HHS Public Access

Author manuscript

J Comput Graph Stat. Author manuscript; available in PMC 2016 June 01.

Published in final edited form as:

J Comput Graph Stat. 2015 June 1; 24(2): 520–538. doi:10.1080/10618600.2014.923777.

Role Analysis in Networks using Mixtures of Exponential Random Graph Models

Michael Salter-Townshend[*] and
Dept. of Statistics, University of Oxford

Thomas Brendan Murphy
School of Mathematical Sciences, University College Dublin

Abstract

A novel and flexible framework for investigating the roles of actors within a network is introduced. Particular interest is in roles as defined by local network connectivity patterns, identified using the ego-networks extracted from the network. A mixture of Exponential-family Random Graph Models is developed for these ego-networks in order to cluster the nodes into roles. We refer to this model as the ego-ERGM. An Expectation-Maximization algorithm is developed to infer the unobserved cluster assignments and to estimate the mixture model parameters using a maximum pseudo-likelihood approximation. The flexibility and utility of the method are demonstrated on examples of simulated and real networks.

Keywords

Exponential Random Graph Model; ego-network; finite mixture model; Expectation Maximisation algorithm

1 Introduction

Most current work on clustering nodes in a network focusses on community finding; that is, finding sets of nodes that are highly connected to each other. For example, in latent space models (e.g. Handcock et al. (2007)), pairs of nodes that are clustered together are more likely to link to each other than nodes in separate clusters. In stochastic blockmodels (e.g. Snijders and Nowicki (1997)) the probability of linking to nodes within a block (or cluster) is often higher than the between block connection probability. It is possible that the between block linkage probabilities may be higher than within block linkage probabilities (Nowicki and Snijders (2001)), but clustering is still performed based on the ratio of between to within linkage densities.

In contrast to the above, the clustering nodes by similarity of *role* is a problem that has been the subject of comparatively little research. Our goal here is to develop a statistical model-based method of clustering nodes based on the role they play in the network. Clustering based on degree or other centrality measure is the simplest example of this type of

[*]salter@stats.ox.ac.uk.

clustering. Core/periphery models (Borgatti and Everett (1999)) go further and partition nodes into a core of nodes that link to each other and also to a periphery of nodes which are in turn connected only to the core. The authors acknowledge that the concept of the core/periphery structure is closely related to degree (and other measures of centrality), although they also note that "all coreness measures are centrality measures, but the converse is not necessarily true". For example, the subgraph of the most central or high degree nodes may not contain any links and would thus constitute an empty core. We are motivated by problems involving roles that may be more complex than those based only on degree, centrality or coreness.

Lerner (2005) presents a comprehensive review of methodologies for assigning roles to nodes. These include structural equivalence, Regular equivalence and other related equivalences. Structural equivalence assigns the same role to two nodes (egos) if and only if they share the same neighbours (alters). Regular Equivalence is closer in spirit to our method. Everett and Borgatti (1994) present Regular Equivalence as a method of graph colouration in which the colours (or roles) are assigned such that the spectrum of colours of the alters is the same for any two egos of the same colour. Everett and Borgatti (1994) states that "A colouration *C* of a digraph *D* is an assignment of colours to the vertices of *D*. Any colouration induces a partition of the vertices which defines an equivalence relation".

For example, if a yellow node has connections to blue and red nodes only, then so must all yellow nodes. i.e. all nodes coloured yellow must have at least one red and one blue neighbour and no links to any non blue or red nodes. The number of connections and the identities of the alters do not need to be the same and any given graph or network may have several such valid colourations. The key difference between our methodology and approaches such as Regular Equivalence is that these methods do not discriminate on different *patterns* of connectivity. In addition, Regular Equivalence scales as $\mathcal{O}(n^3)$.

There is some literature on identifying roles in a more local manner but this is currently restricted to visualisation based methods of discriminating roles and/or methods based on in-degree and out-degree. For example, Fisher (2005) looks at distributions of in-degree and out-degree within ego-networks (see Section 2) to identify "answer people" in newsgroups. Similarly, Fisher et al. (2006) look at role assignment based on in-degree and out-degree and perform visualisations of ego-networks only to refine understanding of these roles.

Welser et al. (2007) develop the idea yet further and refer to "patterned characteristics of communication" which they conceive of as "structural signatures". Again, they restrict their focus to identification of "answer people" for whom the pattern of connectivity is assumed known *a priori*. This can be thought of as clustering people into two groups: the answer people and the rest. Welser et al. (2007) perform formal regression analysis on the role assignments. Specifically, a linear model is postulated that incorporates a measure of the pattern of connectivity in a node's ego-network. However, this is reduced to a scalar measure of whether the node was above average on three characteristics (namely the proportion of out-degree ties to low-degree neighbours, triangles and intense ties). Thus the modelling is restricted to searching for a very specific pattern of connectivity, as acknowledged by the authors when they say "we are less concerned with assigning all actors

to different classes than we are in identifying general structural features that are associated with one particular role".

Gleave et al. (2009) go the furthest in formally identifying what a social role means in terms of a node in a network. Their definition is both "conceptual and operational" and is a "combination of social psychological, social structural, and behavioral attributes." Our interest lies in the social structural part of their definition. Welser et al. (2011) seeks to identify roles of Wikipedia editors based on patterns in their ego-network. Again, analysis of the motif patterns of the ego-networks begins with "broad qualitative explorations" and acknowledge that "researchers could construct wiki related metrics that would help distinguish between different role types" but that "constructing and testing those metrics is beyond the scope of this paper". A novel and flexible framework in which to do precisely that, for a broad range of network types, is the challenge that we undertake in this paper. Furthermore, our interest is in problems for which we do not have *a priori* knowledge of what roles the nodes in the network might play.

Brandes and Lerner (2007) provides a survey of methods for determining role-equivalency in networks. They note that neighbourhood identity and overlap for role analysis "are biased to detecting dense subgraphs and thus do not distinguish conceptually from density based graph clustering". They therefore introduce *structural similarity* as a relaxation of regular equivalency. To this end the describe how to construct characteristic matrices from which they obtain groupings of nodes such that "similar vertices have to be connected to vertices that are themselves similar". A characteristic matrix is called a similarity if it is symmetric and idempotent (i.e. $S^2 = S$). They demonstrate several interesting properties of the approach including the fact that automorphic equivalences induce structural similarities.

Brandes and Lerner (2010) extends the discussion and introduces efficient algorithms for structural similarity computation, based on eigenvector decomposition of the adjacency matrix. In this paper, structural similarity is framed as a relaxation of blockmodelling. They compare their approach with other spectral methods and argue that the difference between the two is the same as the difference between requirements of identical versus equivalent neighbourhoods for equivalent vertices. They note that "which subset actually has to be chosen depends on what the similarity is for, what properties it should satisfy, or what criterion should be optimized" but propose generalised algorithms for spectral 2 and 3 colouring. In our results sections we will compare with results obtained using the spectral 2 and 3 colouring algorithms of Brandes and Lerner (2010).

