Extreme Actors - Outliers and Influential Observations in exponential random graph (p-star) models

Johan Koskinen [*], Peng Wang, Garry Robins, and Philippa Pattison
MelNet Social Networks Laboratory Technical Report 08–05
Department of Psychology, School of Behavioural Science
University of Melbourne, Parkville Victoria 3010, Australia

November 29, 2008

[*] E-mail: johank@unimelb.edu.au.

Abstract

We discuss how the statistical notions of influential observations and outliers may apply to the statistical models for social networks known as exponential family random graph models (ERGMs or p-star). We argue that the most fruitful approach is to consider observations on the level of the actors and that if a removal of an actor would result in a radically different model, then it could be said that the inferred characteristics of the network depend greatly on this actor and hence the actor is in some sense "central" with respect to the structural features that "matter". To define how to "remove" an actor is not straightforward because of the usually high degree of interdependence among tie-variables in ERGMs. We construe the removal in terms of two case deletion strategies: the interaction (tie-variables) of an actor is assumed to be unobserved or, removal results in the subgraph induced by removing the corresponding vertex. 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 parametrisation, representing varying degrees of approximation. We arrive at several measures of influence, three of which do not require refitting of the model to the case deleted data and hence these lend themselves to routine application in the ERGM fitting procedure. The influence measures are applied to a well known data set to illustrate the information they provide.

1. Introduction

Statistical models are simplified stochastic models of what we might observe and the choice of the best fitting model changes when we *do* observe something new. Hence, that estimates should change with additional observations is in the nature of the statistical model. 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, something which might deflate our overall confidence in the model. 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 O'Hara 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; Robins, Pattison, and Wasserman, 1999; Snijders et al., 2006; Hunter and Handcock, 2006), or p^* models, have proved promising for capturing the complex dependencies giving rise to observable tie variables in social networks (Robins and Morris, 2007). The statistical literature on influence has largely drawn on linear regression and therefore been concerned with defining analogies to residuals that may be used to study how, for example, case deletion changes the deviance. 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.

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 differences 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. 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 out-liers - 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.

Nevertheless, we may speak of actors that are central to the structure of the network. Given the premis that a particular ERGM parameterisation adequately describes all the relevant structural processes that gave rise to the network, some actors may be central to this structure whereas others may be peripheral. If the structure - as described by an ERGM - does not change substantively when an actor is, in some way, removed, then this actor is not solely important to the structure of the network. For instance, actors that are structurally equivalent, may still be important to the structure in the sense that the roles of these actor may be crucial, but there may be many actors that could fulfil that role. From this perspective, an influence measure devised for ERGMs could be interpreted as form of centrality index, an index of structural centrality, to identify actors that who alone are important to the structure as described by the model. The concept of centrality is well established in the social networks literature, is well understood and has been rigorously defined (Freeman, 1979; Borgatti and Everett, 2006). Here we leave the connection between centrality and structural influence at the intuitive level and do not offer more than a heuristic interpretation.

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 their corresponding 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. The measures are then applied to a well-known data set with a thoroughly researched model-specification.

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. The entries of y are typically not independent and if Y_A and Y_B are the collection of variables corresponding to distinct 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 to constitute our observations. Considering for example the residual $e_{ij} = y_{ij} - \hat{p}_{ij}$, for a dyad, where $\Pr(Y_{ij} = 1|\hat{\theta}, Y_{\mathcal{E} \setminus \{i,j\}} = y_{\mathcal{E} \setminus \{i,j\}})$ (this is of course residual defined in an intuitive sense and for GLMs many forms may be considered; see, for example, Williams, 1984, and Pierce and Schafer, 1986; tie-based residuals have also been used in the ERGM framework for model selection, e.g. Wasserman and Pattison, 1996). In the case of graphs, in many instances, e_{ij} may be too many to be of any practical use but more serious is that the residuals have to be interpreted with respect to all other observations $y_{\mathcal{E} \setminus \{i,j\}}$ because of the conditioning. 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 (this relates to the permutation invariance induced by the homogeneity assumptions, according to which all isomorphic graphs are equally probable, Frank and Strauss, 1986). Inspecting joint residuals is simply too costly as there are too many combinations to consider.

