesses the implications for statistical inference. While they provide a more nuanced account and more applicable results than Shalizi and Rinaldo (2013), the fact still remains that since ERGM do not marginalise it is not clear how a model for $\mathcal{Y}_{(i)}$ relates to a model for $\mathcal{Y}$.

3.2 Combined influence for $p > 1$

When $p = 1$, the influence on the estimate of θ may simply be investigated by plotting $\widetilde{\theta}_{(i)}$ and $\hat{\theta}_{(i)}$ against $\hat{\theta}$ for each of the i's. When we have more than one parameter we may still plot the individual elements of the parameter vector separately but it will be hard to asses the overall influence of an actor from these partial plots. These plots may not be directly comparable since parameters are likely to be on different scales. Therefore we may not know which parameters are most "important" and what weight should be given to the deviations on the different elements of θ. Additionally, the estimates are typically highly correlated, wherefore it may be hard to parse out the influence of actors on individual elements of θ. A measure corresponding to DFBETA for linear regression (Belsley et al., 1980), that takes the correlation between parameters into account, could be developed for ERGM but we do not pursue that here.

3.2.1 Kullback-Leibler divergence

MD A common way of investigating similarity between distributions on range space $\mathcal{Y}$ with probability mass functions $p(y)$ and $q(y)$, where p is dominated by q, is by using the Kullback-Leibler divergence $D(p||q) = E_{Y|p}\{\log(p(Y)/q(Y))\}$. Note that the Kullback-Leibler divergence may be rewritten $H(p,q) - H(p)$, where $H(p,q) = -\sum_{y\in\mathcal{Y}} p(y)\log q(y)$ is commonly referred to as the cross entropy and $H(p) = -\sum_{y\in\mathcal{Y}} p(y)\log p(y)$ is the entropy. This is of some significance as the ERGM, $p_{\mu_{\mathcal{Y}}^{-1}(\mu)}$, with statistics z, maximises $H(p)$ subject to the constraint that $E_{Y|p}\{z(Y)\} = \mu$. The Kullback-Leibler divergence is, as Handcock (2003) points out, a natural choice for assessing similarity of distributions in the case of ERGMs, in which case it is given by

$$E_{\hat{\theta}}\left\{\log\frac{p_{\theta,x}(Y)}{p_{\phi,x}(Y)}\right\} = (\theta-\phi)^{\mathrm{T}}\mu_{\mathcal{Y}}(\theta;x) + \psi_{\mathcal{Y}}(\phi;x) - \psi_{\mathcal{Y}}(\theta;x),$$

If θ is the MLE, $\hat{\theta}$, $\mu_{\mathcal{Y}}(\hat{\theta};x) = z(y;x)$ and we may define the missing data divergence (DMD) as

$$D(\hat{\theta}||\phi) = (\hat{\theta}-\phi)^{\mathrm{T}}z(y;x) + \psi_{\mathcal{Y}}(\phi;x) - \psi_{\mathcal{Y}}(\hat{\theta};x),$$

which we recognise as half the deviance $2\{\log p_{\hat{\theta},x}(y) - \log p_{\phi,x}(y)\}$ between the two models defined by $\hat{\theta}$ and ϕ, where $D(\hat{\theta}||\phi)$ is taken to mean $D(p_{\hat{\theta},x}||p_{\phi,x})$, when there is no ambiguity. The interpretation is therefore that $D(\hat{\theta}||\phi)$ measures the decreases in likelihood as the *maximum likelihood* estimate is substituted by a less optimal estimate. Construing influence as the degree of change in deviance has also been done for GLMs when $p > 1$ (see e.g. Williams, 1987; Lee, 1988). Cook (1986) also identifies the relationship between the influence statistic defined in terms of differences in fitted values (Cook, 1977) and the deviance or likelihood displacement. Handcock and Gile (2010) used $D(\cdot||\cdot)$ as a general measure of how different the distributions defined by the MDMLEs were to the MLE for a data set were the MDMLE was calculated for snowball sampled subsets of y.

In order to calculate $D(\hat{\theta}||\hat{\theta}_{(i)})$, defined on $\mathcal{Y}$, for each $i \in V$, we need to refit the model by solving (4) n times. In addition, since $\psi_{\mathcal{Y}}$ typically is analytically intractable, we require some numerical approximation to this normalising constant. Hunter and Handcock (2006) proposed to use the path sampler, a generalisation of bridged importance sampling that draws on the principle of thermodynamic integration in statistical physics (Meng and Wong, 1996; Gelman and Meng, 1998; Neal, 1993). In the calculations here, the quantity $\lambda(\phi, \theta) = \psi_{\mathcal{Y}}(\phi; x) - \psi_{\mathcal{Y}}(\theta; x)$, has been estimated by $\hat{\lambda}(\phi, \theta) = \frac{1}{M} \sum_{m=1}^{M} (\phi - \theta) z(y_m; x)$, where y_m has been generated from $p_{\phi_m, x}$, $\phi_m = t_m \theta + (1 - t_m)\phi$, and t_m are i.i.d. uniformly random variates. There is a variety of alternative samplers for approximating $\lambda(\phi, \theta)$ but the path sampler appears to be the most efficient to date. In addition the path sampler has the advantage that it estimates the ratio on the log-scale (for a review see Gelman and Meng, 1998).

AC For the AC approach we define $D(\cdot||\cdot)$ a little differently, namely as $D(\widetilde{\theta}_{(i)}||\hat{\theta})$ with respect to the reduced graph space $\mathcal{Y}_{(i)}$, giving the available case divergence (DAC)

$$(\widetilde{\theta}_{(i)} - \hat{\theta})^{\mathrm{T}} z(y_{(i)}; x_{(i)}) + \psi_{\mathcal{Y}_{(i)}}(\hat{\theta}; x_{(i)}) - \psi_{\mathcal{Y}_{(i)}}(\widetilde{\theta}_{(i)}; x_{(i)}).$$

This statistic hence measures the decrease in fit when the optimal parameter value for the data defined by removing i altogether, $\widetilde{\theta}_{(i)}$, is substituted by the parameter value that is optimal (in the likelihood sense) for the model defined for the data set in its entirety, including i. As for the normalising constants in the MD approach, $\hat{\lambda}(\hat{\theta}, \widetilde{\theta}_{(i)})$ may be estimated using the path sampler, only now the simulated graphs belong to $\mathcal{Y}_{(i)}$.

3.2.2 Taylor series approximations

We may expand $D(\hat{\theta}||\psi)$, around $\hat{\theta}$, and by noting that $\log p_{\hat{\theta}, x}(y) = 0$, disregarding terms of order greater than 2, and rearanging we have the following approximation

to $D(\hat{\theta}||\hat{\theta}_{(i)})$, the missing data generalised Cook's distance (GCD MDMLE)

$$||\hat{\theta}_{(i)} - \hat{\theta}||^2_{I(\hat{\theta})^{-1}} = (\hat{\theta}_{(i)} - \hat{\theta})^{\mathrm{T}} I(\hat{\theta})(\hat{\theta}_{(i)} - \hat{\theta}),$$

saving the effort of calculating ψ. In the sequel we use the notational convention $||u - v||^2_A = (u - v)^{\mathrm{T}} A^{-1}(u - v)$, for $p \times 1$ vectors $u, v \in \mathbb{R}^p$, and positive definite A. In the case of GLM, Lee (1988) states that the likelihood replacement is preferable to this generalised Cooks distance on the ground that there is ambiguity in the choice of scaling matrix. If the expansion is valid it does however justify the use of $I(\hat{\theta})$ (c.p. the use of the normal curvature in Cook, 1986) and using Cook's distance to infer the presence of outliers and influential observations has a long tradition in linear regression and GLMs (cf Hines and Hines, 1995).

Expanding $D(\widetilde{\theta}_{(i)}||\hat{\theta})$, we analogously get

$$||\widetilde{\theta}_{(i)} - \hat{\theta}||^2_{I(\widetilde{\theta}_{(i)})^{-1}}.$$

For the purposes of further approximation, it is a somewhat undesirable feature that the information matrix here depends on $\widetilde{\theta}_{(i)}$. Making the assumption that the curvature in the neighbourhood of $\widetilde{\theta}_{(i)}$ for the model defined on $\mathcal{Y}_{(i)}$ is not too different from the curvature in the neighbourhood of $\hat{\theta}$ for the model defined on $\mathcal{Y}$, we simplify the above expression according to

$$||\widetilde{\theta}_{(i)} - \hat{\theta}||^2_{I(\widetilde{\theta}_{(i)})^{-1}} \approx ||\widetilde{\theta}_{(i)} - \hat{\theta}||^2_{I(\hat{\theta})^{-1}},$$

which we call the available case generalised Cook's distance (GCD ACMLE). These two approximations are expressed in terms of differences in parameter estimates, weighted together by their variation with consideration taken to the association between estimators. We may therefore say that they represent the magnitudes of changes in the effects (self organisation, assortive mixing, etc) we would see as a result of removing an actor.

