TRACTABLE AND CONSISTENT RANDOM GRAPH MODELS

ARUN G. CHANDRASEKHAR[‡] AND MATTHEW O. JACKSON[⋆]

Abstract. We define a general class of network formation models, Statistical Exponential Random Graph Models (SERGMs), that nest standard exponential random graph models (ERGMs) as a special case. We analyze conditions for practical and consistent estimation of the associated network formation parameters, addressing two open issues in the estimation of exponential random graph models. First, there are no previous general results on whether estimates of such a model's parameters based a single network are consistent (i.e., become accurate as the number of nodes grows). Second, a recent literature has shown that standard techniques of estimating ERGMs have exponentially slow mixing times for many specifications in which case the software used for estimating these models will be unreliable. SERGMs reformulate network formation as a distribution over the space of sufficient statistics instead of the space of networks, greatly reducing the size of the space of estimation and making estimation practical and easy. We identify general classes of models for which maximum likelihood estimates are consistent and asymptotically normally distributed. We also develop a related, but distinct, class of models that we call subgraph generation models (SUGMs) that are useful for modeling sparse networks and whose parameter estimates are also consistent and asymptotically normally distributed. We show how choice-based (strategic) network formation models can be written as SERGMs and SUGMs, and illustrate the application of our models and techniques with network data from villages in Karnataka, India.

JEL Classification Codes: D85, C51, C01, Z13.

Keywords: Random Networks, Random Graphs, Exponential Random Graph Models, Exponential Family, Social Networks, Network Formation, Consistency, Sparse Networks, Multiplex, Multigraphs

Date: December 2011, Revision: September 2013.

We thank Isaiah Andrews, Larry Blume, Gabriel Carroll, Victor Chernozhukov, Esther Duflo, Ben Golub, Bryan Graham, Marcel Fafchamps, Randall Lewis, Angelo Mele, Stephen Nei, Elie Tamer, Juan Pablo Xandri and Yiqing Xing for helpful discussions and/or comments on earlier drafts, and especially Andres Drenik for valuable research assistance. Chandrasekhar thanks the NSF Graduate Research Fellowship Program. Jackson gratefully acknowledges financial support from the NSF under grants SES-0961481 and SES-1155302 and from grant FA9550-12-1-0411 from the AFOSR and DARPA, and ARO MURI award No. W911NF-12-1-0509.

[‡]Department of Economics, Stanford University; Microsoft Research, New England.

[⋆]Department of Economics, Stanford University; Santa Fe Institute; and CIFAR.

1. Introduction

> ...[A] pertinent form of statistical treatment would be one which deals with social configurations as wholes, and not with single series of facts, more or less artificially separated from the total picture.
>
> *Jacob Levy Moreno and Helen Hall Jennings, 1938.*

To what extent is someone's proclivity to form relationships influenced by whether those relationships are in public or private? For example, are people of different types, e.g., different castes or races, more reluctant to form relationships across types when they have a friend in common than when they do not? The answer to such a question has implications for communication, learning, inequality, diffusion of innovations, and many other behaviors that are network-influenced. Being able to statistically test whether people's tendencies to interact across groups depends on social context requires allowing for correlation in relationships within a network.

Beyond this illustrative question, correlations in relationships are important in many other social and economic settings: from informal favor exchange where the presence of friends in common can facilitate robust favor exchange (e.g., Jackson et al. (2012)), to international trade agreements where the presence of one trade agreement can influence the formation of another (e.g., Furusawa and Konishi (2007)). Similarly, in forming a network of contacts in the context of a labor market, an individual benefits from relationships with others who are better-connected and hence relationships are not independently distributed (e.g., Calvo-Armengol (2004); Calvo-Armengol and Zenou (2005)); nor are they in a setting of risk-sharing (e.g., Bramoullé and Kranton (2007)).

Once such interdependencies exist, estimation of a network formation model cannot take place at the level of pairs of nodes, but must encompass the network as a whole, as reflected in the quote from Moreno and Jennings (1938) above. Exponential random graph models (henceforth "ERGMs") incorporate such interdependencies and thus have become the workhorse models for estimating network formation.[1] Indeed, as originally shown via a powerful theorem by Hammersley and Clifford (1971), the exponential form can nest *any* random graph model and can incorporate arbitrary interdependencies in connections.[2] Moreover, ERGMs admit a variety of strategic (choice-based) network formation models, as we show below and others have shown in other contexts.

Although ERGMs are widely used and are seemingly natural tools for estimating the formation of networks with interdependent links, there are two critical gaps in the understanding of these models and we address both gaps in this paper.

[1]These grew from work on what were known as Markov models (e.g., Frank and Strauss (1986)) or $p*$ models (e.g., Wasserman and Pattison (1996)). An alternative approach is to simply work with regression models at the link (dyadic) level, but to allow for dependent error terms, as in the "MRQAP" approach (e.g., see Krackhardt (1988)). Although that approach can work theoretically, it is less well suited for identifying the incidence of particular network subgraph structures that may be implied by various social or economic theories of the type that we wish to allow for here.

[2]Their theorem applies to undirected and unweighted networks. See the discussion in Jackson (2008). Of course, the representation can become fairly complicated; but the point is that the ERGM model class is broadly encompassing. Furthermore, as we illustrate below, it can also be adapted to allow for multigraphs such that nodes can have multiple types of relationships.

First, the number of possible networks on a given number of nodes is an exponential function of the number of nodes. Estimating the likelihood of a given network requires having some estimate of its likelihood relative to the other networks that could have appeared instead. Of course, given the exponential explosion of the number of possible networks,[3] it is impossible to directly calculate likelihoods of a given network appearing and so some approximation is necessary. However, this computational hurdle is so formidable that even state-of-the-art algorithms for estimation can be inadequate and inaccurate. This is the subject of a burgeoning literature that shows that standard estimation techniques cannot mix in less than exponential time for large classes of exponential random graph models (e.g., see Bhamidi et al. (2008) and Chatterjee et al. (2010)).[4]

Second, there is little that is known about the consistency of parameter estimates of such exponential random graph models: do estimated parameters converge to the true parameters as the size of the network grows?[5] It is important to emphasize that the relevant asymptotic frame is one in which the number of nodes grows to infinity, rather than the number of networks; since in many applications, the data consist of a single network.[6] If links are all independent of each other, then a single network provides many observations and link probabilities are estimable. However, once there are nontrivial interdependencies between links, standard asymptotic results do not apply, which is why little is known about the consistency of associated estimation techniques. This does not mean that consistency is precluded, (just as it is not precluded in time series or spatial settings) as even though the probability must be specified at the network level there is still a lot of information that can be discerned from the observation of a single large network.[7] Nonetheless, it does mean that we must approach the asymptotics in a

[3]For example with only $n = 30$ nodes, the number of possible networks is 2^{435} (which exceeds most estimates of the number of atoms in the universe).

[4]Apart from estimation issues, there is another issue that concerns which formulations of exponential random graph models are distinguished from independent link models. For example, Chatterjee and Diaconis (2011) show that some classes of ERGMs are indistinguishable in a well-defined sense from independent-link models. Our formulations avoid such issues, and in particular our results on sparse networks apply to classes of models that are well-distinguished from independent-link models and to which the Chatterjee and Diaconis (2011) results do not apply.

[5]Of course, consistency has been examined in the context of some random graph models (e.g., see Bickel et al. (2011); but not for the class of models that we consider here. Consistency of a different sort has been examined by Shalizi and Rinaldo (2012) in the context of ERGMs: whether or not if one observes only a subsample of the network, whether the estimated parameters will accurately reflect the true parameters of the full network. That is also related to work on measurement error by Chandrasekhar and Lewis (2013), but we do not address that separate issue here.

[6]If the asymptotic frame was one in which we observed a growing number of networks with a fixed number of nodes, each generated by the same network formation process, consistency would be follow from standard results. Although there exist such data sets, they are the exception rather than the rule. And, even for such data sets, the above estimation issues of practical calculations that we tackle here are still relevant.

[7]In cases where a time series of link formation is observed or postulated and estimated, then one can take advantage of the sequentiality to see how each link forms conditional on the network in place at the time (e.g., see Christakis et al. (2010)). However, without such information, or in cases where links may further evolve over time, the network perspective again prevails. For example, Mele (2011) considers a sequential model and then shows that it becomes effectively equivalent to a certain ERGM.

way that accounts for the potentially complex correlations and interdependencies that arise in link formation.[8]

To fix ideas and discuss our approach in more detail, let us be more explicit about the issues that ERGMs face. In an ERGM, the probability of a network of observing a particular network described by a graph g with an associated vector of statistics $S(g)$ (for example, the density of links, number of cliques of given sizes, the average distance between nodes with various characteristics, counts of nodes with various degrees, and so forth) is proportional to

$$\exp\left(\theta \cdot S\left(g\right)\right)$$

where θ is a vector of model parameters.[9]

Turning the above expression into a probability of observing a network g requires normalizing this expression by summing across all possible networks, and so the probability of observing g is

$$\mathrm{P}_\theta\left(g\right) = \frac{\exp\left(\theta \cdot S\left(g\right)\right)}{\sum_{g'} \exp\left(\theta \cdot S\left(g'\right)\right)}. \tag{1.1}$$

The challenge is that estimating the parameters this sort model requires estimating how the relative likelihood of a network relates to parameters. This involves either explicitly or implicitly estimating the denominator of (1.1) - and this is true of any of a variety of estimation techniques whether it be maximum likelihood, generalized method of moments, or even Bayesian. However, as we mentioned above, examining all possible networks g' is infeasible even for small n, and thus, one has to take other routes.

The adaptation of Markov Chain Monte Carlo (MCMC) sampling techniques to draw networks and estimate ERGMs, by Snijders (2002) and Handcock (2003), provided a breakthrough. The subsequent development of computer programs based on those techniques led to their widespread use.[10] However, it was clear to the developers and practitioners that the programs had convergence problems for many specifications of ERGMs. Until recently, it remained unknown if or when these techniques would mix accurately in a feasible time. Given the huge set of networks g' to sample, any MCMC procedure can visit only an infinitesimal portion of the set, and it was unclear whether such a technique would lead to an accurate estimate in any practical amount of time. Unfortunately, important recent papers have shown that for broad classes of ERGMs standard MCMC procedures will take exponential time to mix unless the links in the network are approximately independent (e.g., see the discussions in Bhamidi et al. (2008) and Chatterjee et al. (2010)). Of course, if links are approximately independent then was no real need for an ERGM specification to begin with, and so in cases where ERGMs are really needed they cannot be accurately estimated by such MCMC sampling techniques. Such difficulties were well-known in practice to users of software

[8]Clearly, there are other settings with interdependencies, such as time series and spatial settings. The network setting presents a complex set of interdependencies that do not permit off-the-shelf approaches.