We approach the problem of role analysis using ego-networks and within the Exponential (family) Random Graph Model (ERGM) framework (see Robins et al., 2007, for example). ERGMs provide a probability model for a network through a set of network sufficient statistics. We model the set of all ego-networks using a finite mixture of ERGMs in order to perform a model-based clustering of the ego-networks and thus cluster the nodes of the network. We refer to this model as the **ego-ERGM**.

An Expectation-Maximization algorithm is developed to fit the model and thus to simultaneously estimate both the node clustering assignments and the ERGM parameters the

clusters. The development of this mixture model for role analysis is detailed in Section 4. Note that our model and algorithm is loosely connected to, but very different from Steinley et al. (2011), where they use clustering and ERGMs for network data. They do not focus on ego-networks or on nodal roles but perform *k*-means clustering using ERGMs fitted to subsets of a network with a view to partitioning the network into parts that each adhere to different ERGMs.

2 Ego-Networks

Following Harrigan et al. (2012), we focus our role analysis on the ego-networks for each node in the network. The ego-network of a particular node reveals the local structure of the network around that node. Each node gives rise to its own ego-network as follows:

1. Select a node; this is referred to as the ego.
2. Include the alters (nodes connected to the ego).
3. Include the connections between the alters.

It is also possible to extend the ego-network by looking at the alters for each alter, etc. The central hypothesis of this paper is that nodes performing differing local roles will have markedly different ego-networks and nodes performing similar roles will have similar ego-networks, possibly beyond just the size of ego-network which is the degree of the ego. This is due to varying behaviour of the surrounding alters. An illustration of this point is made in Figure 1.

3 Exponential Random Graph Models

Exponential random graph models (ERGMs or p^*) (Holland and Leinhardt (1981)) are a flexible and popular model for statistical network analysis. ERGMs are an exponential family model which uses sufficient network statistics to model the whole network. See Robins et al. (2007) for an introduction and Robins et al. (2007) for more recent developments. Additionally, Goldenberg et al. (2010) and Salter-Townshend et al. (2012) provide general reviews which compare various models for network analysis including ERGMs.

ERGMs do not assume dyadic independence and model the whole network as a single realisation arising from a distribution parametrised by a collection of network statistics, as outlined below. Specifically, the probability of the observed network $\boldsymbol{Y}$ is proportional to the exponent of the sum of the network statistics times some unknown parameters;

$$P(\boldsymbol{Y}|\underline{\theta}) = \exp\{\underline{\theta}^T S(\boldsymbol{Y}) - \gamma(\underline{\theta})\}, \quad (1)$$

where $\underline{\theta}$ are the parameters of the model, $S(\boldsymbol{Y})$ are network summary statistics chosen by the analyst and $\gamma(\underline{\theta})$ is a normalising constant. This normalising constant is difficult to obtain as it involves summing over all possible networks.

Typical choices for $S(\boldsymbol{Y})$ include the number of edges, the number of triangles and the number of k-stars for various k values. Fitting the ERGM then involves finding estimates of the parameters for each of the network statistic terms in the model. There are several approaches to fitting ERGMs without recourse to summing over all possible networks; these include maximum pseudo-likelihood estimation (MPLE) (Besag (1975), Strauss and Ikeda (1990)), Markov Chain MLE (Geyer and Thompson (1992); Snijders (2002)) and Bayesian methods (Caimo and Friel (2011)).

The estimated parameter values for ERGMs are heavily dependent on network size. To help account for this effect, Krivitsky et al. (2011) propose the addition of an offset term to adjust for network size in ERGMs and we include this term in the work presented here. Although Krivitsky et al. (2011) also look at fitting ERGMs to ego-networks, they do not do this in the context of role analysis; their interest is in comparison of maximum likelihood estimates across varying sizes of ego-network.

4 Finite Mixtures of ERGMs

We fit a finite mixture model to the extracted ego-networks, where the component distribution within each cluster is an ERGM with cluster specific parameters. Note that, although there is overlap between the ego-networks extracted from the network, we model them as independent. A simplistic two-stage approach would be to fit an ERGM to each ego-network in turn and then apply a post-hoc clustering algorithm (e.g. k-means clustering) to the coefficients of the fitted ERGMs. We adopt a more formal model based clustering approach based on a finite mixture of ERGMs.

Suppose that there are G clusters (or role exhibited by the nodes). The *a priori* probability of a node taking role g is τ_g for $g = 1, 2, \ldots, G$. In addition, suppose that nodes within role g are modeled by an ERGM (1) with role specific parameter vector $\underline{\theta}_g$. The probability of ego-network $\boldsymbol{Y}_i$ is given by the mixture model

$$P(\boldsymbol{Y}_i|\underline{\tau},\boldsymbol{\theta})=\sum_{g=1}^{G}\tau_g \exp\{\underline{\theta}_g^T S(\boldsymbol{Y}_i)-\gamma(\underline{\theta}_g)\},$$

where $\underline{\tau} = (\tau_1, \tau_2, \ldots, \tau_G)$ and $\boldsymbol{\theta} = (\underline{\theta}_1, \underline{\theta}_2, \ldots, \underline{\theta}_G)$ are the model parameters. It is a feature of the ERGM family of model that the normalising constant $\gamma(\underline{\theta}_g)$ involves a sum over all possible networks and is computationally intractable as it is $\mathcal{O}(2^{n^2})$ for a network with n nodes (note that we will apply ERGMs to the ego-networks so the n here will pertain to ego-network sizes). There are two competing approximations to this likelihood that avoid such computation; the Maximum Pseudo-likelihood Estimation (MPLE) method and the Markov Chain Maximum Likelihood Estimation method. We discuss both methods in the context of inference for our model in Section 5 but for now suppress reference and refer to $P(\boldsymbol{Y}_i|\underline{\tau}, \boldsymbol{\theta})$.

If we assume independence of the ego-networks, we find that the likelihood is the form

$$P(\boldsymbol{Y}|\underline{\tau},\boldsymbol{\theta})=\prod_{i=1}^{N}\left[\sum_{g=1}^{G}\tau_g \exp\{\underline{\theta}_g^T S(\boldsymbol{Y}_i)-\gamma(\underline{\theta}_g)\}\right]. \quad (2)$$

We acknowledge that the ego-networks overlap and are therefore not in fact independent, however we use (2) as an approximation in the spirit of a pseudo-likelihood. The effect of this non-independence of ego-networks is that the overlap of adjacent ego-networks will cause our model to cluster nodes that are "close" in the network to a greater extent. We note that this is also a feature of neighbourhood overlap methods; as Brandes and Lerner (2007) mention "these measures are biased to detecting dense subgraphs and do not distinguish conceptually from density based graph clustering".