3.2.3 Approximate generalised Cook's distances by means of the MVP

Refitting the model for every node to obtain the case-deletion parameter estimates is computationally very costly. For other models, one-step estimators have been used to obtain approximate estimates (Lesaffre and Verbeke,1986; Lee, 1988). Here we draw instead on the relationship between the natural parameter and the mean-value parametrisation.

Starting with AC, consider the MVP form of $||\widetilde{\theta}_{(i)} - \hat{\theta}||^2_{I(\hat{\theta})^{-1}}$, with the natural parameter estimate $\widetilde{\theta}_{(i)}$ and $\hat{\theta}$ substituted by their corresponding MVP estimates $\mu_{\mathcal{Y}_i}(\widetilde{\theta}_{(i)}; x_{(i)})$ and $\mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})$, and the MVP Fisher information $I(\hat{\theta})$. This yields the expression

$$||\mu_{\mathcal{Y}_i}(\widetilde{\theta}_{(i)}; x_{(i)}) - \mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})||^2_{I(\hat{\theta})}.$$

As $\widetilde{\theta}_{(i)}$ is the ACMLE, $\mu_{\mathcal{Y}_i}(\widetilde{\theta}_{(i)}; x_{(i)}) = z(y_{(i)}; x_{(i)})$, and hence

$$||\mu_{\mathcal{Y}_i}(\widetilde{\theta}_{(i)}; x_{(i)}) - \mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})||^2_{I(\hat{\theta})} = ||z(y_{(i)}; x_{(i)}) - \mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})||^2_{I(\hat{\theta})},$$

which is referred to as the *approximate available case generalised Cook's distance in mean value parameterisation* (GCD ACMVP). The vectors $z(y_{(i)}; x_{(i)})$ can readily be calculated as described above, $I(\hat{\theta})$ is obtained from fitting the model to y, and $\mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})$ may be approximated by the ergodic mean $\frac{1}{M}\sum_{m=1}^{M} z(u_m; x_{(i)})$ over an MCMC sample $\{u_m\}$ from the model defined by $\hat{\theta}$ on the graph of order $n-1$ with covariates $x_{(i)}$.

For MD, we may analogously consider substituting the natural parameters in $||\hat{\theta}_{(i)} - \hat{\theta}||^2_{I(\hat{\theta})^{-1}}$ by their corresponding MVP estimates, using $||\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) - \mu_{\mathcal{Y}}(\hat{\theta}; x)||^2_{I(\hat{\theta})}$. As before we may use that $\mu_{\mathcal{Y}}(\hat{\theta}; x) = z(y; x)$, and from (4) we see that for $\hat{\theta}_{(i)}$, $\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) = \mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x)$, and hence

$$||\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) - \mu_{\mathcal{Y}}(\hat{\theta}; x)||^2_{I(\hat{\theta})} = ||\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x) - z(y; x)||^2_{I(\hat{\theta})}.$$

To obtain $\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x)$ we would however have to estimate $\hat{\theta}_{(i)}$ first. Denoting the

MD log likelihood $\ell(\theta; y_{(i)}, x) = \log \sum_{u \in \mathcal{Y}^i(y_{(i)})} p_{\theta,x}(u)$, we may consider the Kullback-Leibler divergence in the "other" direction given by

$$D(\hat{\theta}_{(i)}||\hat{\theta}) = E_{\mathcal{Y}_{(i)}}\left[\ell(\hat{\theta}_{(i)}; U, x)\right] - E_{\mathcal{Y}_{(i)}}\left[\ell(\hat{\theta}; U, x)\right],$$

where the expectation $E_{\mathcal{Y}_{(i)}}(g(U)) = \sum_{u \in \mathcal{Y}_{(i)}} e^{\ell(\hat{\theta}_{(i)}; u, x)} g(u)$. The gradient of $D(\hat{\theta}_{(i)}||\theta)$ as a function of θ is $-E_{\mathcal{Y}_{(i)}}[S(\theta; U, x)]$, where $S(\theta; U, x) = \mu_{\mathcal{Y}(U)}(\theta; x) - \mu_{\mathcal{Y}}(\theta; x)$, is the MD score function evaluated in θ. This motivates the use of $\mu_{\mathcal{Y}(y_{(i)})}(\hat{\theta}; x)$ instead of $\mu_{\mathcal{Y}(y_{(i)})}(\hat{\theta}_{(i)}; x)$, giving the following distance measure, *approximate missing data generalised Cook's distance in mean value parameterisation* (GCD MDMVP)

$$||\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}; x) - \mu_{\mathcal{Y}}(\hat{\theta}; x)||^2_{I(\hat{\theta})} = ||\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}; x) - z(y; x)||^2_{I(\hat{\theta})},$$

which only requires some additional simulations to calculate $\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}; x)$. As the Kullback-Leibler divergence in general is not symmetric we would not expect perfect equivalence between DMD and GCD MDMVP. The distributions $e^{\ell(\theta; u, x)}$ and $p_{\theta,x}(u)$ are furthermore defined on different range spaces. However, seeing as the former is the marginalised form of the latter, large differences in DMD would be mirrored by large differences GCD MDMVP. Note that the sample space over which $\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}; x)$ is calculated is considerably smaller than that of $\mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})$. The former is restricted to graphs in $\mathcal{Y}^i(y_{(i)})$, which has cardinality 2^{n-1}, whereas the latter is defined over the whole of $\mathcal{Y}_i$, with cardinality $2^{(n-1)(n-2)/2}$.

For reference we will use a case deletion measure similar to that used by Snijders and Borgatti (1999), that we may call the Jack-knifed distance measure (JN)

$$||z(y_{(i)}; x_{(i)}) - \bar{z}_{AC}||^2_{\Sigma(\bar{z}_{AC})},$$

where

$$\bar{z}_{AC} = \frac{1}{n}\sum_{i=1}^{n} z(y_{(i)}; x_{(i)}), \tag{5}$$

and $\Sigma(\bar{z}_{AC}) = \frac{1}{n}\sum_{i=1}^{n} z(y_{(i)}; x_{(i)})^{\mathrm{T}} z(y_{(i)}; x_{(i)}) - \bar{z}_{AC}^{\mathrm{T}} \bar{z}_{AC}$. When only subgraph census statistics are included in $z(y, x) = z(y)$, the MVP estimate $\mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)}) = \mu_{\mathcal{Y}_i}(\hat{\theta})$ does only depend on the parameter $\hat{\theta}$. The difference between $||\mu_{\mathcal{Y}_i}(\widetilde{\theta}_{(i)}; x_{(i)}) - \mu_{\mathcal{Y}_i}(\hat{\theta}; x_{(i)})||^2_{I(\hat{\theta})}$ and $||z(y_{(i)}; x_{(i)}) - \bar{z}_{AC}||^2_{\Sigma(\bar{z}_{AC})}$, is then likely to be small meaning that GCD AC and the Jack-knifed distance measure more or less coincide.

3.3 MCMC p-values

For the purposes of testing heterogeneity and whether any actor is extreme, we may want to benchmark the observed values against what we expect under a homogeneous ERGM. Let $S_i(y)$ be the value for node i on the measure of interest. The p-value $\Pr(S_i(Y) > S_i(y_{\text{obs}}))$ is not available in an analytically tractable form nor can we rely on standard approximations (such as χ^2). Instead we propose a direct Monte Carlo-based approach, whereby we generate a sample $\{u_m\}_{m=1}^{M}$ from the ERGM under $\hat{\theta}$, and calculate the MCMC p-value as $\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}(S_i(u_m) > S_i(y))$. Any function of the distribution of values may be investigated. If no covariates are used, the observed maximum value $S^{(n)}(y) = \max_i\{S_i(y)\}$ can be compared to the distribution of the maximum, i.e. $(S^{(n)}(u_m))_m^M$. If covariates are used, the ERGM is no longer homogenous and the maximum of the raw measures might be misleading. It is convenient to standardise the values within actors. An example is given in the next section.

The measure has to be recalculated for each $m = 1, \ldots, M$ meaning that it is time-consuming and is best suited for in-depth investigation. Furthermore, the Monte Carlo test does not perfectly mirror the sampling distribution of the measure as the model is not refitted. Thus, in for example $||\mu_{\mathcal{Y}^i(u_{m,(i)})}(\hat{\theta}; x) - z(u_m; x)||^2_{I(\hat{\theta})}$, while $z(u_m; x)$ is the correct estimate in the mean value parametrisation, the conditional mean $\mu_{\mathcal{Y}^i(u_{m,(i)})}(\hat{\theta}; x)$ is based on $\hat{\theta}$ for y rather than u_m.