[9]The reasons for using the exponential family are clear. It nests many standard distributions such as multivariate normal, Poisson, power, lognormal, gamma, beta, Weibull, Laplace, multinomial, etc. Also, it has many nice properties regarding its cumulants and moment generators, and has been well-studied in the statistics literature.

[10]See Snijders et al. (2006) for more discussion.

programs that perform such estimations, as rerunning even simple models can lead to very different parameter and standard error estimates, but now these difficulties have been proven to be more than an anomaly.

Beyond the computational challenges, it is also not known whether various estimators of ERGMs are consistent. Will maximum likelihood estimates converge to the true parameters as the number of nodes becomes large? Given that data in many settings consist of a single network or a handful of networks, we are interested in asymptotics where the number of nodes in a graph grows. However, it may be the case that increasing the number of nodes does not increase the information in the system. In fact, for some sequences of network statistics and parameters it is obvious that the parameters of the associated ERGM are *not* consistent. For example, suppose that $S(g)$ includes a count of the number of components in the network and the parameters are such that the network consists of a single or a few components. The limited number of components would not permit consistent estimation of the generative model. Thus, there are models where consistent estimation is precluded. On the other extreme where links are all independent, we know that consistent estimation would hold, and so the interesting question is for which models is it that consistent estimation can be obtained.

This paper makes five contributions.

First, we propose a generalization of the class of ERGMs that we call SERGMs: *Statistical ERGMs*. To understand the generalization, note that in any ERGM the probability of forming a network is determined by its statistics: for instance, having a given link density, a given clustering coefficient, specific path lengths, etc. Most importantly, every network exhibiting the same statistics is equally likely.[11] SERGMs nest the usual ERGM models by noting that: (i) we can define the model directly over the statistics and thus greatly reduce the dimensionality of the space, and (ii) we can weight the distribution over the space of statistics in many ways other than simply by how many networks exhibit the same statistics. Some of these reference distributions generate natural models that both allow for realistic features as well as desirable statistical properties such as consistency and asymptotic normality of the estimators of model parameters. This change to the space of statistics rather than networks allows us to develop computationally practical techniques for estimation of SERGMs.

Second, we examine sufficient conditions as well as some necessary conditions for consistent estimation of SERGM parameters (nesting ERGMs as a special case) and identify a class of SERGMs for which it is both easy to check consistency and estimate parameters. Models in this class are based on "count" statistics: for instance, how many links between nodes with certain characteristics exist, how many triangles[12] including certain types of nodes exist, how many nodes have a given degree, and so forth.

Third, we identify a related class of network formation models that are based on the formation of subgraphs that we call SUGMs (*Subgraph Generated Models*).[13] We

[11]This is related to the well-known property of sufficient statistics of the exponential family. As an analogy, a binomial distribution defines the probability of seeing x heads but does not care about the exact sequence under which the x heads arrive.

[12]Triangles refer to triads: cliques of size three; that is, triplets of nodes i, j, k that include all three possible links.

[13] Although some particular examples of random networks have previously been built up from randomly generated subgraphs (Bollobás et al. (2011)), our general specification of SUGMs is new.

can think of such a network as being constructed from building blocks of varying sizes: links, triangles, larger cliques, stars, etc., layered upon each other. We show that if such models are sufficiently *sparse* in a well-defined way, then they are consistently estimable and parameters are asymptotically normally distributed. Such sparse networks appear in many applications as they have realistic features (e.g., average degree that grows at a rate less than n, but still allowing for high clustering, homophily, rich degree distributions, and so forth). These models are also easily to simulate and admit random utility foundations that may be used in economic applications (e.g., counterfactual policy analysis, testing theory).

Our fourth contribution is to provide a set of strategic network formation models that mix utility-based choices of link and subgraph formation by agents with randomness in meeting opportunities. We describe two basic approaches (one based on consent in link and subgraph formation and the other based on noncooperative search intensity choices), showing how these can provide foundations for classes of SERGMs and SUGMs, and illustrate one of them in our applications section.

Our fifth and final contribution is to provide illustrations of the techniques developed here by applying them to data on social networks from Indian villages. We show that many patterns of empirical networks are replicable by a parsimonious SUGM with very few parameters. We also answer the question that we began with above, of whether individuals tend to form cross-caste relationships more frequently when there are no friends in common than when there are. We find that cross caste relationships occur with significantly higher frequency when in isolation than when embedded in triads. Beyond this, we also develop an extension of the models that apply to multigraphs (so individuals may have different sorts of edges between them). This allows us to then test several theories of how multigraphs are formed (why links are correlated or multiplexed): fixed costs of link formation, patterns that foster favor exchange, and correlated (un)-observables. We find evidence consistent with predictions of a theory of favor exchange: being a member of a triangle may substitute for having multiple types of relationships with others.

The connection between ERGMs, SERGMs, and SUGMs is as follows. SERGMs not only provide an alternative way of representing ERGMs by working directly with the network statistics, but also substantially generalize the class by allowing for alternative reference distributions. SUGMs then allow for an additional change relative to SERGMs in terms the way the graph is generated. A SERGM – in order to maintain the nesting of ERGMs – has the likelihood of a network depend on the *observed* counts of various statistics, including subgraphs. A SUGM can be thought of as generating subgraphs, but allowing them to overlap: it is not clear whether a given triangle was generated directly as a triangle or as three separate links. Thus, one needs to infer the true statistics in estimating the parameters of the model. This subtle change allows for a more direct estimation in the case of sparse networks. We provide an exact relationship between SUGMs and SERGMs, showing that these models are related but distinct.

The remainder of the paper is organized as follows. In Section 2 we provide an overview of the paper by way of a simple example. We provide formal definitions of the framework in Section 3. In Section 4 we define and discuss SERGMs, providing estimation and consistency results. In Section 5, we develop a variation of the model

of network formation based on subgraphs, SUGMs, which we show to be easily and consistently estimable in the case of sparse networks, In Section 6 we provide further extensions and applications of these models, including how they may be used in analyzing strategic network formation and extensions to multigraphs. In Section 7 we present simulation exercises to demonstrate asymptotic properties, as well as empirical applications to Indian village networks that illustrate the techniques and some of the results. Section 8 concludes.

2. Preliminaries and an Example

Let $\mathcal{G}^n$ be a set of possible graphs on a finite number n of nodes. The class can consist of undirected or directed graphs and unweighted or weighted graphs. In the unweighted case, we take $\mathcal{G}^n$ to be finite, but it will generally be large as a function of n. For instance, if $\mathcal{G}^n$ is the set of all undirected, unweighted graphs on n nodes, then the cardinality of $\mathcal{G}^n$ is $|\mathcal{G}^n| = 2^{\binom{n}{2}}$.

We often omit notation $\mathcal{G}^n$ and denote a generic network by g. Thus, if we write $\sum_g$ it is understood to mean $\sum_{g\in\mathcal{G}^n}$ for whatever class of networks is relevant. Unless otherwise stated we take $\mathcal{G}^n$ to be the set of undirected, unweighted graphs; but as will be clear, the results extend directly to more general classes such as directed graphs and multigraphs as we illustrate below.

We observe a single (large) graph from which to estimate a network formation model.[14] A family of models is indexed by a vector of parameters of interest β, and can be represented by corresponding probability distributions over graphs $\mathrm{P}_\beta(g)$, which depends on parameters β.

Some of our results concern asymptotic properties of such models, and so at times we consider a sequence of random graphs g_n, $n \in \mathbb{N}$, drawn from a sequence of probability distributions $\mathrm{P}^n_{\beta_n}(\cdot)$. Since everything then carries an n index we suppress it except when we want to highlight dependence.

The models that we develop can be expressed as functions of characteristics of networks. A vector of *statistics* of a network $g \in \mathcal{G}^n$ is a finite (k-dimensional) vector $S(g) = (S_1(g), \ldots, S_k(g))$, where $S_\ell : \mathcal{G}^n \to \mathbb{R}$ for each $\ell \in \{1, \ldots k\}$. For examine, a statistic might be the number of links in a network, the average path length, the number of cliques of a given size, the number of isolated nodes, the number of links that go between two specific types of groups, and so forth.

2.1. **A Leading Example and a Preview of the Paper.**

We begin with an example that minimally complicates an independent-link model, but enough to require modeling link interdependencies. The idea is that instead of links being formed solely on a bilateral basis there are also multilateral opportunities to form relationships. Subgroups of individuals sometimes randomly meet and decide whether or not to form subgraphs (we develop full random utility foundations of such models in Section 6.1). Specifically, individuals meet in pairs, triples, and larger groups in order to determine whether to form relationships. Both the meeting probabilities

[14] In some contexts a researcher may have access to several or many networks drawn from the same distribution. That can obviously help with estimation, but we do not presume that the researcher has such information.

and the preferences for forming relationships may depend on the characteristics of the pair or larger clique of individuals in question (see Sections 7.1 and 7.2).

For illustrative purposes, we work with a simple version of such a model: individuals meet in pairs and triples and for this section we ignore characteristics of the individuals. We also allow for the presence of *isolates*: asocial individuals who do not form relationships with others. We work with this model since despite its simplicity it works remarkably well in fitting networks in some applications as we show in Section 7.1.

The probability of the formation of a network g can be expressed as a function of the network's: number of isolated nodes, $S_I(g)$; number of links, $S_L(g)$; and number of triangles, $S_T(g)$. Such a model can be expressed in a standard exponential random graph model (ERGM) of the following form. The probability of a network g being formed is

$$\mathrm{P}_\theta(g) = \frac{\exp\left(\theta_I S_I(g) + \theta_L S_L(g) + \theta_T S_T(g)\right)}{\sum_{g'} \exp\left(\theta_I S_I(g') + \theta_L S_L(g') + \theta_T S_T(g')\right)}. \tag{2.1}$$

If $\theta_I = \theta_T = 0$ then this reduces to a standard Erdős-Rényi random graph. The more interesting case is where at least one of $\theta_I \neq 0$ or $\theta_T \neq 0$, so that networks become more ($\theta_T > 0$) or less ($\theta_T < 0$) likely based on the number of triangles they contain - or, similarly, of isolates they contain.