We will use an Expectation-Maximization (EM) algorithm (Dempster et al. (1977)) to find maximum likelihood estimates of the model parameters $\underline{\theta}_g$. We introduce unobserved indicator variables $\underline{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iG})$ where $Z_{ig} = 1$ if node i belongs to role g and zero otherwise. Treating these indicators as missing values leads to a complete-data likelihood given by

$$P(\boldsymbol{Y},\boldsymbol{Z}|\underline{\tau},\boldsymbol{\theta})=\prod_{i=1}^{N}\prod_{g=1}^{G}[\tau_g \exp\{\underline{\theta}_g^T S(\boldsymbol{Y}_i)-\gamma(\underline{\theta}_g)\}]^{Z_{ig}},$$

where $\boldsymbol{Z} = (\underline{Z}_1, \underline{Z}_2, \ldots, \underline{Z}_N)$. The complete-data log-likelihood is therefore given by

$$\begin{aligned}\log[P(\boldsymbol{Y},\boldsymbol{Z}|\underline{\tau},\boldsymbol{\theta})] &= \sum_{i=1}^{N}\sum_{g=1}^{G} Z_{ig}\log[\tau_g \exp\{\underline{\theta}_g^T S(\boldsymbol{Y}_i)-\gamma(\underline{\theta}_g)\}] \\ &= \sum_{i=1}^{N}\sum_{g=1}^{G} Z_{ig}\{\log\tau_g+\underline{\theta}_g^T S(\boldsymbol{Y}_i)-\gamma(\underline{\theta}_g)\}.\end{aligned} \quad (3)$$

At convergence the EM algorithm provides maximum likelihood estimates of the mixing proportions $\underline{\hat{\tau}}$, the ERGM parameters $\hat{\boldsymbol{\theta}}$ along with estimates of the expected cluster memberships $\hat{\boldsymbol{Z}}$ and the value of the maximized likelihood.

4.1 Selection of ERGM Terms

Selection of terms for ERGM models is a problem that has received surprisingly little attention. Morris et al. (2008) simply state that "the larger question of how to go about choosing terms wisely is beyond its scope." Hunter et al. (2012) provide a recent review of the ERGM literature (along with several other network models) and discuss model degeneracy, but do not touch on selection of ERGM terms. Model degeneracy refers to the situation in which only a few networks (often the full or empty networks) have appreciable probability given the model and is a challenge in fitting ERGM models. Robins et al. (2007) introduce ERGM terms designed specifically to address model degeneracy but again do not discuss model selection. Hunter et al. (2008) propose graphical methods for model selection in the ERGM framework as "traditional methods (AIC, BIC, etc) entail several problems. For one thing, the assumptions used to justify the AIC and BIC are not met here, because

our observations are not an independent and identically distributed sample. In fact, it is not even clear how to evaluate BIC, because there is no easy way to determine the effective sample size N". Caimo and Friel (2013) address model selection in a Bayesian framework but rather than exploring a space of models defined on sets of ERGM terms, they compare a small number of preselected competing models.

We have experimented with a version of the headlong greedy model selection algorithm based on BIC (to select both terms and number of groups) developed in Murphy et al. (2010) for our model. For the simulation study in Appendix A this algorithm did not perform well. We simulated a set of small ERGMs using the terms edges, mutual, geometrically weighted in-degree distribution and geometrically weighted out-degree distribution. Given a superset of summary statistics including those used to simulate data, the algorithm converged to the terms edges, geometrically weighted edgewise shared partner distribution, geometrically weighted in-degree distribution, geometrically weighted out-degree distribution, triangle, mixed 2-stars and geometrically weighted dyadwise shared partner distribution. Thus the term mutual is erroneously dropped and the terms geometrically weighted edgewise shared partner distribution, mixed 2-stars, triangle and geometrically weighted dyadwise shared partner distribution are erroneously included. Multiple re-runs with freshly simulated data lead to other models being selected, indicating that model choice for ERGMs is a difficult problem. The large overlap between various ERGM statistics on the same network leads to a high collinearity of terms and thus competing models will have similar likelihoods. However, we find that across multiple simulations the clustering performance that is our focus was not significantly effected.

4.2 Two-Stage Method for Initialisation

To initialise the algorithm we first estimate ERGM parameters for each ego-network i independently using maximum pseudo-likelihood (MPLE) which is implemented in the R (R Development Core Team (2010)) package ergm (Handcock et al. (2011)). We then cluster these parameter values using a k-means algorithm. This provides initial estimates of $\boldsymbol{Z}$, the cluster membership vectors. We also set the mixing proportions $\underline{\tau}$ as the group proportions in the k-means clustering and the role parameters $\underline{\theta}_g$ equal to the cluster mean of the individual ERGM parameters; these are denoted as $\hat{\boldsymbol{Z}}^{(0)}$, $\underline{\tau}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ respectively, for the first iteration of an EM algorithm. We will also refer to this initialisation as the **two-stage method** and we compare accuracy of classification across multiple simulation studies in Appendix A. We note that some ego-networks will be poorly specified by our choice of ERGM for the mixture model. In particular, some ego-networks will have undefined or smallest attainable valued counts for some statistics. In this case, our two-stage algorithm is forced to set the coefficients to zero or a large negative number which is somewhat ad-hoc; however the EM fitted mixture model is still valid.

4.3 EM Algorithm

The EM algorithm then uses the following update steps:

0 Let $t = 0$ and let $\underline{\tau}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ be the initial estimates of the model parameters.

1 **E-Step:** Compute the expected values of the Z_{ig} values based on the current model parameters.

$$\hat{Z}_{ig}^{(t+1)} = \frac{\hat{\tau}_g^{(t)} P(\boldsymbol{Y}_i | \underline{\hat{\theta}}_g^{(t)})}{\sum_j \hat{\tau}_j^{(t)} P(\boldsymbol{Y}_i | \underline{\hat{\theta}}_j^{(t)})}.$$

2 **M-Step:** Maximize the expected complete-data log-likelihood to yield new parameter estimates. That is,

$$(\underline{\hat{\tau}}^{(t+1)}, \hat{\boldsymbol{\theta}}^{(t+1)}) = \underset{\underline{\tau}, \boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \sum_{g=1}^{G} \hat{Z}_{ig}^{(t+1)} \log \{\tau_g P(\boldsymbol{Y}_i | \underline{\theta}_g)\}.$$

3 Check for convergence. If converged, stop. Otherwise, increment *t* and return to Step 1.

We use a maximum change in $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{\theta}}$ of less than 10^{-6} to check for convergence.