3.4 Remarks on interpretation of measures

While the measures may be interpreted strictly as measures of the influence of a node on the graph-level statistical inference their interpretation from the perspective

of the ERGM may not be straightforward. We briefly consider here some observations for homogenous ERGM before we investigate the application of the measures to empirical examples.

Consider two two nodes i and j with identical row-vectors $y_{i\bullet}$ and $y_{j\bullet}$ (with adjustment for the elements y_{ij} and y_{ji}). These nodes are structurally equivalent. As a consequence of the permutation invariance of ERGM discussed in the context of residuals in Section 2, for the marginal tie-probabilities $\pi_{ik} = \pi_{jk}$ $(k \neq i, j)$ but for their conditional tie-probabilities we also have $\pi_{ik|-ik} = \pi_{jk|-ik}$. Furthermore, $y_{(i)} = y_{(j)}$ and consequently $\widetilde{\theta}_{(i)} = \widetilde{\theta}_{(j)}$ and $\hat{\theta}_{(i)} = \hat{\theta}_{(j)}$. Hence, for two structurally equivalent nodes the measures developed here will be the same (but may differ between AC and MD).

For a Bernoulli model the MLE is available in closed form and we can write up a closed form expressions for GCD ACMLE and GCD MDMLE. In fact,

$$||\hat{\theta}_{(i)} - \hat{\theta}||^2_{I(\hat{\theta})^{-1}} = ||\widetilde{\theta}_{(i)} - \hat{\theta}||^2_{I(\hat{\theta})^{-1}} = (\bar{L} - y_{i+})^2 n^* L^{-2},$$

where $L = \sum_{1 \leq i < j \leq n} y_{ij}$, $n^* = \binom{n}{2}$, and $\bar{L} = 2L/n$, the average degree. Consequently the influence measure is a curvilinear function of the actor degree and actors with extremely many or extremely few ties are going to be influential. When the degree distribution is skewed to the right this means that high degree nodes are going to be most influential.

For models with more complicated dependence structures it is difficult to say anything about the properties of AC and MD. As discussed in Section 3.1, the model estimated for $\mathcal{Y}$ is misspecified on $\mathcal{Y}_{(i)}$ and as a consequence the interpretation of AC in terms of the traditional case-deletion approach may be confounded by the dependencies of the model.

4. Empirical illustration

Here we provide two examples that help illustrate what type of local structural patterns contribute to large values on the influence statistics. One example flags a truly influential node and the other example a seemingly influential node.

4.1 A collaboration network

Lazega (2001) collected a collaboration network among 36 New England law-firm partners that were located in three different offices and that practiced either litigation or corporate law. The network in Figure 1 displays a high degree of homophily on office but also appears to have homophily on practice.

We fit a so called social circuit dependence model, with $u_{\lambda_t}^{(t)}(y)$ defined as in (2), that has been used for this data set for a number of illustrations (Snijders et al., 2006; Hunter and Handcock, 2006; Handcock and Gile, 2010; van Duijn et al., 2009). The value of the smoothing constant λ_t (Snijders et al., 2006), is commonly set to 2 (Robins and Lusher, 2013:168-171). Here we set $\lambda_t = \exp(0.7781)$ based on the argument in Handcock and Gile (2010) and van Duijn et al. (2009) that this was found to be the MLE by Hunter and Handcock (2006) when λ_t was estimated. The estimation results are provided in Table 1.

The vertex valencies and the calculated influence measures are provided in Table 2. The MCMC error for the path sampler was checked individually for AC, MD and each i to assure that it was negligible in comparison with the respective approximations of the ratios of normalising constants (a total of 2,000 sample points were used and the burn-in for each sample point $150n(n-1).25$).

Throughout we will not attempt to interpret the magnitudes of values in Table 2 (the MVP, for example, tend to be on a different scale to their MCMC equivalents) but focus on the ranking of nodes. Figure 2 plots the values of Table 2 against each other to make consistency across measures more clearly visible. For this particular example, the measures are relatively consistent and pick out node 15 as the top

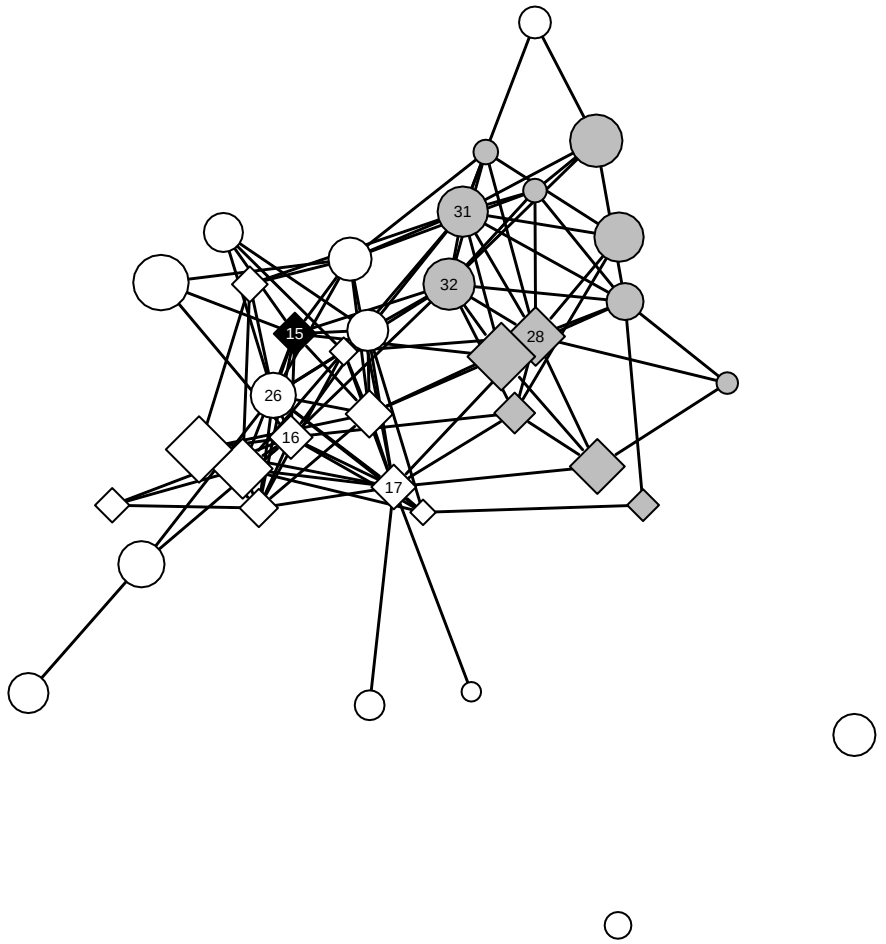

Figure 1: Collaboration network for Lazega's (2001) 36 partners. Colours (white, grey, black) indicating the different offices; size reflecting tenure; practice corporate (diamond) practice litigation (circle). The top 7 highest degree nodes are labeled.

Table 1: Estimates and statistics for Lazega's (2001) partners

	MLE	se	$z(y;x)^*$	$\bar{z}^{**}_{AC}$
density	−6.51	0.571	115	108.611
main seniority	0.852	0.237	130.194	122.961
main practice	0.41	0.115	129	121.833
homophily practice	0.76	0.198	72	68
homophily sex	0.703	0.251	99	93.5
homophily office	1.145	0.19	85	80.278
alternating k-triangle	0.898	0.148	190.306	177.372

* Statistics are defined as in Section 2
** Defined as in Eq. (5)

ranked, though DAC and GCD ACMVP rank node 28 above 15 (more of which will be discussed below). As we would expect, in the scatter plots of Figure 2, the different stages of approximations in the MD approach are largely consistent, and DMD, GCD MDMLE, and GCD MDMVP provide much the same information. Similarly for the AC approach, the measures are internally consistent. The differences between measures corresponding to the MD approach and those of the AC approach echo those of between DAC and DMD, comparing e.g. the top left panel of Figure 2 (GCD MDLE against GCD ACMLE) with the panel in the middle at the far right (DMD against DAC).