2.1.1. *ERGM Estimation.* The difficulty with estimating such a model is that the number of such networks in the calculation of $\sum_{g'}$ is $2^{\binom{n}{2}}$.[15] Thus, the fraction of networks that can be sampled is necessarily negligible, and unless careful knowledge of the model is used in guiding the sampling, the estimation of the denominator can be inaccurate.

Given that estimating the parameters of an ERGM are thus forced to circumvent direct calculation of the denominator, approximation methods such as MCMC techniques have been used.[16] The rough intuition is that such methods sample some networks (picking a few g's) to estimate the relative sizes of $\exp\left(\theta_I S_I(g') + \theta_L S_L(g') + \theta_T S_T(g')\right)$ from which to extrapolate the $\sum_{g'}$ in the denominator of (2.1) and thus develop a rough estimate of the relative likelihood of the observed data under various specifications of θ. Even with this approach, the space of all possible networks is difficult to sample in a representative fashion. For instance, if one samples say 10000 networks, then one samples on the order of 2^{16} networks out of the possible 2^{1225} on 50 nodes, which is about one out of every 2^{1209} networks. Thus, unless one is very knowledgeable in choosing which networks to sample and how many to sample of different types, or one is very lucky, the sample is unlikely to be even remotely representative of the possible configurations that might occur. Formally, draws generated by the sampling need to be well-mixed in a practical amount of time.

Indeed, the time before which an MCMC technique has a chance to sample enough networks to gain a representative sample is generally *exponential* in the number of links

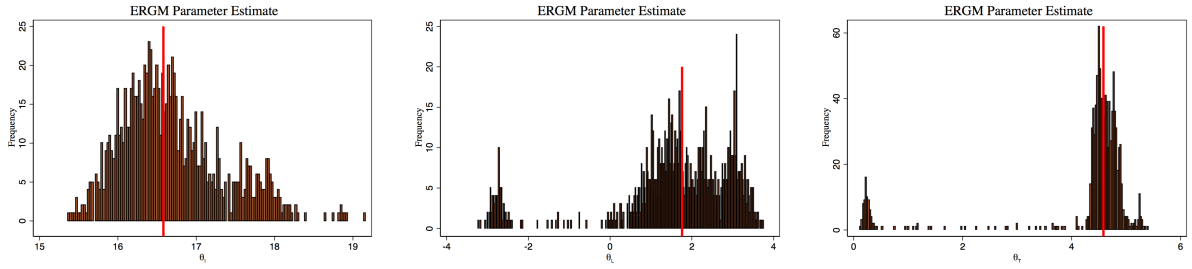
[15]Even with a tiny society of just 30 nodes this is 2^{435}, while estimates of the number of atoms in the universe are less than 2^{258} (Schutz, 2003).

[16]See Snijders (2002), Handcock (2003), as well as discussions in Snijders et al. (2006) and Jackson (2011).

and so is prohibitively large even with a small number of nodes.[17] In particular, an important recent result of Bhamidi et al. (2008) shows that MCMC techniques using Glauber dynamics for estimating many classes of ERGMs mix in less than exponential time *only if* any finite group of edges are asymptotically independent. So, the only time those models are practically estimable is when the links are approximately independent, which precludes the whole reason for using ERGMs: allowing for nontrivial correlations in links.

To illustrate the computational challenges, we a simple model with $n = 50$ nodes. The model consists of some isolated nodes, some randomly generated triangles, and some randomly generated links. In particular, we first select 17 nodes (one third) to be isolated. Next, we generate triangles with a probability of .0014 on each possible triangle on the nodes that are not isolated. Finally, we generate links with probability .0415 on the nodes that are not isolated. Overall, this leads to networks that have on average 20 isolated nodes, 45 links, and 10 triangles (so, $\mathrm{E}[S_I(g)] = 20, \mathrm{E}[S_L(g)] = 45, \mathrm{E}[S_T(g)] = 10$)). We randomly draw 1000 different networks in this manner.

Using standard ERGM estimation software (`statnet` via `R`, Handcock et al. (2003)) we estimate the parameters of an ERGM with isolates, links and triangles for each of these randomly drawn networks. We present the estimates in Figure 1.



(A) Isolate Parameter Estimates

(B) Link Parameter Estimates

(C) Triangle Parameter Estimates

FIGURE 1. Standard ERGM estimation software (`statnet`) output for 1000 draws of networks on 50 nodes, with an average of 20 isolated nodes, 45 links, and 10 triangles. The red lines (on top of each other) are the median left and right 95 percent confidence interval lines (which do not have appropriate coverage).

There are two self-evident issues with the estimation. First, and most importantly, the estimated parameters for links and triangles cover a wide range of values, in fact with the link parameter estimates being both positive and negative and ranging from below -3 to above 3 (Figure 1b) and triangles parameter estimates ranging from just above 0 to more than 5 (Figure 1c). Only the isolates parameter estimates are stable (Figure 1a). Second, despite the enormous variation in estimated parameter values from very similar networks, the reported standard errors are quite narrow and almost

[17]This does not even include difficulties of sampling. For example, as discussed by Snijders et al. (2006), a technique of randomly changing links based on conditional probabilities of links existing for given parameters can get stuck at complete, empty, or other extreme networks.

always report that the parameter estimates are highly significant. Moreover, the median left and right standard error bars essentially coincide and do not come close to capturing the actual variation.

In Appendix D we do some additional diagnostics to confirm that this is really due to the impossibility of practical estimation of ERGMs and not simply due to the variation in simulated networks. There we report the distribution of the statistics from the simulated networks (Figure 9) – they are fairly tightly clustered about the mean values.

As an acid test of the estimation procedure, we also do the following exercise. We randomly generate networks that have *exactly* 20 isolates, 45 links and 10 triangles on 50 nodes. Thus, the statistics of the networks are identical, and only the location of the links and triangles changes. Any two networks with exactly the same statistics should lead to exactly the same parameter estimates as they have exactly the same likelihood under all parameter values. Thus, the only variation comes from imperfections in the software and estimation procedure. As illustrated in Appendix D (Figure 8), although there is slightly less noise in the parameter estimates, they still cover similar ranges and exhibit similar features, and have similar difficulties in the standard error calculations.

Finally, we perform another exercise. Each of the 1000 simulated networks generates parameter estimates. Using those parameter estimates we simulate a network using Statnet's simulation command. We then check whether the simulated networks come anywhere close to matching the original networks. Although the networks turn nearly 20 isolates, they generally have hundreds of links and thousands of triangles (Figure 10), not at all matching the original networks (Figure 9).

2.1.2. *A Prélude to Our Approach.* We develop two new classes of models, both of which are partly built on the following insight, which one can see by rewriting the model above in ways that make it practical to calculate.

Given the model specified in (2.1), any two networks that have the same numbers of isolates, links, and triangles have the same probability of forming. That is, if $(S_I(g), S_L(g), S_T(g)) = (S_I(g'), S_L(g'), S_T(g'))$, then $\mathrm{P}_\theta(g) = \mathrm{P}_\theta(g')$ for any θ. This is simply an observation that $(S_I(g), S_L(g), S_T(g))$ is a sufficient statistic for the probability of the network g. More generally, whichever statistics on which an ERGM is based are sufficient statistics for the probability of a given network forming. This simplifies the calculation above.

Given a vector of statistics S (e.g., $S = (S_I, S_L, S_T)$ in our example), let

$$N_S(s) := |\{g \in \mathcal{G}^n : S(g) = s\}|$$

denote the number of graphs that have statistics s.

We rewrite the denominator of the ERGM in (2.1) as

$$\sum_{s'} N_{S_I,S_L,S_T}(s') \exp\left(\theta_I s'_I + \theta_L s'_L + \theta_T s'_T\right).$$

Note that the denominator now sums across the set of possible numbers of links and triangles. While the denominator of the ERGM in (2.1) was a summation over a number of networks which is of order 2^{n^2}, the summation now is over possible numbers of isolates, links, and triangles which is of order n^6 and thus is polynomial in the number of nodes rather than exponential.

Moreover, instead of considering the probability of observing a particular *network*, we can instead ask what the probability is of observing a particular realization of network *statistics*. For instance, what is the probability of observing a network with a given number of links and triangles? Generally, this is what a researcher is interested in rather than which specific network that had a given list of characteristics was realized. We can then express the model in the following form:

$$\mathrm{P}_\theta\left((S_I, S_L, S_T) = s\right) = \frac{N_{S_I,S_L,S_T}(s)\exp\left(\theta_I s_I + \theta_L s_L + \theta_T s_T\right)}{\sum_{s'} N_{S_I,S_L,S_T}(s')\exp\left(\theta_I s'_I + \theta_L s'_L + \theta_T s'_T\right)}. \tag{2.2}$$

This is an example of what we call a Statistical Exponential Random Graph Model, or SERGM, which are defined in their more general form below.

We have thus reduced the complexity of the estimation problem from something that is exponential in the number of nodes, to something that depends on the size of the space of statistics, which is generally polynomial in the number of nodes.

In addition, there are ways to approximate the denominator of (2.2) which can further ease computation burdens. For example, we can estimate the denominator by summing across some subset of s' that has high probability rather than summing over the full set, as although n^6 is polynomial it still can be a large sum to do exhaustively as n grows. In particular, suppose that for some parameter θ, the probability that the observed statistic ends up taking a value in some set A is at least $1-\varepsilon$: $\mathrm{P}_\theta(s \in A) \geq 1-\varepsilon$. Then by setting

$$\overline{\mathrm{P}}_\theta\left((S_I, S_L, S_T) = s\right) = \frac{N_{S_I,S_L,S_T}(s)\exp\left(\theta_I s_I + \theta_L s_L + \theta_T s_T\right)}{\sum_{s'\in A} N_{S_I,S_L,S_T}(s')\exp\left(\theta_I s'_I + \theta_L s'_L + \theta_T s'_T\right)}.$$

it follows that for any $s \in A$

$$\frac{1}{1-\varepsilon} \geq \frac{\overline{\mathrm{P}}_\theta\left((S_I, S_L, S_T) = s\right)}{\mathrm{P}_\theta\left((S_I, S_L, S_T) = s\right)} \geq 1.$$

Thus, we can work with $\overline{\mathrm{P}}_\theta\left((S_I, S_L, S_T) = s\right)$ which only requires computations over $s' \in A$ in its denominator. [18]

2.2. **Statistical ERGMs (SERGMs) and Subgraph Generation Models (SUGMs).**

(2.2) defines a model over network statistics and, in principle, there is nothing special about the weighting function $N_S(\cdot)$, and at times it can be hard to compute or even approximate. This leads us to our more general representation of SERGMs. By replacing the weighting function N_S with some other function $K_S : A \to \mathbb{R}$ we obtain a *statistical exponential random graph model* (SERGM). The associated probability of