A natural way of grouping the variables in Y is by the nodes. In particular 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 $x_{(i)}$. We assume that z may unambiguously be define 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. It is important to point out however that $z(y_{(i)}; x_{(i)})$ is not the same as evaluating $z(y^*; x)$ for an adjacency matrix whose elements are equal to those of $y_{(i)}$ for all $\{k, \ell\}$ such that $\{k, \ell\} \cap \{i\} = \emptyset$ but with elements $y^*_{ij} = 0$. While for subgraph census statistics that do not depend on the order of the graph this may hold but if, say, a count of the number of isolates is part of z, then this might not be the case. 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 $x_{i\bullet}$, and $y_{i+} = \sum_{j \in V \setminus \{i\}} y_{ij}$.

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

$$\mu_{\mathcal{Y}}(\hat{\theta}; x) = z(y; x), \tag{1}$$

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 (1) 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; 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 have 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, 2007, 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 (2007). Hence the name MD approach.

In the second approach node i is removed entirely from the network as are its covariate values, so that we instead of having the observations y and x 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 little more involved but similarily entails finding the estimate $\hat{\theta}_{(i)}$ that satistfies

$$\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) = \mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x), \tag{2}$$

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. The equation (2) may be solved by stochastic maximisation (Snijders, 2007), or by approximating the expectancies using importance samples (as suggested in Handcock and Gile, 2007). The negative of the Hessian is straightforwad to obtain as $Cov_{\hat{\theta}_{(i)}}\{z(Y; x)|x\} - Cov_{\hat{\theta}_{(i)}}\{z(Y; x)|x, Y_{(i)} = y_{(i)}\}$ (for a Bayesian inference scheme see Koskinen, Robins, and Pattison, 2008b).

AC is easier to understand since $z(y_{(i)}; x_{(i)})$ is still defined in terms of a regular ERGM and the removal principle is straightforward and natural. It is however not well understood how ERGMs for graphs of different order relate to each other (Anderson, Butts, and Carley, 1999; Robins, Pattison & Woolcock, 2005). In some instance it may therefore be hard to motivate comparing two estimates $\widetilde{\theta}_{(i)}$ and $\hat{\theta}$ that are based on networks of different order. Furthermore, if we posit that y is generated from an ERGM then, strictly speaking, $y_{(i)}$ in most cases cannot be (in brief, a subgraph of an ERGM does not necessarily follow the ERGM; see Koskinen, Robins, and Pattison, 2008a).

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 element 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. Furthermore, we may not know what parameters are most "important" so that we we do not know what weight should be given to the deviations on the different elements of θ. Finally, the estimates are typically highly correlated, wherefore it may be hard to parse out the influence of actors on individual dimensions of θ.

3.2.1 Kullback-Leibler divergence

MD A common way of investigating similarity between distributions with probability mass functions $p(u)$ and $q(u)$, where p is dominated by q, is by using the Kullback-Leibler divergence $D(p||q) = E_{U|p}\{\log(p(U)/q(U))\}$. Note that the Kullback-Leibler divergence may be rewritten $H(p, q) - H(p)$, where $H(p, q)$ is commonly referred to as the cross entropy and $H(p)$ 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 $E_{U|p}\{z(U)\} = \mu$. 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 before in GLMs when $p > 1$ (see e.g. Williams, 1987; Lee, 1988). Handcock and Gile (2007) 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 (2) 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 (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 up 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 measure 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 sequal 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. Lee (1988), casually observes that this is a generalised form of the Cook's distance, something which inspires confidence as the use of Cook's distance to infer the presence of outliers and influential observations has a long tradition in linear regression and GLMs (cf O'Hara 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, undoubtedly unrealistic, 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

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 $||\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) - \mu_{\mathcal{Y}}(\hat{\theta}; x)||^2_{I(\hat{\theta})^{-1}}$. As before we may use that $\mu_{\mathcal{Y}}(\hat{\theta}; x) = z(y; x)$, and from (2) we se that for $\hat{\theta}_{(i)}$, $\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) = \mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x)$, and hence

$$||\mu_{\mathcal{Y}}(\hat{\theta}_{(i)}; x) - \mu_{\mathcal{Y}}(\hat{\theta}; x)||^2_{I(\hat{\theta})^{-1}} = ||\mu_{\mathcal{Y}^i(y_{(i)})}(\hat{\theta}_{(i)}; x) - z(y; x)||^2_{I(\hat{\theta})^{-1}}.$$

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 $2^{(n-1)(n-2)/2}$.