Table 2: Influence measures for Lazega's (2001) lawyers 1 through 36

ID	Deg	MDMLE	ACMLE	DMD	DAC	ACMVP	MDMVP	JN
1	1	0.065	0.179	0.038	0.059	0.128	0.054	2.435
2	6	0.156	0.221	0.073	0.074	0.193	0.128	5.437
3	3	0.098	0.1	0.065	0.052	0.079	0.094	2.004
4	9	0.118	0.176	0.059	0.087	0.165	0.124	5.504
5	6	0.371	0.475	0.161	0.2	0.327	0.29	6.173
6	5	0.139	0.157	0.06	0.066	0.18	0.103	3.001
7	2	0.346	0.457	0.169	0.185	0.377	0.236	3.039
8	0	0.192	0.482	0.091	0.194	0.338	0.178	5.811
9	3	0.374	0.807	0.216	0.34	0.432	0.252	1.714
10	5	0.909	1.284	0.51	0.512	0.949	0.674	5.6
11	1	0.112	0.369	0.066	0.148	0.281	0.079	2.39
12	9	0.331	0.654	0.171	0.31	0.705	0.251	8.97
13	2	0.206	0.402	0.073	0.165	0.264	0.153	3.145
14	6	0.056	0.103	0.042	0.043	0.155	0.051	3.211
15	11	3.138	3.009	1.366	1.136	1.754	2.089	27.681
16	13	0.357	0.551	0.184	0.297	0.616	0.257	5.936
17	15	0.309	0.169	0.159	0.086	0.199	0.31	10.555
18	8	0.259	0.314	0.131	0.135	0.295	0.236	4.98
19	10	0.048	0.354	0.018	0.185	0.545	0.043	3.798
20	4	0.027	0.378	0.003	0.145	0.279	0.017	0.826
21	1	0.171	0.66	0.067	0.31	0.601	0.148	2.276
22	9	0.385	0.572	0.184	0.267	0.444	0.308	4.942
23	0	0.283	1.106	0.112	0.468	1.061	0.245	5.811
24	9	0.255	0.374	0.116	0.148	0.34	0.206	5.523
25	5	0.274	0.723	0.14	0.343	0.777	0.219	5.155
26	12	0.542	0.605	0.264	0.276	0.6	0.459	13.28
27	3	0.319	0.169	0.15	0.089	0.149	0.208	6.095
28	13	0.977	1.961	0.456	1.179	2.172	0.711	17.65
29	9	0.957	0.681	0.435	0.264	0.464	0.637	19.178
30	4	0.137	0.441	0.047	0.19	0.233	0.096	1.689
31	13	1.456	1.924	0.65	0.686	1.295	0.831	13.006
32	12	0.659	0.71	0.318	0.314	0.609	0.42	7.975
33	5	0.303	0.815	0.138	0.347	0.82	0.244	5.328
34	6	0.387	1.216	0.226	0.432	0.532	0.282	8.749
35	7	0.576	1.547	0.332	0.666	1.536	0.295	14.015
36	3	0.63	2.691	0.303	0.983	2.044	0.444	2.116

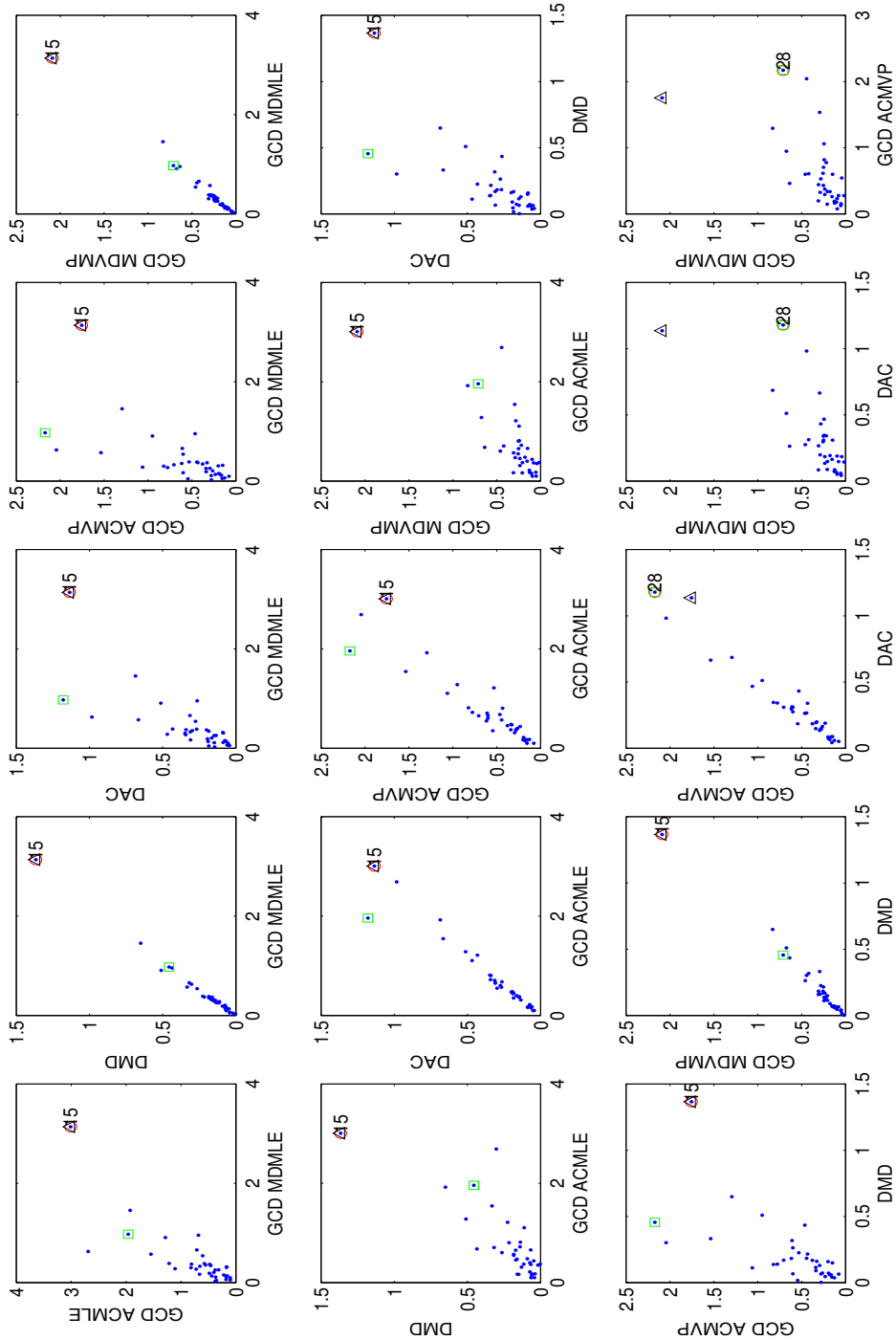

Figure 2: Comparison of influence measures for Lazega's (2001) 36 partners. Actor 15 marked by triangle and actor 28 by square

The influence measures are non-trivial in the sense that they do not merely reflect differences in actor degrees as can be seen from the left hand panels of Figure 3. The ranking is also not immediately visible in the sociogram of Figure 1.

Interpreting the difference between the AC and the MD measures, it is informative to study a plot of GCD ACMLE against GCD MDMLE with marker size proportional to the Jack-Knifed distances as in Figure 4. The two main differences between AC and MD can firstly be said to be that AC, in addition to measuring the extremeness of $y_{i,\bullet}$, also indicates whether an observation has great influence because it has a covariate vector $x_{i,\bullet}$ that is extreme. This is in analogy with GLM where observations may be extreme in terms of the response variable or in the design space. Some care may however be taken in translating this to ERGMs since no clear distinction can be made between exogenous covariates and response variables. This example nonetheless illustrates that this general idea provides insight into the difference between AC and MD. Secondly, something which is harder to parse out, is the fact that the AC model is misspecified under the assumption that the network actually consists of n nodes. If there is evidence of Markov (and social circuit) dependence in y we may rule out "long-range" dependencies in the data generating process (Snijders, 2010). The action of removing an actor i does however induce dependencies among the tie-variables that are not of the type of dependence, that were assumed for y (Markov and social circuit). Loosely speaking, the MD approach is able to pick out interdependencies between tie variables, that should be conditionally independent according to a model defined on the induced subgraph, as stemming from unobserved potential ties, the AC approach is unable to cope with this since it does assume that there are no unobserved tie variables (Koskinen, Robins, and Pattison, 2010). This may also explain why more nodes appear to have large values on AC than MD. These matters are highlighted by a closer inspection of the actors 28, 36, and 35, that have high values on GCD ACMVP but not on GCD MDMVP.