[18] We have to worry about determining A since it depends on θ which is presumed to be unknown to the researcher. However, in many models, the probability of various statistics concentrates in a small neighborhood around the observed statistics with high probability, so the above approximation becomes quite useful. By observing s, and then choosing A to be a large enough neighborhood around the observed s, one can be sure that under the true (unobserved) θ, $\mathrm{P}_\theta(A) \geq 1-\varepsilon$. In particular, it is easy to choose a small set A based on the observed s over which to sum the denominator *without knowing* θ, and which with arbitrarily high probability will give an arbitrarily accurate estimate for large enough n. A general version of such a lemma appears as Lemma B.1 in the appendix.

seeing realized number of links and triangles $(S_I, S_L, S_T) = s$ is:

$$\widehat{\mathrm{P}}_\theta\left((S_I, S_L, S_T) = s\right) = \frac{K_{S_I,S_L,S_T}(s)\exp\left(\theta_I s_I + \theta_L s_L + \theta_T s_T\right)}{\sum_{s'\in A} K_{S_I,S_L,S_T}(s')\exp\left(\theta_I s'_I + \theta_L s'_L + \theta_T s'_T\right)}.$$

This is a model that states that the probability that a network exhibits a specific realization of statistics $S = s$ is given by an exponential function of the statistics s. This is an example of the class of models called SERGMs that we develop here, and which nests ERGMs as a special case.

$$\frac{\widetilde{S}_I}{n} \text{ and } \frac{\widetilde{S}_T}{\binom{n-\widetilde{S}_I}{3}} \text{ and } \frac{\widetilde{S}_L}{\binom{n-\widetilde{S}_I}{2}}. \tag{2.3}$$

The other main approach that we develop is as follows. As described above, we can also think of a model in which isolates, links and triangles are formed directly at random (for instance, subgroups of individuals meet and decide whether they want to form a subgraph). The model is then governed by the probabilities p_I that isolates are directly generated, p_L that any given link is directly generated (on non-isolated nodes), and p_T that any given triangle is directly generated (on non-isolated nodes). This model is what we call a Subgraph Generation Model or SUGM.

The challenge in estimating a SUGM is that we observe the resulting network and not the directly generated isolates, links and triangles. For example, if the three links $12, 23, 13$ are all directly generated, then we would observe the triangle 123 in the graph g and not be sure whether it was generated as three links or as a triangle. Nonetheless, by examining the graph we can back out the probabilities for many such models.

Just to a bit more explicit, suppose that under the model there are some numbers of truly (directly) generated isolates $\widetilde{S}_I$, links $\widetilde{S}_L$, and triangles $\widetilde{S}_T$. If we could see these statistics of truly generated links and triangles, then we would estimate

$$p_I = \frac{\widetilde{S}_I}{n} \text{ and } p_T = \frac{\widetilde{S}_T}{\binom{n-\widetilde{S}_I}{3}} \text{ and } p_L = \frac{\widetilde{S}_L}{\binom{n-\widetilde{S}_I}{2}}. \tag{2.4}$$

However, generally we do not observe these statistics $\widetilde{S}_I$, $\widetilde{S}_L$ and $\widetilde{S}_T$. Instead count two things: the number of observed isolates, S_I, triangles, S_T, and links *not in triangles* – which we refer to as *unsupported* links S_U. We then can take two approaches. One is to use these to estimate the number of truly generated isolates, links and triangles, $\widetilde{S}_I$, $\widetilde{S}_L$ and $\widetilde{S}_T$, by calculating the rates at which incidental isolates triangles would be generated. This can be done in many settings, and we develop an algorithm to do this as described in Section 5.3.

Another approach is to use S_I, S_T, S_U to estimate the probability that a triangle forms and the probability that a link forms by simply computing

$$\widehat{p}_I := \frac{S_I}{n} \text{ and } \widehat{p}_T := \frac{S_T}{\binom{n-S_I}{3}} \text{ and } \widehat{p}_L := \frac{S_U}{\binom{n-S_I}{2} - 3S_T}.$$

These will be accurate estimates of the true parameters p_I, p_T, p_L provided that the network is sparse enough, as we show in Theorem 2.[19,20] Sparseness ensures that the fraction of observed subgraphs that are incidentally generated is vanishing. The main idea in direct estimation is that if the subgraphs we care about are relatively sparse in a precise way, then although three links (or some set of links and links in other triangles) can combine to incidentally generate a triangle, it is much more likely that the triangle forms directly. This is not restrictive for many applications as networks in economic and sociological data sets are often sparse.

Let us now return to the example presented in Section 2.1.1 that provided headaches for standard techniques for estimating ERGMs.

We can estimate that either as a SERGM or a SUGM. The SUGM delivers direct estimates for the parameters based on (2.4). For all of the networks

$$\widehat{p}_I = \frac{\widetilde{S}_I}{n} = \frac{20}{50} = .4,\ \widehat{p}_T = \frac{\widetilde{S}_T}{\binom{n-\widetilde{S}_I}{3}} = \frac{10}{\binom{30}{3}} = .002 \text{ and } \widehat{p}_L = \frac{15}{\binom{30}{2} - 30} = .037. \tag{2.5}$$

If we work with a SERGM (on unsupported links) that has weights

$$K_I(s_I) = \binom{50}{s_I} \text{ and } K_T(s_T) = \binom{\binom{30}{3}}{s_T} \text{ and } K_U(s_U) = \binom{\binom{30}{2} - 30}{s_U}, \tag{2.6}$$

then as we show in Theorem 1 and 3, the SERGM parameters can be directly obtained as from the SUGM binomial calculations, with an adjustment for the exponential:

$$\widehat{\theta}_I = \log \frac{\widehat{p}_I}{1-\widehat{p}_I} = -.17,\ \widehat{\theta}_T = \log \frac{\widehat{p}_T}{1-\widehat{p}_T} = -2.7 \text{ and } \widehat{\theta}_U = \log \frac{\widehat{p}_U}{1-\widehat{p}_U} = -1.4.$$

Thus we directly and easily obtain parameter estimates for the same networks that gave the ERGM estimation troubles.

These estimates are obtained from the fixed values of 20 isolates, 45 links and 10 triangles on 50 nodes. Each of the 1000 simulated networks with these expected values will have slightly different realized values, and so we report the full distribution of those estimated parameters in Appendix D in Figures 11 and 12. The distributions of estimated parameters are tightly grouped around their means.

We remark that due to the finite sample, the parameter estimates above exhibit some slight biases. For example, the networks that were generated were generated to have 17 isolated nodes, so the true p_I was .34, while the estimate is .40. This occurs since in some of the networks, so of the other nodes that were not designated to be isolated end up not being a part of any links or triangles. On average in our simulations, this happens to about 3 nodes. Similarly, there may be extra triangles generated incidentally by links.

Several things are worth noting. First, as we show in our results below, for appropriate models this bias disappears for large n and the estimates are consistent. Second, we provide an algorithm in Section 5.3 for improving the estimation. There are several

[19]Additionally, we prove that the estimators, appropriately normalized, are asymptotically normally distributed.

[20]Theorem 2 does not explicitly include isolates, as we define subgraphs as connected objects for ease of notation. However, the theorem extends easily to this case. In particular, in the case of isolates, 'sparse' actually puts a *lower* bound on the probability of links - so that links are not so sparse as to generate extra isolated nodes.

ways to do finite sample corrections and we discuss them in Section 5.3, but to build intuition here we follow a simple heuristic argument.

The probability a node (that is truly not chosen as an isolate) is isolated in the resulting graph is roughly[21]

$$(1 - p_L - p_T (n - n_I - 2))^{n-n_I-1}.$$

Simply plugging in the estimates of the n_I, p_L and p_T gives us a rough calculation of the probability that any given node would be isolated *incidentally*, and multiplying by the number of nonisolated nodes give us an estimate of the expected number of incidentally generated isolates:

$$(1 - p_L - p_T (n - n_I - 2))^{n-n_I-1} \approx (1 - 0.037 - 0.002 (30 - 2))^{30-1} = .06$$

Thus, the probability that any given node is isolated is approximately $p_I + .06$. To get 20 isolated nodes, $50(\widehat{p}'_I + .06) = 20$, where $\widehat{p}'_I$ is the corrected estimate. This solves to $\widehat{p}'_I = .34$: a correct estimate.

To sum up, we develop two classes of tractable models. One are subgraph generation models (SUGMs) in which we think of subgraphs as being directly generated by subgroups of nodes. The second is a more general statistical exponential random graph model (SERGM), in which a network is drawn based on its properties (e.g., a vector of sufficient statistics such as subgraph counts). We provide theorems on asymptotic estimation of each of these classes of models, and also describe techniques that provide for tractable estimation even with large numbers of nodes in many cases. We then also clarify the relationship between SUGMs and SERGMs, via Theorem 3. Indeed, for sparse networks there is a close correspondence between SUGMs and SERGMs (and ERGMs). As ERGMs are special cases, as a corollary we provide first consistency theorems for those models and show how tractable estimation can be achieved via statistic counts rather than network counts.

We next provide our models and results in their full generality, along with theorems on asymptotic properties of these models and estimators.

3. Definitions

We first present some needed definitions before describing our results.

3.1. **SERGMs.**

The general set of SERGMs that we define is as follows. Consider a vector of network statistics $S = (S_1, \ldots, S_k)$ that takes on values in some set $A \subset \mathbb{R}^k$.[22] A weighting function $K_S : A \to \mathbb{R}$, together with a set of parameters $\beta \in \mathcal{B} \subset \mathbb{R}^k$, define a

[21]There are $n - n_I - 1$ other nodes that it could be linked to, and must end up with none of those links. Roughly the probability that it ends up with a link to any given one of those is $p_L - p_T (n - n_I - 2)$ - the sum of the probability that it ends up with a link, or in a triangle with that node and any of the other $n - n_I - 2$ nodes.