To complement these measures, it is also useful to consider the differences between the $z(y_{(i)};x_{(i)})$'s, with respect to observed variation in these statistics among the actors. This is a case deletion strategy similar to that used by Snijders and Borgatti (1999), and we may call it the Jack-knifed distance measure (JN)

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

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

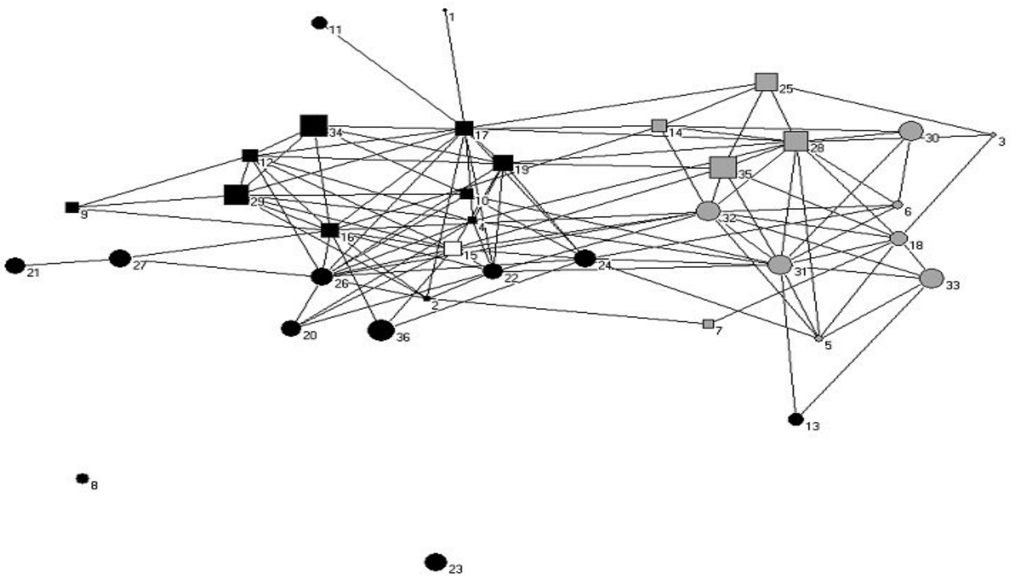

Figure 1: Collaboration network for Lazega's (2001) 36 partners. Office Boston (black), Hartford (grey), and Providence (white); size proportional to seniority; practice corporate (ellipse) practice litigation (box). Actors 27, 29, and 34 are women, all other actors men

4. Empirical illustration

A model was fitted to Lazega's (2001) lawfirm partners (sociogram in Figure 1), which yielded the estimates in Table 1. The model is a so called social circuit dependence model and has together with this data set been used to illustrate various aspects of ERGMs (Snijders et al., 2006; Hunter and Handcock, 2006; Handcock and Gile, 2007; van Duij et al., 2007). The value of the smoothing constant λ (Snijders et al., 2006), typically set arbitrarily by the researcher (Robins and Morris, 2007), is here set to exp(0.7781) (using the same argument as in Handcock and Gile, 2007, and van Duij et al., 2007, this value was set to the corresponding MLE for the model when λ is treated as a free parameter; for details see Hunter and Handcock, 2006).

	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
hom practice	0.76	0.198	72	68
hom sex	0.703	0.251	99	93.5
hom office	1.145	0.19	85	80.278
alt. k-tri	0.898	0.148	190.306	177.372

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

The vertex valencies and the calculated influence measures are provided in Table 2. A combination of stochastic approximation (Snijders 2002, 2007) and a Geyer and Thompson (1992) MCMC importance sampling approach in line with Handcock (2003) was used. The accuracy of each set of estimates obtained were checked using independent importance samples, following roughly the practice of applying what in Snijders (2002) algorithm is called the third phase. 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 $150n(n-1).25$).

For this particular example the measures are relatively consistent (although we expect them to differ a lot more in general) and consistently picks out node 15 as the top 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 the different measures of 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).

ID	Deg	GCD MDMLE	GCD ACMLE	DMD	DAC	GCD ACMVP	GCD 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