The M-step of the EM algorithm is computationally expensive because the $P(\boldsymbol{Y}_i|\underline{\theta}_g)$ term is an ERGM which is in itself difficult to fit. However, the expected complete-data log-likelihood is of the form

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{g=1}^{G} \hat{Z}_{ig}^{(t+1)} \log \left\{\hat{\tau}_g P(\boldsymbol{Y}_i|\underline{\hat{\theta}}_g)\right\} &= \sum_{i=1}^{N}\sum_{g=1}^{G} \hat{Z}_{ig}^{(t+1)} \left\{\log \hat{\tau}_g + \underline{\hat{\theta}}_g^T S(\boldsymbol{Y}_i) - \gamma(\underline{\hat{\theta}}_g)\right\} \\
&= \sum_{i=1}^{N}\sum_{g=1}^{G} \hat{Z}_{ig}^{(t+1)} \log \hat{\tau}_g + \sum_{i=1}^{N}\sum_{g=1}^{G} \hat{Z}_{ig}^{(t+1)} \left\{\underline{\hat{\theta}}_g^T S(\boldsymbol{Y}_i) - \gamma(\underline{\hat{\theta}}_g)\right\} \\
&= \sum_{g=1}^{G} \hat{N}_g^{(t+1)} \log \hat{\tau}_g + \sum_{g=1}^{G} \left\{\underline{\hat{\theta}}_g^T \left[\sum_{i=1}^{N} \hat{Z}_{ig}^{(t+1)} S(\boldsymbol{Y}_i)\right] - \hat{N}_g^{(t+1)} \gamma(\underline{\hat{\theta}}_g)\right\} \\
&= \sum_{g=1}^{G} \hat{N}_g^{(t+1)} \log \hat{\tau}_g + \sum_{g=1}^{G} \hat{N}_g^{(t+1)} \left\{\underline{\hat{\theta}}_g^T \hat{S}_g^{(t+1)} - \gamma(\underline{\hat{\theta}}_g)\right\},
\end{aligned}$$

where $\hat{N}_g^{(t+1)} = \sum_{i=1}^{N} Z_{ig}^{(t+1)}$ is the number of observations assigned to role *g* and $\hat{S}_g = \sum_{i=1}^{N} Z_{ig} S(\boldsymbol{Y}_i) / \hat{N}_g^{(t+1)}$ is a weighted average of the sufficient network statistics. The first term of the expression can be easily maximized in closed-form to yield $\hat{\tau}_g^{(t+1)} = \hat{N}_g^{(t+1)} / N$ and the second term has the same functional form as the log-likelihood of a single ERGM. Therefore, the methods for finding estimates of ERGM parameters can be used in the M-step of the algorithm.

The next section details some issues regarding the computational complexity of our method, including the need to resort to the maximum pseudo-likelihood estimator, we note that the algorithm scales as $\mathcal{O}(n \times 2^{n_e^2})$ where *n* is the number of nodes in the network and n_e is the average number of nodes in each ego-network. This is because there are *n* such egos and all ERGM calculations are performed on the extracted ego-networks only. Thus the method

scales linearly with the number of nodes in the network rather than quadratically or even cubically as does Regular Equivalence.

5 Inference

There is a consensus in the literature on ERGMs that estimation should be approached using Monte-Carlo Maximum Likelihood Estimation (MCMLE, Snijders (2002) and Hunter and Handcock (2006)). This involves sampling many networks from a model with a current best estimate of the ERGM parameters in order to approximate the normalising constant $\gamma(\underline{\theta}_g)$.

An algorithm to find the MCMLE would consist of the following steps: Initialise using the MPLE fit. Until convergence, iterate the E and M steps as per Section 4. For the E-step we require the log-likelihood for network *i* given current $\underline{\theta}_g$. To do that we can use the so called bridge log-likelihood difference estimator:

$$\hat{L}(\underline{\hat{\theta}}_g) = r(\underline{\hat{\theta}}_g, \underline{\theta}_g^*) - r(0, \underline{\theta}_g^*) - \log(M),$$

where

$$r(\underline{\theta}_a, \underline{\theta}_b) = (\underline{\theta}_a - \underline{\theta}_b)^T S(\boldsymbol{Y}_i) - \log\left(\frac{1}{M}\sum_{m=1}^{M} \exp\left((\underline{\theta}_a - \underline{\theta}_b)^T S(\boldsymbol{Y}_i^{(m)})\right)\right),$$

where the $\boldsymbol{Y}_i^{(m)}$ are networks simulated using parameters $\underline{\theta}_b$. We can use the current guess for $\underline{\theta}_g$ for both $\underline{\theta}_g^*$ in *r* so that:

$$\begin{aligned} P(\boldsymbol{Y}_i | \underline{\hat{\theta}}_g^{(t)}) &\simeq \hat{L}(\underline{\hat{\theta}}_g) \\ &= r(0, \underline{\hat{\theta}}_g) - \log(M) \\ &= -\underline{\hat{\theta}}_g S(\boldsymbol{Y}_i) - \log\left(\frac{1}{M}\sum_{m=1}^{M} \exp\left(-\underline{\hat{\theta}}_g S(\boldsymbol{Y}_i^{(m)})\right)\right). \end{aligned}$$

Thus we must simulate *M* networks for each combination of *g* and *i*. We may be able to reduce the computational overhead somewhat by simulating for each unique combination of **size** and **density** of ego-network as the offset term will be the same for two networks with the same numbers of both nodes and edges. We can also re-use the simulated networks required by the M-step.

For the M-step we must maximize the complete data log-likelihood w.r.t. $\underline{\theta}_g$ (actually, maximising $r(\underline{\theta}_g, r(\underline{\hat{\theta}}_g, \underline{\theta}_g^*)$ will do the job). The issue here is that we want to maximise w.r.t. $\underline{\theta}_g$, therefore we will need to simulate many new networks for each value of $\underline{\theta}_g$ that we examine. Thus we must simulate $M \times N \times G \times A$ networks for *each* iteration *t* of our EM algorithm where *A* is the number of iterations to find the current MCMLE $\underline{\theta}_g$. We are maximising:

$$r(\hat{\underline{\theta}}_g^{(a+1)}, \hat{\underline{\theta}}_g^{(a)}) = (\hat{\underline{\theta}}_g^{(a+1)} - \hat{\underline{\theta}}_g^{(a)}) \times \sum_{i=1}^{N} \left(\frac{\hat{Z}_{ig} S(\boldsymbol{Y}_i)}{\hat{N}_g^{(t+1)}} - \log \left(\frac{1}{M} \sum_{m=1}^{M} (\hat{\underline{\theta}}_g^{(a+1)} - \hat{\underline{\theta}}_g^{(a)}) S(\boldsymbol{Y}_i^{(m)}) \right) \right).$$

where $\boldsymbol{Y}_i^{(m)}$ is simulated networks with size and density corresponding to network *i* using parameters $\hat{\underline{\theta}}_g^{(a)}$. *a* iterates from 1 to *A* (or until convergence) given the current estimate of $\boldsymbol{Z}$. This procedure is computationally intractable and so we resort to Maximum Pseudo-Likelihood Estimation (MPLE) for both the E and the M steps. It is crucial to note that although this procedure has been shown to result in poor estimates (van Duijn et al. (2009)), our focus here is not in finding ERGM parameter estimates that are “correct”, but rather in finding a clustering of ERGMs that is useful.