[22]Given the finite number of possible networks, A is taken to be finite. The dimension of A can easily be generalized to be larger than k, as the dimension plays no role in our results. If one wishes to work with weighted networks, then obvious extensions to continuous ranges and integrals apply.

statistical exponential random graph model (SERGM). The associated probability of seeing realized statistics $S = s$ is:

$$\mathrm{P}_{\beta,K_S}(s) = \frac{K_S(s)\exp(\beta \cdot s)}{\sum_{s' \in A} K_S(s')\exp(\beta \cdot s')}. \tag{3.1}$$

This is a model that states that the probability that a network exhibits a specific realization of statistics $S = s$ is given by an exponential function of the statistics s. Thus, the model is based directly on the properties of the network rather than the actual realized network. A distribution on properties $S \in A$ drives the network formation process, and those network properties are what the researcher ultimately cares about with respect to the social or economic theories being studied. Which network forms given the realized statistics is secondary and could be uniform at random, or according to some other conditional distribution, so long as given the realized s a network g such that $S(g) = s$ is drawn. (Unless otherwise stated we take it to be uniform at random.)

It is important to note that node characteristics can also be included in statistics. For example, nodes might be classified into some finite number of groups based on some characteristics, and then one can track the number of links between various types of nodes, various clustering and cohesion measures by types of nodes, and so forth. This permits the fitting of choice-based models, where the utility that an individual derives from a link to another depends on node characteristics and network position.

In the language of exponential families of random variables, $K_S(\cdot)$ is simply a *reference distribution*. Varying the reference distribution, of course, changes the resulting odds of various values of s being drawn and can affect whether the model is consistently estimable.

Recalling that $N_S(s) = |\{\, g \in \mathcal{G}^n : S(g) = s \,\}|$ is the number of graphs that have the same statistic value s; the special case in which $K_S(\cdot) = N_S(\cdot)$ corresponds to a standard ERGM.

Two remarks are in order. First, note that there is no reason to maintain that $K_S(\cdot)$'s must approximate $N_S(\cdot)$'s. Nature may choose properties of networks (S's) according to some alternative weighting. The instance of studying $N_S(\cdot)$-weighted SERGMs may be a historical one: on another planet, people may have first modeled SERGMs with general K's and would see those as natural with the ERGMs being a special case where the weights are specialized to the N_S's.

Second, even if one is interested in a sub-class of these models wherein the $K_S(\cdot)$'s approximate (or are) the $N_S(\cdot)$'s, the statistical representation greatly reduces the dimensionality of the space over which relative likelihoods must be estimated to the point at which practical estimation of SERGMs becomes feasible.

3.2. **Estimation.**

The maximum likelihood estimator $\widehat{\beta} := \operatorname{argmax}_\beta \log\left(\mathrm{P}^n_\beta(s)\right)$ solves

$$\widehat{\beta} = \operatorname*{argmax}_\beta \beta \cdot s - \log\left[\sum_{s' \in A} K_S(s')\exp(\beta \cdot s')\right].$$

It follows that, except for extreme cases, the maximum likelihood estimator satisfies

$$0 = s - \frac{\nabla \sum_{s' \in A} K_S(s')\exp\left(\widehat{\beta} \cdot s'\right)}{\sum_{s' \in A} K_S(s')\exp\left(\widehat{\beta} \cdot s'\right)}.$$

Thus, under regularity conditions such that the SERGM is sufficiently identified (so that $\beta \neq \beta'$ implies that $\mathrm{E}_\beta[S] \neq \mathrm{E}_{\beta'}[S]$), the maximum likelihood estimator $\widehat{\beta}$ of a SERGM of the form (3.1) solves

$$s = \frac{\sum_{s' \in A} K_S(s') \exp\left(\widehat{\beta} \cdot s'\right) s'}{\sum_{s' \in A} K_S(s') \exp\left(\widehat{\beta} \cdot s'\right)} = \mathrm{E}_{\widehat{\beta}}[S]. \tag{3.2}$$

For extreme values of s this will not be well-defined.[23] Here, we implicitly assume that the model is specified so that the probability of observing extreme statistics for which this is not satisfied is negligible, which will be true of the asymptotic specifications that we work with provided that the β's do not tend to extremes too quickly.[24]

Beyond maximum likelihood estimation, we may also be interested in the more general family of generalized method of moments (GMM) estimators. These are standard classes of estimators associated with the first order condition equations normalized by the rate of growth of the associated network statistics S or some other normalization. For example, given some diagonal matrix C with positive diagonal entries relative to which we are interested in the estimator:

$$\widehat{\beta}_n = \arg\min \left(S(g) - \mathrm{E}_{\widehat{\beta}}[s]\right)' C_n \left(S(g) - \mathrm{E}_{\widehat{\beta}}[s]\right). \tag{3.3}$$

We do not explicitly discuss Bayesian estimation, but the conditions that we define here to ensure practical and consistent estimation for MLE and GMM estimators also provide for straightforward extensions to Bayesian estimation with appropriate regularity conditions on priors.

3.3. **Subgraph Generation Models: SUGMs.**

Our approach to defining SERGMs is that the generation process is one based on network properties rather than networks.

One general class of network properties are counts of subgraphs: how many links does a network have, how many triangles, how many cliques of size x, how many star configurations of given sizes, how many isolated nodes, and so forth.

The idea behind a "Subgraph Generation Model," SUGM, is that subgraphs are directly generated by some process. Classic examples of this are Erdos-Renyi random networks in which each link is randomly generated, and the generalization of that model, stochastic-block models, in which links are formed with probabilities based on the nodes' attributes.

The more interesting generalization of those linked-based models to SUGMs is to allow richer subgraphs to form directly, and hence to allow for dependencies in link formation. It is not only links that are generated directly, but also other subgraphs:

[23]For example, for a simple Erdős-Renyi random network where the count statistic is simply the number of links in the network, then if turns out that all links are present so that $s = n(n-1)/2$, then the β that maximizes the likelihood of the ERGM formulation is essentially infinite (the $\beta = \log\left(\frac{p}{1-p}\right)$ corresponding to the maximum likelihood estimator of the link probability ($p = 1$) is not well-defined). For more on the non-existence of well-defined maximum likelihood estimates for extreme networks see Rinaldo et al. (2011).

[24]Parameters can still approach extremes. The requirement here can be fairly weak. For example, if one were counting links it must be that the probability of having absolutely no links (or all links) realized vanishes, which is true even if the probability of a link is larger than $1/n^x$ for some $x < 2$.

triangles, cliques of 4, stars of 5 nodes, etc. That is, groups of agents meet according to some random process possibly related to their attributes, and then decide whether to form a subgraph (with those choices also being potentially dependent on node attributes). So, for instance in an example with links and triangles, two villagers meet at random and decide whether to form a link, with the meeting and link formation probability potentially dependent on their castes (or other characteristics). It might be that people of the same caste meet more frequently or are more likely to form a relationship when they do meet. Similarly, groups of three (or more) randomly meet and can decide whether to form a triangle, with the meeting probability and decision potentially driven by their castes and/or other characteristics. The model can then be described by a list of probabilities, one for each type of subgraph. This results in networks with various distributions of subgraph counts depending on parameters of the model.

As we show in Theorem 3, SUGMs have a representation in a SERGM form, but in some relevant cases SUGMs are easier and more intuitive to work with directly, and so we distinguish them from their SERGM representation.

SUGMs are formally defined as follows. There is a a finite number of different types of nonempty subgraphs, indexed by $\ell \in \{1, \ldots, k\}$, on which the model is based.[25] In particular, a subgraph generation model (SUGM) on n nodes is based on some list of k subgraph types: $(G_\ell^n)_{\ell \in \{1,\ldots,k\}}$ where each G_ℓ^n is a set of possible subgraphs, which are identical to each other (including node covariates) up to the relabeling of nodes.[26]

As an example, the set G_ℓ^n for some ℓ could be all triangles such that two nodes have characteristics X and one node has characteristics X'. These could also be directed subgraphs in the case of a directed network.

A SUGM is then defined by the n nodes, their covariates, a list of subgraph types $(G_\ell^n)_{\ell \in \{1,\ldots,k\}}$, and a list of corresponding parameters $p^n = (p_1^n, \ldots, p_k^n) \in [0,1]^k$ that governing the likelihood that a particular subgraph appears.

A network is randomly formed as follows. First, each of the possible subnetworks in G_1^n is independently formed with a probability p_1^n. Iteratively in $\ell \in \{1, \ldots, k\}$, each of the possible subnetworks in G_ℓ^n that is not a subset of some subgraph that has already formed is independently formed with a probability p_ℓ^n.

We consider two variations of the model. The first, as just defined, is one in which we only keep track of subnetworks in G_2^n that are not already part of a subnetwork in G_1^n that already formed. The other variation is one in which we allow for redundant formation, and simply form subgraphs of each type disregarding the formation of any other subgraphs.

To see the issue, consider the formation of triangles and links. Let G_1^n be a list of all possible triangles and G_2^n be a list of all possible links. First form the triangles with the corresponding probability p_1^n. This then leads to the creation of some of the links in G_2^n. Do we allow those links to also form on their own? Whether we then allow links that are already formed as part of a triangle to form again as links is inconsequential

[25]The definition does not admit isolated nodes as we define subgraphs to be nonempty, but those can also easily be admitted but with notational complications.

[26]Formally, there is a set $\mathcal{H} = \{H_1, ..., H_k\}$ of *representative subgraphs*, possibly depending on node covariates, each having m_ℓ nodes for $\ell = 1, .., k$. Then G_ℓ^n contains all subgraphs that are homomorphic to H_ℓ.

in terms of the network that emerges, and really is an accounting choice. It turns out sometimes to be easier to count subgraphs as if they can form in multiple ways, and at other times it is easier to keep track of smaller subnetworks that form only on their own and not already as part of some larger subnetwork.

When a subgraph g' is generated in the ℓ^{th}-phase, we say that it is *truly generated*. This results in a network g, which is the union of all the truly generated subgraphs. The resulting g can also contain some *incidentally generated* subgraphs that result from combinations of links of unions of truly generated subgraphs, and we provide further definitions concerning this below.

This model differs from a SERGM because the truly generated subnetworks are not directly observed. The actual counts of statistics under the resulting g can differ from the number that were formed directly under the process. Backing out how many of each type of subnetwork was truly generated is important in estimating the true parameters of the model, the p_ℓ^n's, and is something that we discuss at length below.

4. SERGMs and Estimation Techniques

Let us first discuss the estimation of SERGMs. Under what conditions does an estimator of a SERGM converge to the correct estimate in probability as n grows? To our knowledge there are no general results on consistency of ERGM estimators. The primary challenge is that the data consists of a single network and the asymptotics are in terms of the number of nodes, but the relationships are correlated and so the data can be far from independent.

To prove results regarding estimating SERGMs, we consider sequences of SERGMs $(S^n, K_S^n, A^n, \beta_n)$, with $n \to \infty$. We now include notation for the index n since some of our results relate the number of nodes n to the accuracy of the estimation.