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

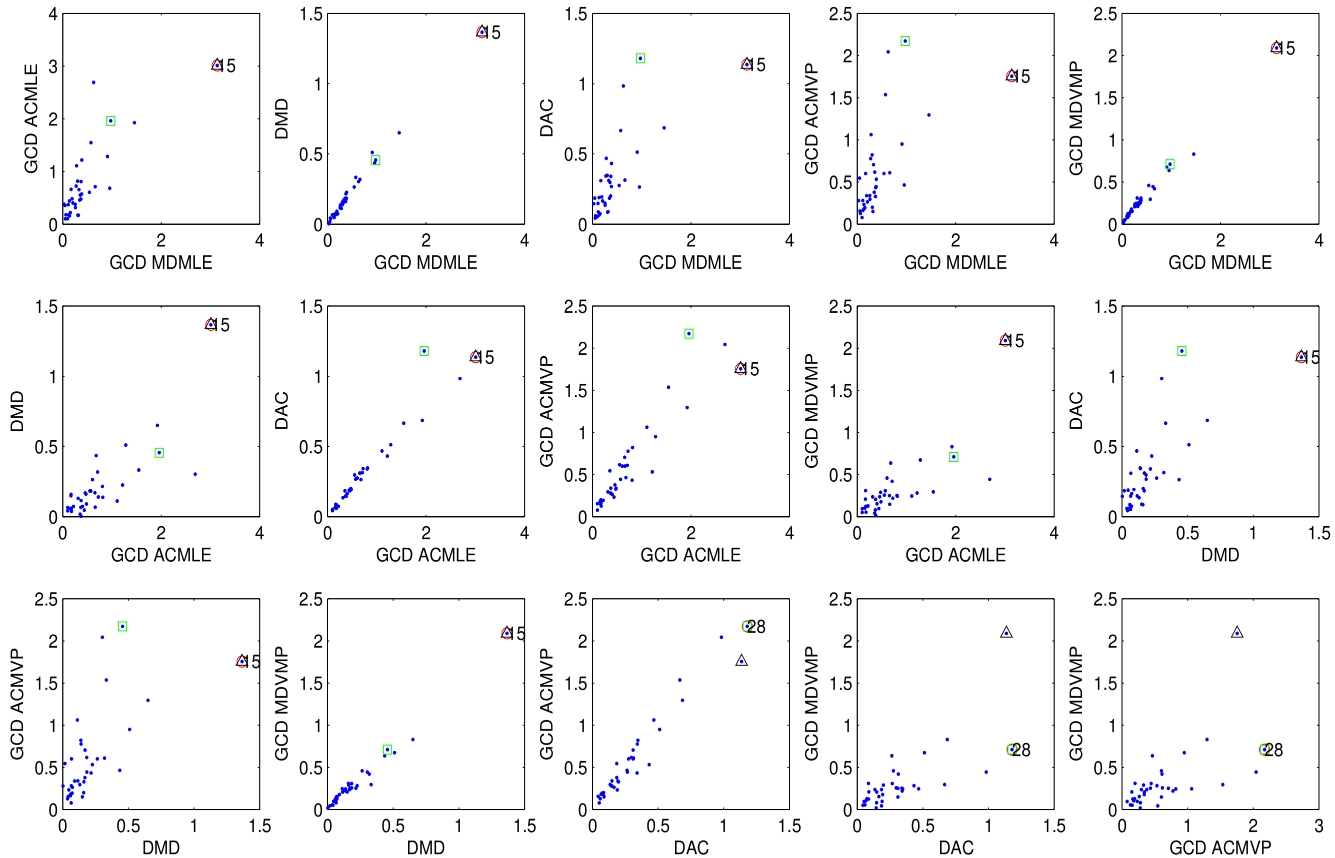

Figure 2: Comparison of influence measures for Lazega's (2001) 36 partners. Actor 15 marked by triangle and actor 28 by square

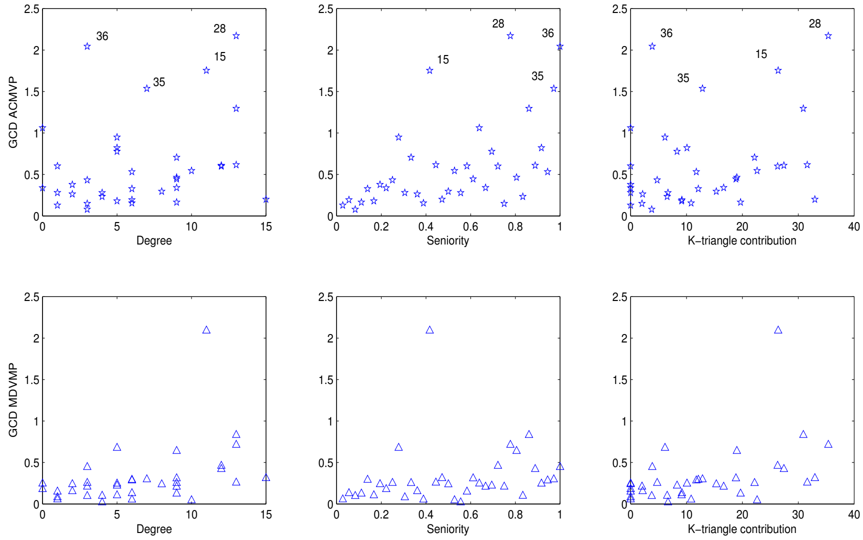

Figure 3: Influence measures against individual attributes (k-triangle contribution given by $z(y_{(i)}; x_{(i)}) - z(y; x)$) for Lazega's (2001) 36 partners with some key actors indicated

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.