The MPLE method examines change statistics calculated by taking the difference in the observed summary statistics when switching each possible link from on to off. The pseudo-likelihood is the product of the full conditionals for each dyad given the rest of the observed network and inference proceeds in a manner similar to logistic regression (see Strauss and Ikeda (1990); Robins et al. (2007)). A large amount of the computation required for fitting ERGMs lies with the calculation of the network summary statistics. Unlike the MCMLE approach, which would require recalculation of such statistics for the *M* simulated networks required at each iteration *t* of the EM algorithm for each value of both *g* and *i* and for each iteration *a* of the MCMLE algorithm, the MPLE approach only requires a single one-off calculation of the change statistics done at the beginning of the algorithm.

It is worth noting two things here; firstly, when we ran non-mixture based independent ERGM fitting to our examples we did not observe a marked difference between the MPLE and MCMLE results (see Appendix A). Secondly, there are examples in the mixture modelling literature of approximate methods that perform well in terms of clustering despite yielding poor estimates of the model parameters. One such example is in k-means clustering; as Friedman (1997) notes:

> Whereas the squared estimation error degrades by over a factor of 35 as the number of irrelevant inputs is increased by a factor of 20, the corresponding increase in classification error is less than a factor of six.

Similarly, our focus is on a computationally convenient algorithm that has enough power to infer a useful clustering of the ego-networks. As MCMLE is impractical, we resort to MPLE and we demonstrate that it can indeed provide usable clusterings of the ego-networks/nodes.

We use BIC to select the the number of clusters *G*, with BIC defined as

$$2\hat{\ell} - (GM + G - 1)\log(N).$$

where $\hat{\ell}$ is the value of the maximised pseudo-log-likelihood, *G* is the number of groups/roles and *M* is the number of summary statistics in the model. The use of pseudo-likelihood

in combination with BIC has previously been exploited by Stanford and Raftery (2002) where they developed a criterion called the Pseudo-Likelihood Information Criterion (PLIC).

6 Results

We present results for two real world datasets for which annotated underlying roles are of interest and are available. A simulation study is presented in Appendix A to assess the quality of the mixture estimates and the clusterings. In both cases we use the following four commonly used ERGM terms: edges, mutual, geometrically weighted in-degree distribution and geometrically weighted out-degree distribution. We set the decay rate of the degree based statistics as 0.8. We present results based on the two-stage method and the fitted ego-ERGM model along with Regular Equivalence as well as spectral 2-colouring and 3-colouring algorithms based on the structural similarity matrices of Brandes and Lerner (2010). We acknowledge that more general structural similarity approaches involve the selection of a subset of eigenvectors but restrict our analysis here to spectral 2-colouring and 3-colouring. These methods are linear in the number of nodes *plus* the number of edges and thus scale similarly to our method.

6.1 Lazega's Lawyers

We now apply our method to the familiar Lazega's Lawyers dataset (Lazega (2001)). This network consists of 71 lawyers in a firm in New England. We examine the advice network in which the lawyers named other members of the firm that they asked for work advice. This directed network has 892 links and there is covariate information on the lawyers. We are interested in the status covariate as this may be an indicator of role; all members are classed as either a partner or an associate and the split is approximately equal (36 partners and 35 associates).

BIC selects the 3 group model and Fisher's Exact test for contingency between the status of the lawyers and the maximal clustering assignments yielded a p-value of 2.88×10^{-6}. Table 1 shows the maximal breakdown into the 3 groups, the 2 groups when $G = 2$ is modelled and a breakdown by Freeman degree only. Assuming groups of approximately equal size, we group the nodes into low and high as defined by below and above median degree. The p-value associated with comparing the 2 groups to the observed status of the lawyers was 1.25×10^{-6} and for the degree only method it is 0.0042.

Our EM algorithm found estimates of $\hat{\tau}$ and $\hat{\boldsymbol{\theta}}$ for the 4 group model equal to:

$$\hat{\underline{\tau}} = \begin{bmatrix} 0.14 \\ 0.36 \\ 0.38 \\ 0.12 \end{bmatrix} \quad \hat{\boldsymbol{\theta}} = \begin{bmatrix} 2.87 & 1.02 & -5.70 & -5.37 \\ 2.65 & 1.42 & -5.09 & -4.66 \\ 3.11 & 0.38 & -4.62 & -4.08 \\ 3.38 & -0.40 & -1.80 & -3.21 \end{bmatrix}.$$

Our method does not simply return a hard classification but assigns a probability for each node of belonging to each group. We observe that this soft clustering is useful as those actors that are assigned to the same (most overlapping) groups as recorded legal status have

a higher average λ than those that are mis-classified. For example, in the 3-group model of Table 1 the 20 associates that are maximally assigned to group 1 have a mean $\hat{Z} = 0.970$ whereas the 7 associates that are maximally assigned to group 2 have a mean $\hat{Z} = 0.911$. Similarly, the 28 partners that are maximally assigned to group 2 have a mean $\hat{Z} = 0.967$ whereas the 6 parters maximally assigned to group 1 have a mean $\hat{Z} = 0.906$.

Similar results hold for the two-group fit: The 28 partners that are maximally assigned to group 1 have a mean $\hat{Z} = 0.989$ whereas the 8 partners that are maximally assigned to group 2 have a mean $\hat{Z} = 0.937$. Similarly, the 28 associates that are maximally assigned to group 2 have a mean $\hat{Z} = 0.973$ whereas the 7 associates maximally assigned to group 1 have a mean $\hat{Z} = 0.918$. Thus a less certain soft clustering is correlated with potential mis-classification.

6.2 Prosper Microfinance

We now apply our method to a larger, real world dataset. Prosper.com is a microfinance website that allows users to request and provide small, short terms loans. Borrowers list loans they would like to get and provide personal details along with details pertaining to the loan, such as purpose and duration; this is a listing. Potential lenders then bid to contribute fractions of these loans; these are bids. If enough bids are made to contribute the entire amount of the loan then the loan is made. Users may be both borrowers and lenders on different loans. Groups also exist based on religion, interests, geographical area, etc, with loans within groups more common.

All data on the site is publicly available for download. We downloaded the entire dataset since the sites inception in 2005. We can construct many social networks based on various connection types. For the analysis we looked at all completed loans made in 2010. This is a directed network with 2649 nodes and 4619 edges. We fit a mixture model with 5 groups, as selected by BIC. We also have the recorded roles that the actors play; however, most users have several labelled roles.

We find that Regular Equivalence, spectral 2-colouring and 3-colouring, the two-stage approach and the ego-ERGM mixture model all succeed in clustering the actors such that there is a very strong contingency between the assigned roles and the reported roles (all methods have a χ^2 test for contingency p-value less than 2.2×10^{-16}). Table 3 gives the Lambda measure of directed association (Goodman and Kruskal (1954)). This is a measure of proportional reduction in error in cross tabulation analysis that varies from 0 (no association) to 1 (perfect association). This measure was chosen as we are interested in predictive performance for the underlying latent roles.