4.1. **Count SERGMs.**

We begin by focusing on a natural subclass of SERGMs that we call "count SERGMs", and which have parameters that are consistently as well as easily estimable.

Let $S^n = (S_1^n, \ldots, S_k^n)$ be a k-dimensional vector of network statistics whose ℓ-th entry takes on non-negative integer values with a maximum value $\overline{S}_\ell^n \to \infty$. We call such a SERGM specified with $K^n(s) = \prod_\ell \binom{\overline{S}_\ell^n}{s_\ell}$ a *count SERGM*. Let let $D_n = \text{Diag}\left\{\overline{S}_\ell^n\right\}_{\ell=1}^k$ be the associated normalizing matrix.

In a count SERGM, each statistic can be thought of as counting some aspect of the network: the number of links between nodes of various types, various types of cliques, other subgraphs, the number of pairs of nodes at less than some distance from each other, etc. It includes counts of subgraphs, but also allows for other counts as well (e.g., the number of pairs of nodes at certain distances from each other, as just mentioned; or the number of nodes that have more than a certain degree - so a degree distribution).

Associated with any vector of count statistics S^n on n nodes is a possible range of values. It could be that there are cross restrictions on these values. For example, if we count links S_L^n and isolates S_I^n, then S_L^n cannot exceed $\binom{n-S_I^n}{2}$. In that case the set of possible statistics is a set A^n where

$$A^n = \left\{(s_L, s_I) : s_I \in \{0, 1, \ldots, n\}, s_L \in \left\{0, 1, \ldots, \binom{n - s_I}{2}\right\}\right\}.$$

Given that the set of possible statistics A^n might not be a product space, in estimating count SERGMs, it will be helpful to know whether the realized statistics are likely to be close to having binding restrictions on the cross counts. If a model truly generates a third of its nodes as isolates, and then generates less than half of all possible links, then in a wide band around the expected values, there would be no conflict in the counts.

A sequence of count SERGMs (S^n, K^n, A^n, β^n) have statistics that are *not conflicted* if there exists some $\varepsilon > 0$ such that

$$\prod_{\ell} \left\{ \lfloor \mathrm{E}_{\beta^n_\ell}[S^n_\ell](1-\varepsilon) \rfloor, \ldots, \lfloor \mathrm{E}_{\beta^n}[S^n_\ell](1+\varepsilon) \rfloor \right\} \subset A^n$$

for all large enough n.[27]

The "not conflicted" condition simply asks that at least in some small neighborhood of the unconstrained expected values of the statistics - as if they were each counted completely on their own, they are not conflicted so that they are jointly feasible. Essentially, a local neighborhood of the expected statistics contains a product space. So for example, if one expects one third of the nodes to be isolated, and one out of five links to be present, then with large numbers of nodes, the statistics are unlikely to be in conflict. This condition is quite easy to satisfy.[28]

For models where counts are not conflicted, with a high probability the realized statistics lie in a product subspace, which helps us in proving the following consistency result.

THEOREM **1** (Consistency and Asymptotic Normality of Count SERGMs)**.** *A sequence of count SERGMs that are not conflicted is consistent;* $|\widehat{\beta}^n - \beta^n| \xrightarrow{\mathrm{P}} 0$.

Moreover, if $\exp \beta^n_\ell / (1 + \exp \beta^n_\ell) \cdot \bar{S}^n_\ell \to \infty$ *for every* ℓ*, the parameter estimates are asymptotically normally distributed:*

$$D^{1/2}_{n,\ell} \left(\widehat{\beta}^n_\ell - \beta^n_\ell \right) \rightsquigarrow \mathcal{N} \left(0, \frac{1}{\frac{\exp \beta^n_\ell}{1+\exp \beta^n_\ell} \cdot \left(1 - \frac{\exp \beta^n_\ell}{1+\exp \beta^n_\ell} \right)} \right).$$

Finally, letting $\widehat{p}_\ell = s_\ell / \overline{S}^n_\ell$*, an approximation of the MLE estimator can be found directly as*

$$\widehat{\beta}_\ell := \log \left(\widehat{p}_\ell / (1 - \widehat{p}_\ell) \right) = \log \left(s_\ell / (\overline{S}^n_\ell - s_\ell) \right).$$

The proof of Theorem 1 works via showing that the model can be locally approximated by a product of appropriately defined binomial random variables. In fact those

[27] $\mathrm{E}_{\beta^n_\ell}[S^n_\ell]$ refers to the expectation taken with respect to the one dimensional distribution of S^n_ℓ ignoring the other statistics: i.e., with respect to a SERGM $\frac{K^n_{S_\ell}(s_\ell) \exp(\beta \cdot s_\ell)}{\sum_{s'_\ell \leq \overline{S}^n_\ell} K^n_{S\ell}(s'_\ell) \exp(\beta \cdot s'_\ell)}$. This takes expectations with respect to the unconstrained range of S^n_ℓ rather than cross restrictions imposed under A^n.

[28] There are other things also embodied in the condition, as there are certain counts of statistics that might not be feasible: for instance it is not possible to have a network with only one triangle missing: once one triangle is removed it also removes many others from the network. Thus, the range of some statistics is not a connected (containing all adjacent entries) subset of the integers. Nonetheless, for lower values of triangles, this is not an issue. Generally, in the relatively sparse ranges of networks that are often of empirical interest, this condition can be easily satisfied.

binomial random variables provide a direct estimator for count SERGMs. Our proof shows that following what would seem to be a naive technique of estimation is valid. One can simply estimate parameters p_ℓ as if the subgraphs were generated according to a binomial distribution with a maximum number of possible realizations $\overline{S}_\ell^n$ and s_ℓ as its realization.

It is important to emphasize that count SERGMs still allow for strong interdependencies and correlations in link appearances, both within and across statistics. What our proof takes advantage of is a local approximation of such count SERGM distributions in unconflicted regions.

Theorem 1 tells us that unconflicted count SERGMs form a consistently estimable class whose statistical properties we understand very well.

4.2. **Consistency in General.**

The above results apply to a fairly general class of SERGMs, count SERGMs, for which we can derive explicit asymptotic distributions and simple estimators. We also provide results about consistency for the more full class of SERGMs in Appendix E.

Briefly, there are two sorts of conditions that we outline as being sufficient for consistency (and, effectively, necessary). One is an identification condition that requires that different parameters distinguish themselves with different expected statistics. It is a minimal condition (essentially necessary) since if two different parameter values generate very similar expected statistics, then observing the realized statistic will not allow us to distinguish the parameters. The second condition requires that the (appropriately normalized) statistics concentrate around their means. If the statistics are not concentrated, then even though different parameters lead to different expected statistics, observing a statistic would not allow one to back out the parameters. Various combinations of such conditions (see Appendix E) ensure consistent estimation.

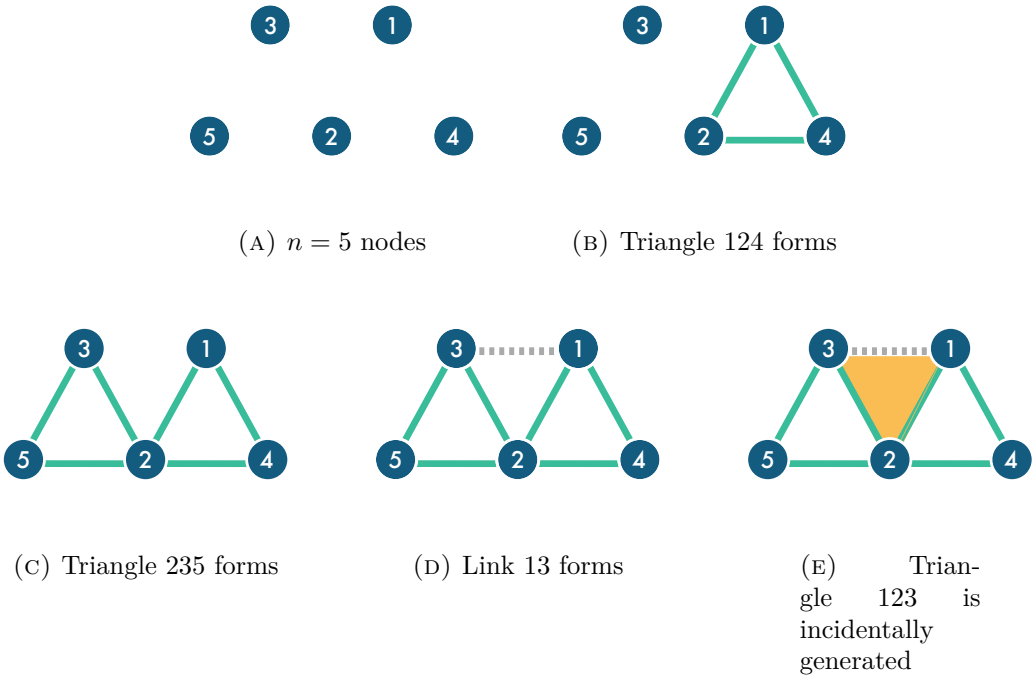
5. Subgraph Generation Models (SUGMs)

Next, we discuss the estimation of SUGMs. The main challenge here is that subnetworks can be incidentally generated: forming links can lead some triangles to form indirectly. Thus, to estimate the actual true generation rates, we need to estimate incidental formation. We take two approaches. One is that we show in large and sparse enough networks, incidental generation does not significantly bias estimation. The second is to provide explicit finite correction methods for estimation in smaller networks where incidentals may be nontrivial, which we return to in Section 5.3.

5.1. **Incidentally Generated Subgraphs.**

To see the issue of incidental subgraph generation in SUGMs, consider the following example. Suppose that the subgraphs in question are triangles and single links, so that $G_1^n(g)$ is the set of all triangles possible among the n nodes, and $G_2^n(g)$ is the set of links on n nodes. The triangle $\{12, 23, 31\}$ could be incidentally generated by the subgraphs g^1, g^2, g^3 where $g^1 = \{12, 24, 41\}$, $g^2 = \{23, 25, 53\}$ and $g^3 = \{31\}$. Figure 2 provides an illustration.

This presents a challenge for estimating a parameter related to triangle formation since some of the triangles that we observe were truly generated in the formation process, and others were "incidentally generated;" and similarly, it presents a challenge



(A) $n = 5$ nodes

(B) Triangle 124 forms

(C) Triangle 235 forms

(D) Link 13 forms

(E) Triangle 123 is incidentally generated

FIGURE 2. An incidentally generated triangle.

to estimating a parameter for link formation since some truly generated links end up as parts of triangles.