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 illustrate that this general idea provides insight into the difference between AC and MD. Secondly, something which is considerably more hard 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. The action of removing an actor i does however induce dependencies among the tie-variables that are not of the type, Markov (and social circuit) dependence, that were assumed for y. 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, 2008a). These matters are highlighted by a closer inspection of the actors 28, 36, qnd 35, that score highly on GCD ACMVP but not on GCD MDMVP.

Judging by Figure 4 actor 15 scores highly 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 ((a) and (g) respectively) in Figure 5 for both MDMLE and ACMLE. As 15 is the only actor in the Providence office (see Figure 1), none of the ties in $y_{15,\bullet}$ contribute towards the homogeneity 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 edges past 15 in AC can be summarised as that while 28 contributes greatly to density and clustering (Figure 3), a big player, 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 triangulate region of the graph and when he is removed using AC many ties are left unexplained and 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 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.

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 on 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.

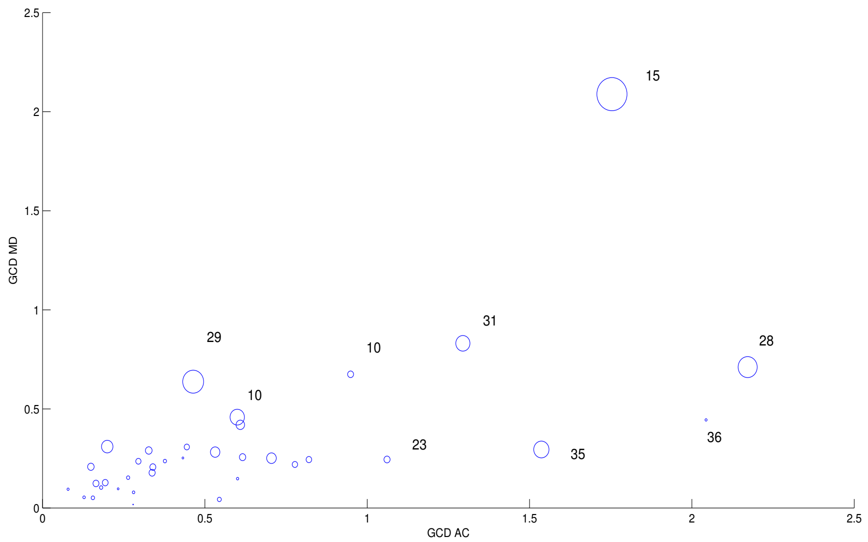

Figure 4: The two approximate generalised Cook's distances for AC and MD, with circle size proportional to Jack-knifed distance, with key actors indicated

When applying these measures it is important to keep in mind that the influence measures only make sense if the model for y is in some sense "true". By true we mean a model that adequately describes the data and is suitably parsimonious. In other words, for ERGMs, studying structural influence is only meaningful if the model has a good fit (Hunter, Goodreau, and Handcock, 2008), not too many parameters and most of them statistically significant. Known computational difficulties of ERGMs (Corander et al., 1998,2002; Snijders, 2002; Handcock, 2003) may also mean that there are some actors for which some of the influence measures are not defined or, at least, very hard numerically to obtain.