We again examine the usefulness of the soft clustering capability of our method. We proceed assuming that the reported roles are what we are interested in recovering using our model. We further assume that for each role type, the group with the highest rate of maximally assigned actors to that role is the correct classification. This assumption is based on the result that there is a strong, statistically significant association between the reported roles and the maximal clusterings. We then measure for each reported role the average $\hat{Z}$ values for the correctly and incorrectly classified actors that belong to the next largest

maximal group for that role. We find that the mean $\hat{Z}$ for correctly classified actors is 0.879 whereas it is just 0.727 for each next largest group of mis-classified actors. Therefore the soft clustering is useful for predicting which actors have been mis-classified in the absence of a ground truth reference.

7 Conclusion

We have presented a novel model and method for clustering the nodes in a network by role, as defined by patterns of local connectivity which we name the ego-ERGM. We develop and fit a mixture of Exponential Random Graph Models to the extracted ego-networks of the network such that nodes with similar local patterns of connectivity are clustered together. We employ an Expectation-Maximization algorithm that makes use of a pseudo-likelihood approximation to estimate the clustering assignment probabilities and the maximum-likelihood estimates of the cluster-specific ERGM parameters.

Our method compares favourably with existing methods for the analysis of unknown a-priori roles. Importantly, it scales as $\mathcal{O}(n \times 2^{n_e^2})$ where n is the number of nodes in the network and n_e is the average number of nodes in each ego-network. This is because there are n such egos and all ERGM calculations are performed on the extracted ego-networks only.

We show via a simulation study (see Appendix A) that the estimates for the ERGM parameters are poor due to the computationally imposed use of the MPLE approximation and the noisiness of the ERGM framework in general. However, we have also demonstrated that the ability of our approach to correctly cluster ERGM networks under a mixture model is excellent despite this. The method is unable to accurately determine the correct number of underlying clusters using BIC, favouring too many clusters. However, when too many clusters are modelled the true clusters are split and not merged together in the modelled clusters.

A debate we do not enter here is whether ERGMs in general should be considered as a generative model (see Blei et al. (2007)). Certainly, random networks may be simulated from an ERGM given a set of summary statistics and values for the parameters/coefficients. However, we note that our mixture of ERGMs model for the ego-networks is definitely not generative. This is due to the practically convenient decision to model the ego-networks as independent of each other given the cluster memberships and cluster-wise ERGM parameters. The ego-networks overlap, to varying degrees, with each other and thus it is not possible to simulate a coherent overall network comprising the ego-networks arising from independent ERGMs. We have illustrated our ego-ERGM approach using several example datasets with contrasting roles being played in the network, thus demonstrating the flexibility of the method. We consign the use of available node covariates and model selection in terms of which ERGM summary statistics to include to future work.

Acknowledgments

This material is based upon on works supported by the Science Foundation Ireland under Grant No. 08/SRC/I1407: Clique: Graph & Network Analysis Cluster (MST and TBM). Revisions between the initial submission and the

final version were performed under funding from NIH grant R01 HG006399 (MST) and The Insight Centre for Data Analytics, supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289 (TBM).

References

Besag J. Statistical analysis of non-lattice data. The Statistician. 1975; 24(3):179–195.

Blei, DM.; Fienberg, SE.; McCallum, A.; Shalizi, CR.; Handcock, MS. Panel discussion. In: Airoldi, E.; Blei, DM.; Fienberg, SE.; Goldenberg, A.; Xing, EP.; Zheng, AX., editors. Statistical Network Analysis: Models, Issues, and New Directions, Volume 4503 of Lecture Notes in Computer Science. Berlin: Springer; 2007. p. 209-222.

Borgatti SP, Everett MG. Models of core/periphery structures. Social Networks. 1999; 21:375–395.

Brandes, U.; Lerner, J. Role-equivalent actors in networks. ICFCA Satellite Workshop on Social Network Analysis and Conceptual Structures: Exploring Opportunities; 2007.

Brandes U, Lerner J. Structural similarity: Spectral methods for relaxed blockmodeling. Journal of classification. 2010; 27(3):279–306.

Caimo A, Friel N. Bayesian inference for exponential random graph models. Social Networks. 2011; 33(1):41–55.

Caimo A, Friel N. Bayesian model selection for exponential random graph models. Social Networks. 2013; 35(1):11–24.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B. 1977; 39(1):1–38.

Everett MG, Borgatti SP. Regular equivalence: General theory. Journal of Mathematical Sociology. 1994; 19(1):29–52.

Fisher D. Using egocentric networks to understand communication. Internet Computing, IEEE. 2005; 9(5):20–28.

Fisher, D.; Smith, M.; Welser, HT. You are who you talk to: Detecting roles in usenet newsgroups. Proceedings of the 39th Annual Hawaii International Conference on System Sciences; 2006. p. 59b

Friedman JH. On bias, variance, 0/1loss, and the curse-of-dimensionality. Data mining and knowledge discovery. 1997; 1(1):55–77.

Geyer CJ, Thompson EA. Constrained Monte Carlo maximum likelihood for dependent data. Journal of the Royal Statistical Society Series B. 1992; 54(3):657–699.

Gleave, E.; Welser, HT.; Lento, TM.; Smith, MA. A conceptual and operational definition of ‘social role’ in online community. Proceedings of the 42nd Annual Hawaii International Conference on System Sciences; 2009. p. 1-11.

Goldenberg A, Zheng AX, Fienberg SE, Airoldi EM. A survey of statistical network models. Foundations and Trends in Machine Learning. 2010; 2:129–233.

Goodman LA, Kruskal WH. Measures of association for cross classifications. Journal of the American Statistical Association. 1954; 49(268):732–764.

Handcock, MS.; Hunter, DR.; Butts, CT.; Goodreau, SM.; Krivitsky, PN.; Morris, M. ergm: A package to fit, simulate and diagnose exponential-family models for networks. 2011. Version 2.4-2

Handcock MS, Raftery AE, Tantrum JM. Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A. 2007; 170(2):1–22.

Harrigan M, Archambault D, Cunningham P, Hurley NJ. Egonav: exploring networks through egocentric spatializations. AVI. 2012:563–570.

Holland PW, Leinhardt S. An exponential family of probability distributions for directed graphs. Journal of the American Statistical Association. 1981; 76(373):33–50.

Hunter DR, Goodreau SM, Handcock MS. Goodness of fit of social network models. Journal of the American Statistical Association. 2008; 103:248–258.

Hunter DR, Handcock MS. Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics. 2006; 15:565–583.

Hunter DR, Krivitsky PN, Schweinberger M. Computational statistical methods for social network models. Journal of Computational and Graphical Statistics. 2012; 21(4):856–882. [PubMed: 23828720]

Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995; 90(430): 773–795.

Krivitsky PN, Handcock MS, Morris M. Adjusting for network size and composition effects in exponential-family random graph models. Statistical Methodology. 2011; 8(4):319–339. [PubMed: 21691424]

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

Lerner, J. Role assignments. In: Brandes, U.; Erlebach, T., editors. Network Analysis, Volume 3418 of Lecture Notes in Computer Science. Springer; Berlin/Heidelberg: 2005. p. 216-252.

Morris M, Handcock MS, Hunter DR. Specification of exponential-family random graph models: terms and computational aspects. Journal of statistical software. 2008; 24(4):1548.