To better understand the differences between the measures, we may consider the influence of different nodes on particular parameter estimates. Figure 5 plots the

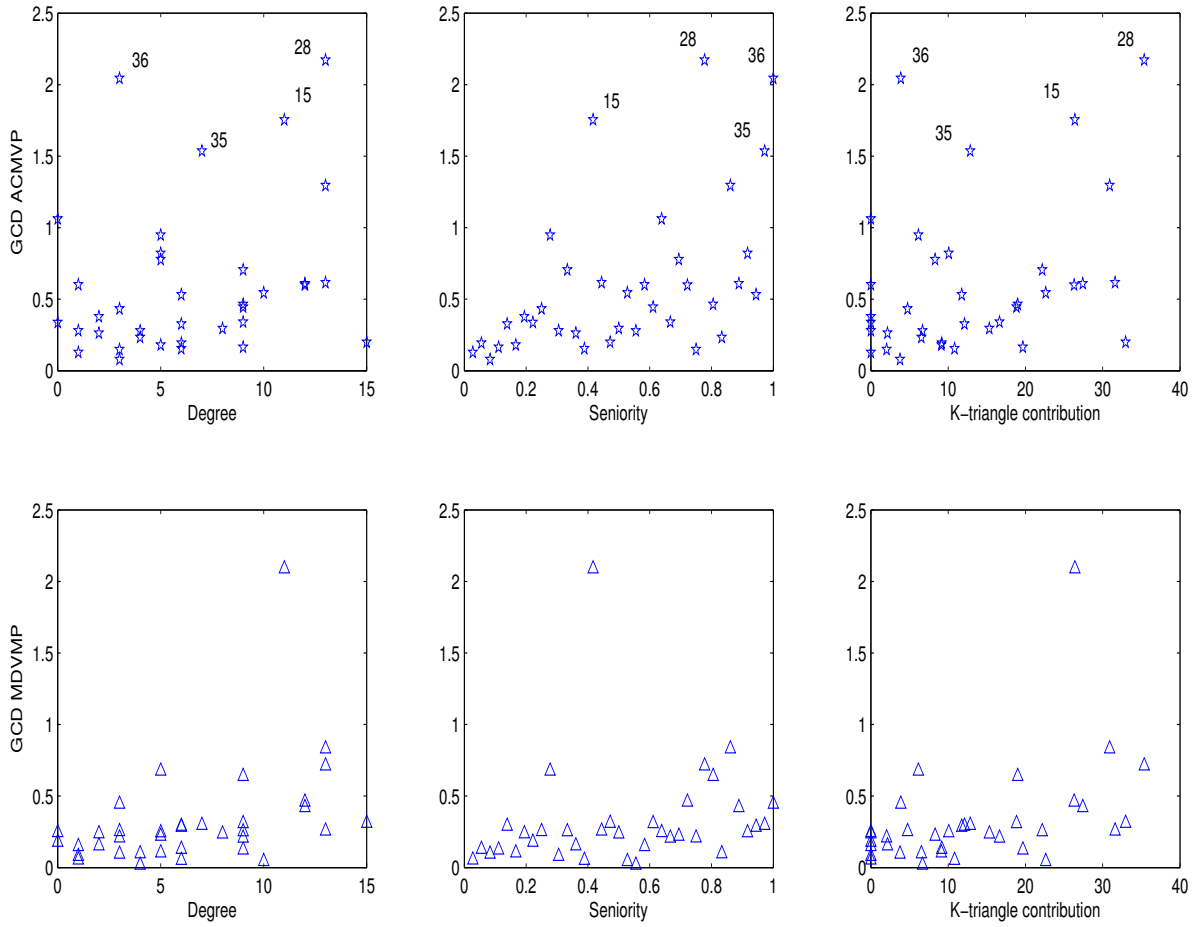

Figure 3: Influence measures against individual attributes (k-triangle contribution given by $z(y_{(i)}; x_{(i)}) - z(y; x)$) for Lazega's (2001) 36 partners with some key actors indicated

case-deletion maximum likelihood estimates for all parameters and nodes. As noted in Section 3.1, we could base a DFBETA-like measure (Belsley et al., 1980) on these individual estimates. Here we only use the estimates of Figure 5 contribute in a different way for AC and MD.

Judging by Figure 4 actor 15 has high values on all of GCD ACMVP, GCD MDMVP, and JN. Because of the "response" $y_{15,\bullet}$, the parameter estimates change a great deal when 15 is removed by either AC or MD. As seen from Figure 3, 15 contributes highly to the density and the clustering (as measured by contribution to the k-triangle count), which is reflected in the corresponding estimates in Figure 5, panels (a) and (g), respectively, for both MDMLE and ACMLE. As 15 is the only actor in a particular office, none of the ties in $y_{15,\bullet}$ contribute towards the homophily effect for office meaning that the estimate for this effect is greatly increased when 15 is removed. This is clearly demonstrated in panel (f) of Figure 5. The change in estimates, and thereby the improved fit, is not greatly altered by the choice of removal method and both AC and MD rate 15 highly influential. The differences in contribution to $z(y; x)$ of 15 is also the most "unusual" given the model, wherefore 15 also has the greatest JN.

The reason 28 inches past 15 in AC can be summarised: 28 contributes greatly to density and clustering (Figure 3) but when, as in the MD approach, the attributes of 28 (high seniority, corporate practice, male, etc) are taken into account $y_{28,\bullet}$ is not as extreme. Actor 28 is still highly influential according to both methods and changes the estimates greatly (Figure 5). Actor 28 sits in a highly triangulated region of the graph and when removed using AC many ties are left unexplained. A symptom of this could be that the change in the k-triangle statistic is much greater in the AC approach for 28 than for MD (Figure 5 g).

Actor 36 is ranked 2nd by GCD ACMVP and 7th by GCD MDMVP, and 35 is ranked 4th and 11th, respectively. Both actors have low JN, 36 in particular has extremely low JN. The reasons for the discrepancy between AC and MD are the same for these two actors but the tendencies are stronger for 36. Looking at Figure 3 we

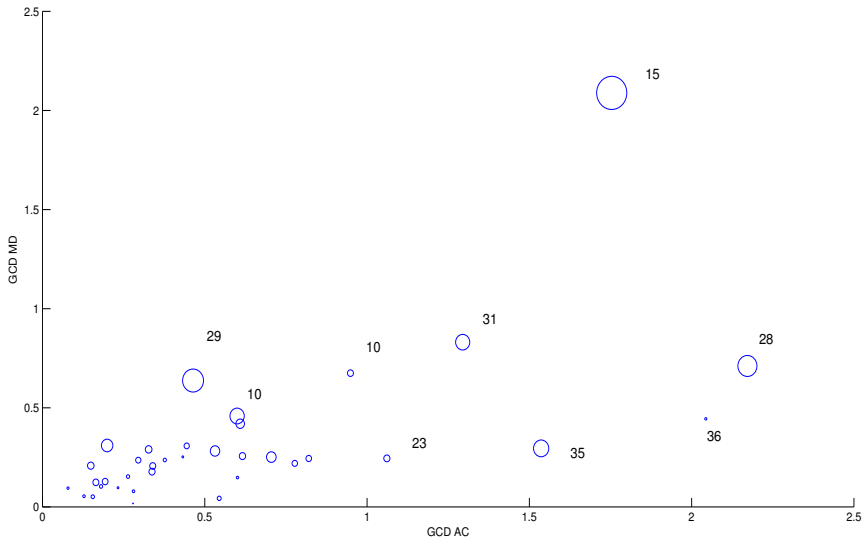

Figure 4: The two approximate generalised Cook's distances for AC and MD, with circle size proportional to Jack-knifed distance, with key actors indicated

see that their degrees are low and the contribution towards clustering small. Actors 36 and 35 are extreme in the attribute space since they are the most and second most senior partners in the firm (middle upper panel of Figure 3). The extreme seniority in combination with relatively few ties means that removal of these actors would result in a substantial increase in the estimate of the main effect of senority (panel (b) Figure 5). In the case of 36, JN is low since the extreme seniority is counteracted by the low contribution to the main effect of seniority.