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 influen-

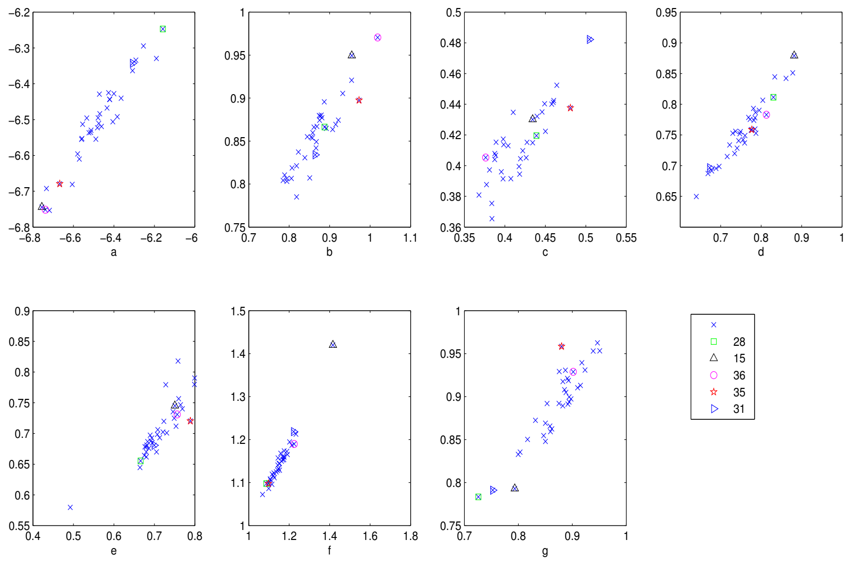

Figure 5: ACMLE (horizontal axes) against MDMLE (vertical) for parameters a through g with key actors indicated

tial actors and outliers in the case of the ERGM and how this relates to the concept of centrality (Freeman, 1979), 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? 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 (cp the closely related so called boundary issue, Laumann et al., 1983; see also discussions in Koskinen, Robins, and Pattison, 2008a,b).

We believe that the proposed influence measures will prove useful tools in further investigating these issues. The properties of these influence measures also warrants further investigation to asses what information they might provide beyond what may be motivated strictly from a statistically principled perspective. Some properties are immediately obvious, such that for example, in the absence of covariates, we have that for structurally equivalent actors i and j, $\hat{\theta}_{(i)} = \hat{\theta}_{(j)}$ and $\widetilde{\theta}_{(i)} = \widetilde{\theta}_{(j)}$. For a Bernoulli model $p_\theta(y) = e^{\theta L}/(1+e^\theta)^{n^*}$, where L is the number of edges and $n^* = \binom{n}{2}$, it is easy to show that GCD ACMLE and GCD MDMLE are identical and equal to $(\bar{L} - x_{i+})^2 n^* L^{-2}$, for $\bar{L} = 2L/n$, 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 most influential.

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.

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.

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.

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

Geyer, C. J., and Thompson, E. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society*, Series B, 54, 657–699.

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. (2007). "Modeling social networks with sampled or missing data," CSSS working paper 75, University of Washington. Obtainable from: http://www.csss.washington.edu/Papers/wp75.pdf.

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. (2007). "Imputation of missing network data: Some simple procedures," *Journal of Social Structure*. Forthcomming.

Hunter, D. R, Goodreau, S. M., and Handcock, M. S., (2008), "Goodness of Fit of Social Network Models," *Journal of the American Statistical Association*, 103, 248–258.

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.

Koskinen, J., Robins, G., Pattison, P. E., (2008a), "Missing data in social networks: Model-based inference," MelNet Social Networks Laboratory Technical Report 08–03, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia. (Available from http://www.sna.unimelb.edu.au/publications/MelNet_Techreport_08_03.pdf)

— (2008b), "Analysing Exponential Random Graph (p-star) Models with Missing Data Using Bayesian Data Augmentation," MelNet Social Networks Laboratory Technical Report 08–04, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia. (Available from http://www.sna.unimelb.edu.au/publications/MelNet_Techreport_08_04.pdf)

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.

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.

Neal, R. M. (2005), “Estimating Ratios of Normalizing Constants Using Linked Importance Sampling,” Technical Report No. 0511, Department of Statistics, University of Toronto. (available from http://arxiv.org/abs/math.ST/0511216)

O’Hara Hines, R. J. and Hines, W. G. S. (1995), “Exploring Cook’s Statistic Graphically,” *The American Statistician*, 49, 389–394.

O’Hara Hines, R. J., Lawless, J. F., and Carter, E. M. (1992), “Diagnos- tics for a Multinomial Generalized Linear Model, with Applications to Grouped Toxicological Mortality Data,” *Journal of the American Statistical Association*, 87, 1059–1069.

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 Morris, M. (2007), “Advances in Exponential Random Graph (p*) Models,” *Social Networks*, 29, 169–172 .

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.

Snijders, T. A. B. (2007), "The Robbins-Monro algorithm for estimation of ERGMs with missing or sampled data," Working paper: Nuffield College, University of Oxford.

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.

van Duijn, M., Gile, K., and Handcock, M., (2007), "Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models," CSSS working paper 74, University of Washington. Obtainable from: http://www.csss.washington.edu/Papers/wp74.pdf.

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.

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.