Murphy TB, Dean N, Raftery AE. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Annals of Applied Statistics. 2010; 4:396–421. [PubMed: 20936055]

Nowicki K, Snijders TAB. Estimation and prediction of stochastic blockstructures. Journal of the American Statistical Association. 2001; 96(455):1077–1087.

R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

Robins G, Pattison P, Kalish Y, Lusher D. An introduction to exponential random graph (p*) models for social networks. Social Networks. 2007; 29(2):173–191.

Robins G, Snijders T, Wang P, Handcock M, Pattison P. Recent developments in exponential random graph (p*) models for social networks. Social Networks. 2007; 29(2):192–215.

Salter-Townshend M, White A, Gollini I, Murphy TB. Review of statistical network analysis: models, algorithms, and software. Statistical Analysis and Data Mining. 2012; 5(4):243–264.

Snijders TAB. Markov Chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure. 2002; 3:1–40.

Snijders TAB, Nowicki K. Estimation and prediction for stochastic bockmodels for graphs with latent block structure. Journal of Classification. 1997; 14(1):75–100.

Stanford D, Raftery AE. Approximate Bayes factors for image segmentation: the Pseudolikelihood Information Criterion (PLIC). IEEE Transactions on Pattern Analysis and Machine Intelligence. 2002; 24(11):1517–1520.

Steinley D, Brusco MJ, Wasserman S. Clusterwise p* models for social network analysis. Statistical Analysis and Data Mining. 2011; 4(5):487–496.

Strauss D, Ikeda M. Pseudolikelihood estimation for social networks. Journal of the American Statistical Association. 1990; 85(409):204–212.

van Duijn MAJ, Gile KJ, Handcock MS. A framework for the comparison of maximum pseudo-likelihood and maximum likelihood estimation of exponential family random graph models. Social Networks. 2009; 31:52–62. [PubMed: 23170041]

Welser, HT.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; Smith, M. Finding social roles in Wikipedia. Proceedings of the 2011 iConference; New York, NY, USA. ACM; 2011. p. 122-129.

Welser HT, Gleave E, Fisher D, Smith M. Visualizing the signatures of social roles in online discussion groups. Journal of Social Structure. 2007; 8(2):1–32.

A Simulated Dataset

We first present results of a simulation study that does not correspond to ego-networks extracted from a single overall network. We thus avoid the issues detailed in Section 7 about generative models. The ground truth is available for this dataset which facilitates direct

performance assessment of our code, algorithms and approximations. Unlike the results for the other datasets, we do not employ the offset term of Krivitsky et al. (2011) thus allowing for more direct comparison between the results and the ground truth.

We simulated 50 undirected networks from each of 3 ERGM types defined on 3 commonly used network sufficient statistics: the number of edges, the geometrically weighted edgewise shared partner distribution and the geometrically weighted degree distribution (Hunter and Handcock (2006)). We use model parameters equal to

$$\underline{\tau} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} -3 & 0 & 1 \\ -1 & -2 & -1 \\ -2 & 0 & 2 \end{bmatrix}$$

to generate data.

We then apply our algorithm for fitting a mixture of ERGMs as described in Section 4. When we ran this experiment 50 times and (correcting for label switching) the mean number of mis-clustered networks out of 150 was just 1.78. In fact, in 24 runs there were no mis-assigned networks and in one case there were 50 which was due to an empty cluster being formed. The EM algorithm found averaged estimates of $\hat{\tau}$ and $\hat{\boldsymbol{\theta}}$ equal to:

$$\underline{\hat{\tau}} = \begin{bmatrix} 0.3331 \\ 0.3388 \\ 0.3389 \end{bmatrix} \quad \hat{\boldsymbol{\theta}} = \begin{bmatrix} -0.02 & 0.94 & -0.29 \\ 1.97 & -5.54 & -0.96 \\ 1.01 & -0.00 & 1.92 \end{bmatrix}.$$

The standard deviations across the 50 repetitions were:

$$SD(\boldsymbol{\theta}) = \begin{bmatrix} 0.18 & 0.06 & 0.16 \\ 0.13 & 0.99 & 0.14 \\ 0.18 & 0.03 & 0.51 \end{bmatrix}.$$

This shows that although our method performs extremely well at correctly clustering the ERGM networks and estimating the mixture proportions, the ERGM parameters are not accurately estimated. We note here that the standard deviations for non-mixture based estimates are high and that the use of the superior MCMLE algorithm does not reduce these uncertainties. The averaged estimates for the ERGM parameters under MPLE and MCMLE across 50 runs were:

$$\hat{\boldsymbol{\theta}}_{MPLE} = \begin{bmatrix} -3.28 & 1.01 & 0.76 \\ -1.27 & -5.61 & -0.68 \\ -2.28 & -0.04 & 3.33 \end{bmatrix} \quad \hat{\boldsymbol{\theta}}_{MCMLE} = \begin{bmatrix} -2.60 & 1.07 & -0.65 \\ -1.35 & -11.23 & -0.60 \\ -2.06 & -0.05 & 2.59 \end{bmatrix}.$$

The standard deviations for the MPLE and MCMLE algorithms for networks simulated for these ERGMs across 50 simulations are:

$$SD_{MPLE}(\boldsymbol{\theta}) = \begin{bmatrix} 1.44 & 0.49 & 5.26 \\ 1.20 & 0.79 & 1.49 \\ 1.20 & 0.21 & 3.40 \end{bmatrix} \quad SD_{MCMLE}(\boldsymbol{\theta}) = \begin{bmatrix} 1.91 & 0.64 & 4.01 \\ 1.11 & 3.03 & 1.39 \\ 1.03 & 0.21 & 2.69 \end{bmatrix}.$$

This is due to the high correlation between the ERGM terms leading to a highly multi-modal likelihood surface. Thus points on the likelihood that have similar values may have parameters that are quite different. The overlap of terms in $S(\boldsymbol{Y})$ causes highly correlated entries in the $\underline{\theta}_g$ vector.

For these reasons BIC isn't always able to correctly identify the number of groups G (see Figure 3). BIC selects 4 clusters in this example, although evidence for 4 over 3 clusters is not strong and the 3 cluster model has the second highest BIC value. The difference in BIC from 3 to 4 clusters is just 6.44; Kass and Raftery (1995) suggest a difference of 10 for strong evidence of one model over another.

Encouragingly, no nodes/egos were incorrectly assigned to another group; when $G = 3$, all nodes are typically correctly clustered and when $G > 3$, the ground truth clusters are split, however they are not merged (i.e. if a node that should belong to cluster g is assigned to cluster g' then only other nodes that should belong to g are assigned by our method to g'). Thus when BIC selects a model with too many clusters, grouping the clusters together restores the correct underlying grouping.

For selection of ERGM terms, BIC was not found to perform well. This is due to the multi-collinearity of ERGM terms which causes competing models to have similar likelihoods. See Hunter et al. (2008) for more details. Fortunately, given a mis-specified ERGM model that includes terms that are related to the generative model, the clustering performance of our algorithm was not statistically significantly effected.