To test whether the observed value for actor 15 is extreme, we calculate the MCMC p-value for the influence statistic of all actors. For each u_m $(m = 1, \ldots, 1000)$, let $\delta_m^i = \text{abs}(S_i(u_m) - \bar{S}_i)/\tilde{S}_i$, and let $\delta_m^{(1)}, \ldots, \delta_m^{(n)}$ be the ordered values in increasing order. The mean $\bar{S}_i = 1/M \sum_m^M S_i(u_m)$ and standard deviation $\tilde{S}_i = [1/M \sum_m^M (S_i(u_m) - \bar{S}_i)^2]^{1/2}$ are the MCMC equivalents of the node-specific expected values and standard deviations, respectively. Figure 6 provides the MCMC distribution of the maximum $\delta_m^{(n)}$. The MCMC p-value for node 15 is 0.045 suggesting

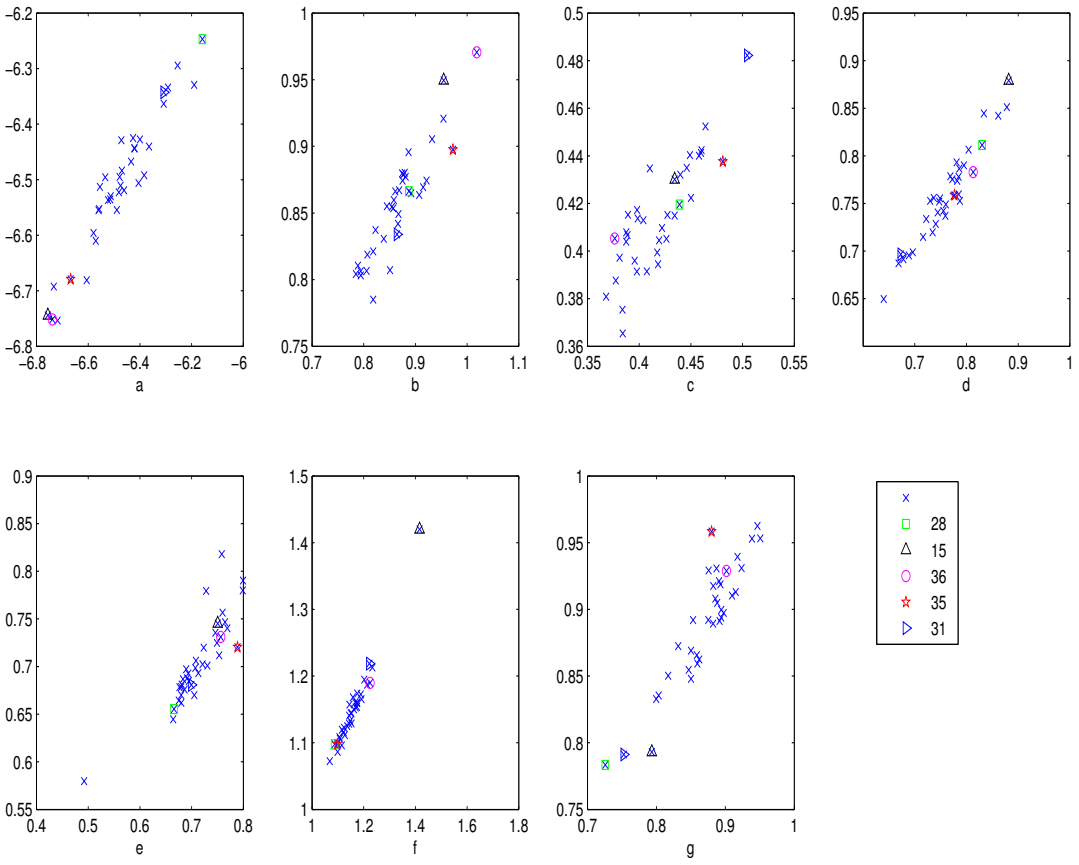

Figure 5: ACMLE (horizontal axes) against MDMLE (vertical) for parameters (a) density; main effects of (b) seniority; and, (c) practice; homophily on (d) practice; (e) sex; and, (f) office; and, (g) alternating k-triangle. through g with key actors indicated

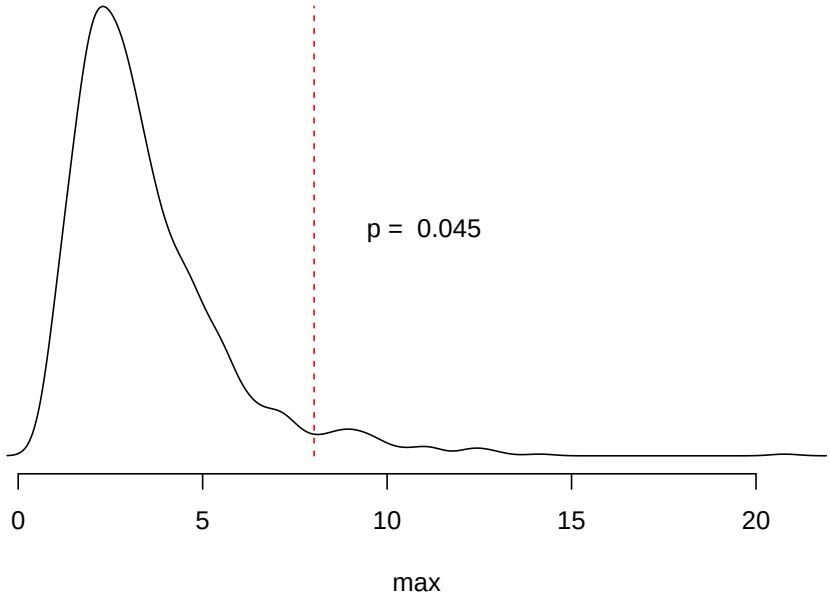

Figure 6: Distribution of node-standardised maximal GCD MDMVP with MCMC p-value and reference line for actor 15

that the actor has a significant impact on the model. The MCMC p-value for $i = 15$ marginally, $\Pr(S_{15}(Y) > S_{15}(y))$, is 0.001.

4.2 A covert network

As a second example we consider the case of the revolutionary organisation November 17 which was active in Greece between 1975 and 2002 (Nomikos, 2007). The network (November 17 Organization Aggregate Attack Series) was obtained from the John Jay & ARTIS Transnational Terrorism Database (JJATT 2009) and consists of 18 actors with a total of 46 ties. Ties are defined as present if there in 1995-2002 was evidence of two individuals being either (1) Acquaintances/Distant family; (2)

Friends/Moderately close family; or (3) Close Friends/Family and Tight-knit operational cliques[1]. The network sociogram is provided in Figure 7. From the sociogram it seems clear that node 1 is the most central node in the network.

Table 3 provides the parameter estimates for a social circuit ERGM fitted to the network that provided adequate fit (Robins and Lusher, 2013:184-185) for different structural features of the network. We focus here only on GCD MDMVP. The values

Table 3: Estimates and statistics* for the November 17 network

	MLE	se	$z(y)$
density	-1.407	1.258	46
alt. k-stars	-1.1974	0.423	120.91
alt. k-tri	2.1890	0.462	75.33

* Statistics are defined as in Section 2

on GCD MDMVP for the actors are provided in Table 4.2 along with the values of $\mu_{Y^i(y_{(i)})}(\hat{\theta}, x)$.

The observed GCD MDMVP for actor 1 (2.96) is almost three times as large as the second highest value, which is recorded for actor 3 (1.095). Actor 1 also has the largest degree (and is in fact connected to all other actors) and actor 3 has the second highest degree of 10. As the assumed model is homogenous we can test the influence statistic using the MCMC p-value $\Pr(S^{(n)}(Y) > S^{(n)}(y_{\text{jobs}}))$. Here this simulated p-value is 0.133 which is illustrated with reference to the CDF in Figure 8. The conclusion can be said to be that while actor 1 clearly is different from the other actors we would expect the model to generate such extreme actors.

The difference between the two examples is revealing. In the first case of the lawyer network, we identified an individual who we inferred (using the MCMC p-value) was

[1]Ties were interactions that were (1) limited to radical organisation activities; (2) extend beyond radical organisations to include such categories as co-workers and roommates; (3) those that would die for each other

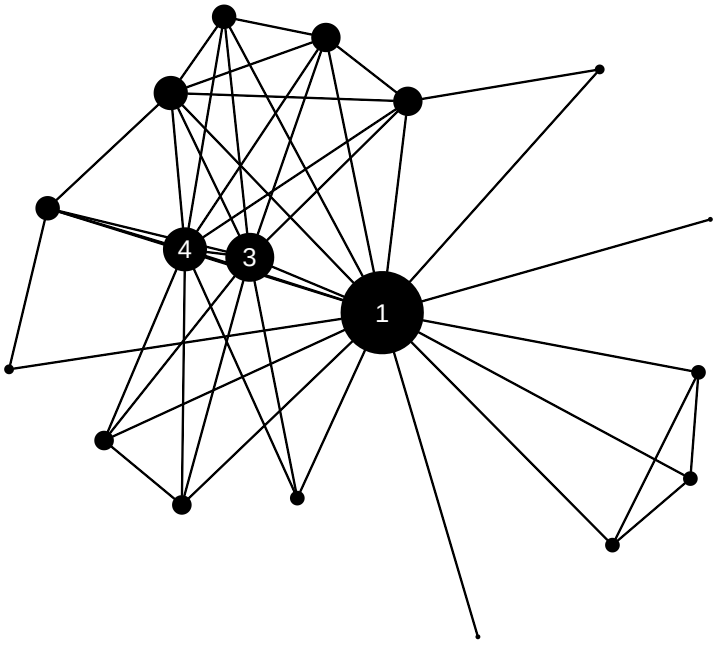

Figure 7: November 17 network. Node-size proportional to degree centrality

Table 4: GCD MDMVP and $\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}, x)$ in the November 17 network