The key to our estimation in this section is that in cases where networks are sparse enough, then the fraction of incidentally generated subgraphs compared to truly generated subgraphs is negligible. As many applications will satisfy the sparsity conditions, the estimation techniques are applicable in many cases of interest.

To state results on the estimation of sparse SUGMs, we first need a few definitions.

Consider a sequence of SUGMs indexed by n, each with some k sets of subgraphs that are counted, $G^n = (G^n_1, \ldots G^n_k)$, where k is fixed for the sequence.

We say that the vector of sets of subgraphs $G^n = (G^n_1, \ldots G^n_k)$ is *nicely-ordered* if the subnetworks in G^n_ℓ cannot be a subnetwork of the subnetworks in $G^n_{\ell'}$ for $k \geq \ell' > \ell \geq 1$:

$$g_\ell \in G^n_\ell \text{ and } g_{\ell'} \in G^n_{\ell'} \text{ implies that } g_\ell \not\subset g_{\ell'}.$$

Note that any vector of sets of subgraphs can be nicely ordered: simply order them so that the number of links in the subgraphs are non-increasing in ℓ: so that $\ell' > \ell$ implies that the number of links in a subnetwork of type ℓ' is no more than the number of links in a subnetwork of type ℓ. For example, triangles precede links.

We then follow our accounting convention so that statistics count subgraphs in order and those which are not part of any previous subgraph:
$S^n_{\ell'}(g) = |\{g_{\ell'} \in G^n_{\ell'} : g_{\ell'} \subset g \text{ and } g_{\ell'} \not\subset g_\ell \text{ for any } g_\ell \in G^n_\ell \text{ such that } g_\ell \subset g \text{ for some } \ell < \ell'\}|$.

We now define incidental generation and sparsity.

Consider a realization of a SUGM process in which the truly generated subgraphs are given by $\Gamma \subset \cup_\ell G^n_\ell$, and let g denote the realized network $g = \cup_{g' \in \Gamma} g'$. The researcher observes g and must make some inferences about Γ.

Fix a specific subgraph $g' \subset g$. We say that g' is *incidentally generated* by a subset of the (truly generated) subgraphs $\{g^j\}_{j\in J} \subset \Gamma$, indexed by J, if:

(i) g' was not *truly generated* ($g' \notin \Gamma$),
(ii) $g' \subset \cup_{j\in J} g^j$, and
(iii) there is no $j' \in J$ such that $g' \subset \cup_{j\in J, j\neq j'} g^j$.

Part (ii) states that the subgraph is incidentally generated, and part (iii) of the condition ensures that the set of generating subgraphs is minimal.

Despite minimality, a subgraph could still be generated in multiple ways. For example, in Figure 2e, if the researcher only observes the resulting network, there are various possibilities to be considered: the triangle $\{12, 23, 31\}$ could have been truly generated, it could also have been incidentally generated by the subgraphs g^1, g^2, g^3 where $g^1 = \{12, 24, 41\}$, $g^2 = \{23, 25, 53\}$ and $g^3 = \{31\}$, it could have been incidentally generated by the subgraphs g^1, g^2, g^3 where $g^1 = \{12\}$, $g^2 = \{23\}$ and $g^3 = \{31\}$, and still other possibilities.

5.1.1. *Generating Classes.*

In order to define sparsity, we have to keep track of the various ways in which a subnetwork could have been incidentally generated.

Out of the many ways in which some $g_\ell \in G^n_\ell$ could be incidentally generated, some of them are equivalent up to relabelings. For instance, in a large graph any different combinations of triangles and edges could incidentally generate a triangle $g_\ell = \{12, 23, 31\}$, however there are only eight ways in which it can be done if we ignore the labelings of the nodes outside of g_ℓ: link 12 could be generated either by a triangle or link, and same for links 23 and 31, leading to $2^3 = 8$ ways in which this could happen.

Consider some $g_\ell \in G^n_\ell$ that is incidentally generated by a set of subnetworks $\{g^j\}_{j\in J}$ with associated indices ℓ_j and also by another set $\{g^{j'}\}_{j'\in J'}$. We say that $\{g^j\}_{j\in J}$ and $\{g^{j'}\}_{j'\in J'}$ are equivalent generators of g_ℓ if for each g^j there is $g^{j'}$ such that $\ell_j = \ell_{j'}$ and $g_j \cap g_\ell = g_{j'} \cap g_\ell$. So the generating sets play the same roles in g_ℓ but might involve different nodes outside of $N(g_\ell)$. Note that equivalent sets of generators must have the same cardinality as they must both be minimal and involve the same intersections with g_ℓ.

Given this equivalence relation, there are equivalence classes of generating sets of networks for any g_ℓ. There are at most $\left(\sum_{\ell'=1}^k m_{\ell'}\right)^{m_\ell}$ equivalence classes of (minimal) generating sets for any subnetwork g_ℓ.[29] For each equivalence class J of generating sets of some ℓ, we have some list $(\ell_j, h_j)_{j\in J}$ of the types of subnetworks and the number of nodes that the each subnetwork has intersecting with g_ℓ. We call these the *(minimal) generating classes* of a subgraph g_ℓ and note that these are the same for all members of G^n_ℓ, and so we refer to them as the generating classes of ℓ.

So, for a links and triangles example, where $G^n = (G_T, G_L)$ are triangles and links respectively, there are four generating classes of a triangle: a triangle could be incidentally generated by three other triangles, two triangles and one link, two links and

[29]For each link in g_ℓ there are at most $\sum_{\ell'=1}^k m_{\ell'}$ links that could generate that link out of various subgraphs, and then the power is just the product of this across links in g_ℓ, producing an upper bound.

one triangle, or three links.[30] Here, then we would represent a generating class of two triangles and a link as $(T, 2; T, 2; L, 2)$.

5.1.2. *Relative Sparsity.*

Consider a set of nicely ordered subgraphs $G^n = (G_1^n, \ldots, G_k^n)$ and any $\ell \in \{1, \ldots, k\}$ and any generating class of some ℓ, denoted $J = (\ell_1, h_1; \ldots; \ell_{|J|}, h_{|J|}$. Let[31]

$$M_J = (\sum_{j \in J} h_j) - m_\ell.$$

For example, in forming a triangle from any combination of triangles and links, each $h_j = 2$ and so $M_J = 6 - 3 = 3$.

We say that a sequence of models as defined in Section 3.3 with associated nicely-ordered subgraphs $G^n = (G_1^n, \ldots, G_k^n)$ and parameters $p^n = (p_1^n, \ldots, p_k^n)$ is *relatively sparse* if for each ℓ and associated generating class J with associated $(\ell_j, h_j)_{j \in J}$:

$$\frac{\prod_{j \in J} \mathrm{E}_{p^n}(S_{\ell_j}^n(g))}{n^{M_J} \mathrm{E}_{p^n}(S_\ell^n(g))} \to 0.$$

This is a condition that limits the relative frequency with which subgraphs will be incidentally generated (the numerator) to directly generated (the denominator).

To make this concrete, consider our example with triangles and links. A triangle can be generated by other combinations of links and triangles. The expected number of triangles that nature generates directly is $\mathrm{E}_{p_T}[S_T^n(g)] = p_T \binom{n}{3}$ and the number of links not in triangles is (approximately) $\mathrm{E}_{p_L}[S_L^n(g)] = p_L \left(\binom{n}{2} - O\left(p_T \binom{n}{3}\right)\right)$. Thus it must be that for each generating class,

$$\frac{\prod_j \mathrm{E}_{p^n}(S_{\ell_j}^n(g))}{p_T n^6} \to 0.$$

For the generating class of all triangles, this implies that $p_T^2 n^3 \to 0$, so $p_T = o(n^{-3/2})$. For the generating class of all links, this implies that[32] $p_L^3 / p_T \to 0$, which is the obvious condition that triangles formed by independent links are rare compared to triangles formed directly. This implies that (but is not necessarily implied by) $p_L = o(n^{-1/2})$. The conditions on the remaining generating classes (some links and some triangles) are implied by these ones.

For example, letting $p_T = a(n)/n^2$ and $p_L = b/n$, where $a(n) = o(n^{1/2})$ satisfies the sparsity conditions.[33]

---

[30] Here, our upper bound $\left(\sum_{\ell'=1}^k m_{\ell'}\right)^{m_\ell}$ is 4^3, which is quite loose.

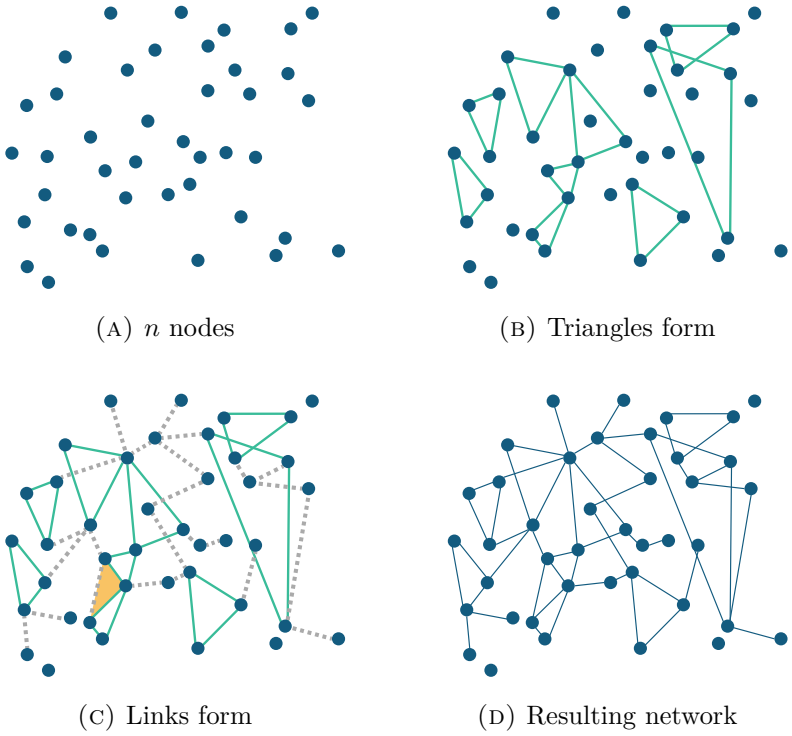
[31] Note that $M_J \geq 1$ since $|J| \geq 2$ and each set of h_j nodes intersects with at least one other set of $h_{j'}$ nodes for some $j' \neq j$. Recall that under the nice ordering, smaller subgraphs cannot be generated as a subset of some single larger one .