We also compared the results for the two-stage method across these 50 simulations. We found that the average standard deviation for the θ estimates was 0.671 for the two-stage model and 0.34 for the mixture model after EM convergence. However, as we have discussed above, our interest is not in these estimates but in the ability of a method to correctly capture the underlying patterns of local connectivity. The average number of mis-classified egos for the mixture model as 1.78 but for the two-stage method it rises to 51.42, indicating that the EM algorithm greatly improves the model fit.

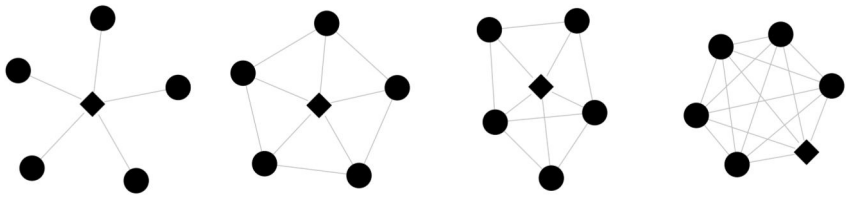

Figure 1.
Differing Alters' Behavior: The ego is square and differing ego roles is illustrated via motifs. In all four cases the ego has a degree of 5 but the pattern of connectivity between the alters varies.

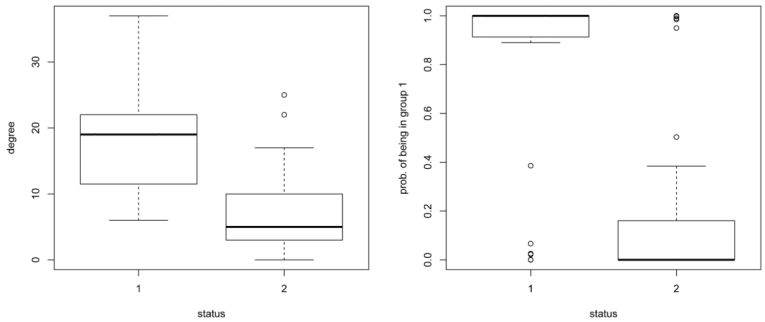

Figure 2.

Boxplots of the degree (left) and probability of belonging to one of two groups (right) under our model split by the status of the lawyers. 1 indicates a partner and 2 an associate. It can clearly be seen that our model picks up a strong signal for indicating status in the law firm, based on the link patterns of the extracted ego-networks that is not wholly captured by splitting into high and low degree nodes.

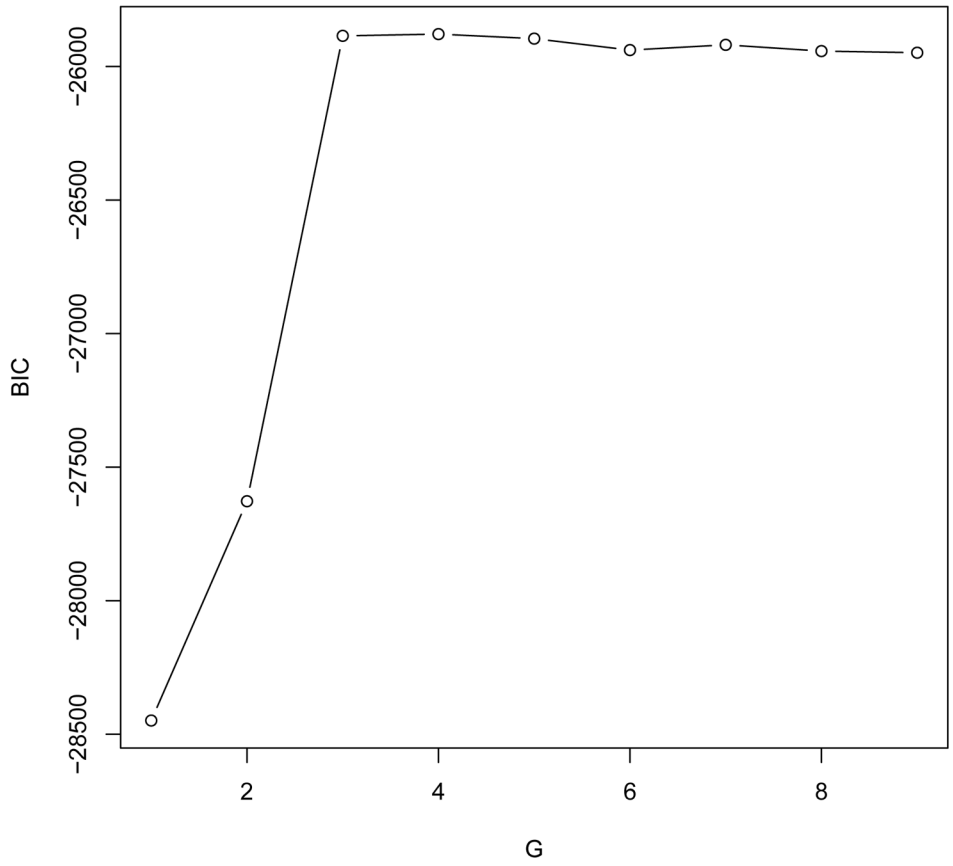

Figure 3.

A plot of BIC against *G* for the simulated data example, based on 3 underlying clusters. The method selects more clusters (4) than were used to generate the data, however clusters tended to split but not merge for models with too many clusters. i.e. the method successfully separates the underlying clusters when too many clusters are modelled.

Table 1

Summary of the results for the Lazega Lawyers dataset. 3 models clusterings are compared to the observed status of the lawyers in the network; ego-ERGM with 3 groups, ego-ERGM with 2 groups and a descriptive model using degree only.

	ego-ERGM$_{G=3}$			**ego-ERGM$_{G=2}$**		**Degree**	
	G_1	G_2	G_3	G_1	G_2	**low**	**high**
Partner	6	28	2	28	8	13	23
Associate	20	7	8	7	28	25	10

Table 2

Lazega Lawyers run times and association as measured by Goodman and Kruskal's Lambda between various role assignments. Although some methods appear have low Lambda values, all methods have a p-value less than 2.2×10^{-16} in Pearson's χ^2 test for contingency.

	Reg. Equiv.	2-colour	3-colour	2-stage	ego-ERGM
Time(s)	< 1	< 1	1	5	121
χ^2 p-value	0.548	0.0498	0.00237	0.00177	7.035×10^{-6}
Lambda	0.03	0.06	0.37	0.34	0.57

Table 3

Prosper run times and association as measured by Goodman and Kruskal's Lambda between various role assignments. Although some methods appear have low Lambda values, all methods have a p-value less than 2.2×10^{-16} in Pearson's χ^2 test for contingency.

	Reg-Equiv	2-colour	3-colour	2-stage	ego-ERGM
Time(s)	1800	210	214	243	1081
Lambda	0.39	0.00	0.04	0.17	0.23