ID	GCD	density	alt. k-stars	alt. k-tri
1	2.963	33.20	78.05	50.37
2	0.822	49.85	133.72	84.45
3	1.095	41.66	103.49	64.03
4	1.062	41.68	103.57	64.13
5	0.150	46.06	120.08	75.15
6	0.242	45.10	116.43	72.48
7	0.069	45.87	119.90	75.18
8	0.116	44.87	116.03	72.70
9	0.580	48.98	130.88	82.59
10	0.168	48.18	128.84	79.24
11	0.187	48.20	128.93	79.24
12	0.063	47.09	124.15	77.22
13	0.165	48.12	128.66	79.14
14	0.759	49.76	133.46	84.22
15	0.483	48.72	130.02	81.96
16	0.318	47.78	126.58	79.71
17	0.064	47.08	124.12	77.16
18	0.413	44.27	113.17	70.40

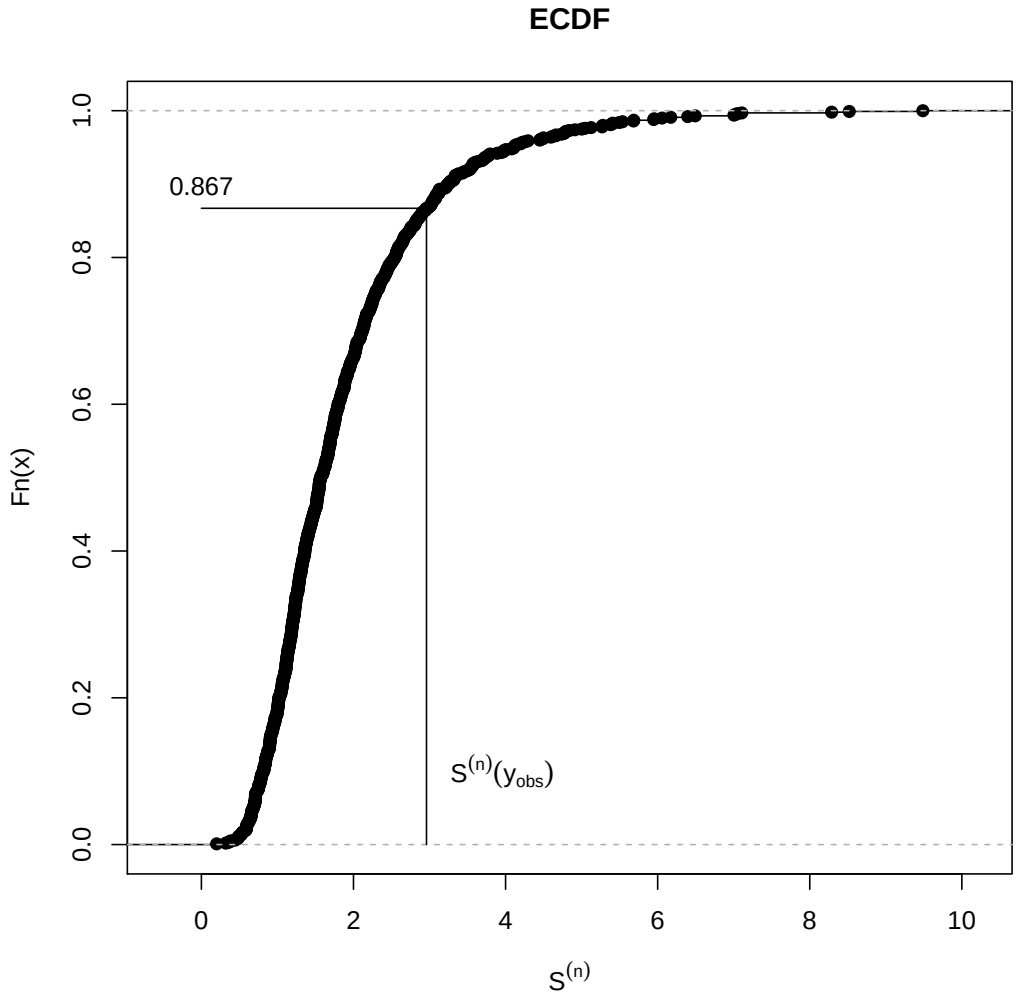

Figure 8: ECDF of maximal GCD MDMVP with MCMC p-value and reference line for actor 1 for November 17 network

extreme compared to other actors. In such a case we might conclude that there is something quite distinctive about this individual that determines a significant influence on the network structure. In the second case, the method identifies one actor who in fact has ties to all others: that is, the highest possible degree centrality. Yet, the MCMC p-value is not significant. So we do not have convincing evidence that this actor is significantly extreme, despite the highest possible degree. As with most statistical tests, the lack of evidence could be related to poor statistical power and the November 17 network is considerably smaller than the Lazega law-firm partner network. The inference we make is that, assuming the social processes that produced the network continue to operate, the "removal" of this actor would likely see another actor move into a highly central position. Indeed, when we simulate networks from the model, it is not unusual to produce one highly central node, even though the highest degree centrality is not always achieved. In short, in the first case, the network is altered substantially; in the second, the network largely reconstitutes itself. These examples illustrate how a model-based analysis of extreme actors goes beyond conclusions based on examination of standard centrality scores.

5. Concluding discussion

We have proposed a methodology for studying the influence of observations on parameter estimates in ERGMs. The methodology relies on defining observations at the level of the actor and thus investigating the influence on the model exerted by the individuals in the network. The influence is measured as the change in parameter estimates that would result under either one of two case deletion strategies: the missing data (MD) approach and the available case (AC) approach. For each of the case deletion strategies we have defined two influence measures that approximate the decrease in deviance, or equivalently, the Kullback-Leibler divergence, for the model defined by the estimates obtained from the respective case deletion strategies. The two measures are particularly useful in investigating influence as a routine application

when fitting ERGMs since these do not require refitting of the model. The AC approach offers a heuristic interpretation in terms of what the structure of the network would be if an actor was completely exogenous to the network and simply were left out. The MD approach has the benefit of being more principled, measuring the difference between networks of the same size. While the AC approach does not take into account that an actor may be extreme only because of their covariates, the MD approach does. The MD approach is implemented in MPNet (Wang, et al., 2014) as is the MCMC p-value scheme. The former can used routinely in ERGM analysis and the latter, which is computationally intensive, can be used for further investigations.

The proposed influence measures may heuristically be thought of as indices of model-based centrality - to what degree does our analysis depend on specific individuals in the sense that their exclusion would change the estimates greatly in the directions that "matter". Delving deeper into the issues of what constitutes influential actors and outliers in the case of the ERGM and how this relates to the concept of centrality (Freeman, 1978; Borgatti and Everett, 2006; Schoch and Brandes, 2015), raises some fundamental issues regarding statistical models for social networks. When fitting an ERGM to a network, what does it mean that an actor is atypical or typical? Robins, Pattison and Woolcock (2005) provide numerous examples of how normal ERGMs may generate extreme nodes. Embedded in these questions are issues of how the ERGM scales and how a stochastically homogeneous network relates to larger networks in which it might be embedded (Schweinberger et al., 2017; c.p. the closely related so called boundary issue, Laumann et al., 1983; and in the context of ERGM, see Koskinen, Robins, Wang, and Pattison, 2013). From the perspective of interpreting the ERGM as a data-generating process, an actor with a large value on the statistic is 'playing by different rules' to the other actors. As such, identifying an 'extreme actor' may indicate that the node should be treated as exogenous in the analysis. Considering a government agency, an ERGM is not a suitable model for explaining the ties of the president - people may have ties to the president because this is the president, not because of endogenous tie-mechanisms. In our empirical

illustrations using a collaboration network, there was one actor that clearly affected the structure of the network. In this case it might mean that if this actor were removed, then the collaborations would be organised differently. For the Revolutionary 17 November network we can however conclude that if the actor that appeared most extreme were removed, then his position would be replaced by another actor.

We believe that the proposed influence measure will prove a useful tool in further investigating these issues. The properties of these influence measures also warrant further investigation to assess what information they might provide beyond what may be motivated strictly from a statistically principled perspective.

References

Anderson, B. S., Butts, C., and Carley, K. (1999), "The interaction of size and density with graph-level indices," *Social Networks*, 21 , 239-267.

Barndorff-Nielsen, O. E. (1978), *Information and Exponential Families in Statistical Theory*, New York: Wiley.

Belsley, D.A., Kuh, E., Welsh, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley Series in Probability and Mathematical Statistics. New York: John Wiley.

Besag, J. 1974. "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society B*, 36,96–127.

Block, P., Koskinen, J.H., Stadtfeld, C. J., Hollway, J., Steglich, C. (in press), "Change we can believe in: Comparing Longitudinal Network Models on Consistency, Interpretability and Predictive Power," *Social Networks*.

Borgatti, S. P., and Everett, M. G., (2006), "A Graph-theoretic perspective on centrality," *Social Networks*, 28, 466-484.

Chatterjee, S., and Hadi, A. S. (1988), *Senisitivity Anialysis in Linear Regression*, New York: John Wiley.

Cook, R.D. (1977), "Detection of influential observations in linear regression," *Technometrics*, 19, 15–18.

Cook, R.D. (1986), "Assessment of local influence," *Journal of the Royal Statistical Society*, Series B, 48, 133–169.

Corander, J. and Dahmström, K. and Dahmström, P. (1998), "Maximum likelihood estimation for Markov graphs," Research report, 1998:8, Stockholm University, Department of Statistics.

Corander, J., and Dahmström, K., and Dahmström, P. (2002), "Maximum likelihood estimation for exponential random graph model," pp:1-17 in Jan Hagberg (ed.), *Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank*, University of Stockholm: Department of Statistics.

Crouch, B., Wasserman, S., and Trachtenberg, F. (1998), "Markov Chain Monte Carlo maximum likelihood estimation for p* social network models," Paper presented at the Sunbelt XVIII and Fifth European International Social Networks Conference, Sitges (Spain), May 28–31, 1998.

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81, 832–842.

Freeman, L.C. (1978), "Centrality in social networks conceptual clarification," *Social networks*, 1, 215–239.

Gelman, A., and Meng, X. L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Handcock, M. S. (2003). "Assessing degeneracy in statistical models of social networks," Working Paper no. 39, Center for Statistics and the Social Sciences, University of Washington. (Available from http://www.csss.washington.edu/Papers/wp39.pdf)

Handcock, M, and Gile, K. (2010). "Modeling social networks from sampled data," *The Annals of Applied Statistics*, 4, 5–25.

Hines, R.O.H., and Hines, W.G.S. (1995), "Exploring Cook's Statistic Graphically," *The American Statistician*, 49, 389–394.

Hines, R.O.H., Lawless, J.F., and Carter, E.M. (1992), "Diagnostics for a cumulative multinomial generalized linear model, with applications to grouped toxicological mortality data," *Journal of the American Statistical Association*, 87, 1059–1069.

Holland, P., and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs" (with discussion), *Journal of the American Statistical Association*, 76, 33–65.

Huisman, M. (2009). "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*, 10(1).

Hunter, D. R., and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Jonasson (1999), " The random triangle model.," *Journal of Applied Probability* , 36, 852–876.

The John Jay & ARTIS Transnational Terrorism Database, JJATT 2009. `http://doitapps.jjay.cuny.edu/jjatt/data.php`

Koskinen, J., Robins, G., Wang, P., Pattison, P. E., (2013), "Bayesian analysis for partially observed network data, missing ties, attributes and actors," *Social Networks*, 35(4), 514–527.

Koskinen, J., Robins, G., Pattison, P. E., (2010), "Analysing Exponential Random Graph (p-star) Models with Missing Data Using Bayesian Data Augmentation," *Statistical Methodology*, 7(3), 366–384.

Koskinen, J., and Snijders, T.A.B., (2013). Simulation, Estimation and Goodness of Fit, pp141-166 in Lusher, Koskinen, Robins, (Eds.). *Exponential Random Graph Models for Social Networks: Theory, Methods and Applications*, Cambridge University Press, NY.

Kuhnt, S. (2004), "Outlier identification procedures for contingency tables using maximum likelihood and L_1 estimates," *Scandinavian Journal of Statistics*, 31, 431–442.

Laumann, E. O., Marsden, P. V., Prensky, D., (1983), "The boundary specification problem in network analysis," In:Burt, R. S., Minor, M. J. (Eds.), *Applied Network Analysis*, Sage Publications, London, pp. 18–34.

Lazega, E. (2001), *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*, Oxford University Press, Oxford.

Lee, A. H. (1988), "Partial Influence in Generalized Linear Models," *Biometrics*, 44, 71–77.

Lehmann, E.L. (1983), *Theory of Point Estimation*, New York: Wiley.

Lesaffre, E., and Albert, A. (1989), "Multiple-Group Logistic Regression Diagnostics," *Applied Statistics*, 38, 425–440.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*,New York: John Wiley.

Lusher, D., Koskinen, J., and Robins, G.L. (2013), *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge, UK .

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001), "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 27, 415–444.

Meng, X.-L., and Wong, W. H. (1996), "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Neal, R. M. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Technical Report CRG–TR–93–1, Department of Statistics, University of Toronto. (available from http://www.cs.utoronto.ca/r̃adford/)

Nomikos, John M. (2007), "Terrorism, Media, and Intelligence in Greece: Capturing the 17 November Group," *International Journal of Intelligence and CounterIntelligence*, 20 (1), 65–78.

Pattison, P. E., Wasserman, S. (1999), "Logit Models and Logistic Regressions for Social Networks: II. Multivariate Relations," *British Journal of Mathematical and Statistical Psychology*, 52, 169–193.

Pierce, D. A. and Schafer, D. W. (1986), "Residuals in generalized linear models," *Journal of the American Statistical Association*, 81, 977–986.

Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705–724.

Robins, G. L., and Daraganova, G. (2013), "Social selection, dyadic covariates, and geospatial effects" pp. 91–101 in: Lusher, Koskinen, Robins (Eds.) *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge, UK .

Robins, G. L., and Lusher, D. (2013), "Illustrations: Simulation, Estimation, and Goodness of Fit," pp. 167–185 in: Lusher, Koskinen, Robins (Eds.) *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge, UK .

Robins, G. L., and Morris, M. (2007), "Advances in Exponential Random Graph (p*) Models," *Social Networks*, 29, 169–172 .

Robins, G.L., Elliott, P., & Pattison, P.E. (2001), "Network models for social selection processes," *Social networks*, 23, 1–30.

Robins, G. L., Pattison, P. E., and Elliot, P. (2001), "Network Models for Social Influence Processes," *Psychometrika*, 66, 161–190.

Robins, G. L., Pattison, P. E., and Woolcock, J. (2005), "Small and other worlds: Global network structures from local processes," *American Journal of Sociology*, 110, 894–936.

Rubin, D. B. (1976), "Inference and Missing Data (with discussion)," *Biometrika*, 63, 581–592.

Schoch, D., and Brandes, U. (2015). "Stars, neighborhood inclusion, and network centrality," *SIAM Workshop on Network Science*

Shalizi, C.R., Rinaldo, A., (2013)." Consistency under sampling of exponential random graph models," *The Annals of Statistics*, 41, 508–535.

Snijders, T.A.B., (2010). "Conditional marginalization for exponential random graph models," *Journal of Mathematical Sociology*, 34, 239–252.

Snijders, T.A.B., (2002), "Markov chain Monte Carlo estimation of exponential random graph models," *Journal of Social Structure*, 3(2), April.

Snijders, T.A.B., and Borgatti, S. P., (1999), "Non-parametric standard errors and tests for network statistics," *Connections*, 22, 61–70.

Snijders, T.A.B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153.

Schweinberger, M. (2011), "Instability, sensitivity, and degeneracy of discrete exponential families," *Journal of the American Statistical Association*, 106, 1361–1370.

Schweinberger, M., Krivitsky, P. N. , and Butts, C. T. (2017), "Foundations of Finite-, Super-, and Infinite-Population Random Graph Inference," *arXiv:1707.04800v1*

Strauss, D. (1986), " On a general class of models for interaction," *SIAM Review*, 28, 513–527.

van Duijn, M.A.J. and Gile, K.J. and Handcock, M.S., (2009), "A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models," *Social Networks*, 31(1), 52–62.

Wang, P., Robins, G., Pattison, P., and Koskinen, J. (2014). *MPNet, Program for the Simulation and Estimation of (p^*) Exponential Random Graph Models for Multilevel Networks: USER MANUAL*. Melbourne School of Psychological Sciences The University of Melbourne Australia.

Wang, P., Pattison, P., Robins, G. (2013), "Exponential random graph model specifications for bipartite networks – A dependence hierarchy, " *Social networks*. 35(2), 211–222.

Wasserman, S., Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.

Wasserman, S., and Pattison, P. E. (1996), "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p*," *Psychometrika*, 61, 401–425.

Waternaux, C., Laird, N.M., and Ware, J.H. (1989), "Methods for analysis of longitudinal data: blood-lead concentrations and cognitive development," *Journal of the American Statistical Association*, 84, 33–41.

Weiss, R.E., and Lazaro, C.G. (1992), "Residual plots for repeated measures," *Statistics in Medicine*, 11, 115–124.

Williams, D. A. (1984), "Residuals in generalized linear models," In Proceedings of the XIIth International Biometric Conference, Tokyo, 59–68.

Williams, D. A. (1987), "Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions," *Applied Statistics*, 36, 181–191.