[32] Given that $p_T = o(n^{-3/2})$, it follows that $\mathrm{E}_{p_L}[S_L^n] = p_L \left(\binom{n}{2} - O\left(p_T \binom{n}{3}\right)\right)$ is proportional to $p_L n^2$.

[33] This leads to an expected degree of $b + a(n)/3$ and an average clustering of roughly $\frac{a(n)}{6(b+a(n)/3)(b+a(n)/3+1)}$. This can be consistent with various clustering rates, and admits rates of links and triangles found various observed networks. To match very high clustering rates the model can be altered to include cliques of larger sizes.

5.2. **Estimation of Sparse Models.**

Let $\widetilde{S}^n$ denote the vector of the numbers of subnetworks of various types that are truly generated; this is not observed by the researcher since the resulting g may include incidentally generation. Let $S^n(g)$ the observed counts including the incidentally generated subnetworks. In Figure 3, $\widetilde{S}^n_T = 9$ but $S^n_T(g) = 10$ and from observing g there is no way to know exactly what the true $\widetilde{S}^n_T$ is, we just have an upper bound on it. Meanwhile, $\widetilde{S}^n_U = 23$, but as one truly generated link becomes part of an incidentally generated triangle, it follows that $S^n_U = 22$.



(A) n nodes (B) Triangles form

(C) Links form (D) Resulting network

FIGURE 3. The network that is formed on n nodes and eventually observed is shown in panel D. The process can be thought of as first forming triangles form independently with probability p^n_T as in (B), and then forming links independently with probability p^n_L on the remaining part of the graph as in (C). In (C) we see that there is one incidence of an extra triangle generated by this process. In this network we would count $S^n_T(g) = 10$ and $S^n_U(g) = 22$ from (D), while the true process generated $\widetilde{S}^n_T(g) = 9$ and $\widetilde{S}^n_U(g) = 23$.

Nonetheless, as we prove, under the sparsity condition we can accurately estimate the true statistics and thus the true parameters.

To state our next result, we need the following notation. Let $\overline{S}^n_\ell(g)$ be the maximum count of S^n_ℓ that is possible on network g. If we are counting triangles and links not in triangles, then $\overline{S}^n_T(g) = \binom{n}{3}$ and $\overline{S}^n_U(g) = \binom{n}{2} - L_T(g)$ where $L_T(g)$ is the number of

links that are part of triangles in g.[34] Let

$$\widehat{p}_\ell^n(g) = \frac{S_\ell^n(g)}{\overline{S}_\ell^n(g)}. \tag{5.1}$$

So, $\widehat{p}_\ell^n(g)$ is the fraction of possible subgraphs counted by S_ℓ^n that are observed in g out of all of those that could possible exist in g. This is a direct estimate of the parameter p_ℓ^n, as if these subgraphs were each independently generated and not incidentally generated.

Let $D_n = \text{Diag}\left\{\widehat{p}_\ell^n(g) n^{m_\ell}\right\}_{\ell=1}^k$. D_n is a normalizing matrix.

In order to have $\widehat{p}_\ell^n(g)$ be an accurate estimator of $p_\ell^n(g)$ in the limit two things must be true. First, the network must be relatively sparse, which limits the number of incidentally generated subgraphs. And, second, it must be that the potential number of observations of a particular kind of subgraph grows as n grows. This would happen automatically in a sparse network setting if we were simply counting triangles and links not in triangles. However, if nodes have different characteristics (say some demographics), and we are counting triangles and links by node types, then it will also have to be that the number of nodes that have each demographic grows as n grows. If there are never more than 20 nodes with some demographic, then we will never have an accurate estimate of link formation among those nodes.

We say a SUGM is *growing* if the probability that $\widetilde{S}_\ell^n(g) \to \infty$ for each ℓ goes to 1.

THEOREM **2** (Consistency and Asymptotic Normality)**.** *Consider a sequence of growing and relatively sparse SUGMs with associated nicely-ordered subgraph statistics* $S^n = (S_1^n, \ldots, S_k^n)$ *and parameters* $p^n = (p_1^n, \ldots, p_k^n)$. *The estimator* (5.1) *is ratio consistent:*[35] $\frac{\widehat{p}_\ell^n(g)}{p_\ell^n} \xrightarrow{\text{P}} 1$ *for each* ℓ. *Moreover,*[36]

$$D_n^{1/2}\left((\widehat{p}_1^n, ..., \widehat{p}_k^n)' - (p_1^n, ..., p_k^n)'\right) \rightsquigarrow \mathcal{N}(0, I).$$

Theorem 2 states that growing and relatively sparse SUGMs are consistently estimable via easy estimation techniques: ones that are direct and trivially computable.

The proof of the theorem involves showing that under the growing and sparsity conditions the fraction of incidentally generated subnetworks vanishes for each ℓ, and so the observed counts of subnetworks converge to the true ones. Given that these are essentially binomial counts, then, as the second part of the theorem states, a variation on a central limit theorem applies and then normalized errors in parameter estimation are normally distributed, and we know the rates at which the parameters converge to their limits. For inference and tests of significance for single parameter values we note that analytic estimates of the variances are directly computable from the analytic expression of the diagonal of the variance matrix. Of course, more complex inferential procedures and tests can be executed through a standard parametric bootstrap as the model is easily simulated.

[34] In sparse networks, $L_T(g)$ would be vanishing relative to $\binom{n}{2}$ and so could be ignored. And typically, in sparse networks, $\overline{S}_L^n(g)$ will be well approximated by $y_\ell \binom{n}{m_\ell}$, where y_ℓ is the number of possible different subgraphs of type ℓ that can be placed on m_ℓ nodes (e.g., y_ℓ is 1 for a triangle, m for a star on m nodes, etc.).

[35] This, of course, implies consistency. But given that the parameters might be converging to 0, the ratio version is important.

[36] I is the k-dimensional identity matrix.

To make the convergence rates concrete, consider the example with links and triangles and let $p_T = a/n^2$ and $p_L = b/n$. These are well within the bounds that would be needed to satisfy sparsity, but provide an example of a realistic level of sparsity that satisfies our conditions for asymptotic normality. Then one can check the incidental generations for triangles is $o_p(n^{1/2})$, which means that the *fraction* of incidentally generated triangles is $o_p(n^{-1/2})$. Here, the normalization D means that the errors on link estimation will be of order $n^{-1/2}$ and on triangle estimation of order $n^{-3/2}$, and so parameter estimates converge very quickly.

Again, we emphasize that although the estimator here is based on binomial approximations, a SUGM still incorporates interdependencies directly through the subgraphs that are generated. The results make use of the fact that in sparse settings, the picture of interdependencies is clear and are measured by the statistics one-by-one.

5.3. **An Algorithm for Estimating SUGMs without Asymptotic Sparsity.**

As the sparsity results are asymptotic, a natural question to ask is whether there exist finite-sample corrections to help estimation. Moreover, it may be that sparsity is not satisfied at all, and we might still be interested in estimation. Specifically, given that in practice incidental statistics could be generated, albeit with probability tending to zero in a sparse case, it is useful to have techniques for estimating SUGMs to correct for bias of using the observed counts. For instance, Figure 3 showed that while there were $\widetilde{S}_U^n = 23$ truly generated unsupported links and $\widetilde{S}_T^n = 9$ truly generated triangles, the researcher counts $S_U^n = 22$ unsupported links and $S_T^n = 10$ triangles.

We now describe an algorithm. The idea behind the algorithm is that we create a network by randomly building up subgraphs in a way that ends up matching the observed network, and we keep track of how many *truly* generated subgraphs of each type were needed to get to a network that matched the observed statistics.

In order to estimate the truly generated subnetworks of each type, $\widetilde{S}_\ell(g)$, we carefully construct a simulated network g_{sim} and keep track of both its truly generated subgraphs $\widetilde{S}_\ell(g_{sim})$ and its observed subgraphs $S_\ell(g_{sim})$. We construct g_{sim} to have $S_\ell(g_{sim})$ match $S_\ell(g)$ as closely as possible, and then use its true subgraphs $\widetilde{S}_\ell(g_{sim})$ to infer the true subgraphs of g, $\widetilde{S}_\ell(g)$.

Consider a SUGM with nicely-ordered subgraphs indexed by $\ell \in \{1, \ldots, k\}$. The algorithm is described for the case where it is presumed that subgraphs of type k (the smallest subgraph - links in most models) cannot be incidentally generated by other subgraphs.[37]

Algorithm

0. Start with an empty graph g_{sim}^0 and set counts $S_\ell(g_{sim^2}) = 0$ and $\widetilde{S}_\ell(g_{sim}^0) = 0$ for all ℓ.
1. Place $S_k(g)$ subgraphs uniformly at random (these will be links in most models). Call the new network g_{sim}^1. This may generate some incidental subgraphs. Update counts of each $S_\ell(g_{sim}^1)$ and $\widetilde{S}_\ell(g_{sim}^1)$ (with the latter only having truly generated links so far).

[37] If the smallest subgraphs can be generated incidentally (for instance if a model only included triangles and cliques of size 4), then begin the algorithm at step t and treat subgraphs of type k symmetrically with all other subgraphs (so drop the first part of step t).

t. – If $S_k(g_{sim}^{t-1}) < S_k(g)$, then place $S_k(g) - S_k(g_{sim}^{t-1})$ subgraphs down uniformly at random. Call the new network g_{sim}^t and proceed to step $t+1$.
– Otherwise, pick subgraph of type ℓ with the minimal ratio $\widetilde{S}_\ell(g_{sim})/S_\ell(g)$. Add one subgraph of type ℓ uniformly uniformly at random.[38] Call the new network g_{sim}^t and proceed to step $t+1$.
– If $\widetilde{S}_\ell(g_{sim}) \geq S_\ell(g)$ for all ℓ, stop.

- The estimates are $\widehat{p}_\ell = \widetilde{S}_\ell(g_{sim})/\overline{S}_\ell(g)$.

To get the basic intuition behind the algorithm, consider a case with just links and triangles. The algorithm takes advantage of the fact that links can generate triangles, but not the other way around. First the algorithm generates unsupported links up to the number observed in g. This might lead to some triangles, and lowering the number of observed links. The algorithm then tops up the links and keeps doing so until the correct observed number of links are present. If there are fewer triangles than in g, it begins adding triangles one at a time (as they might incidentally generate more). At each step, if the number of links drops below what are in g, then new links are added. It continues until the correct number of links and triangles are obtained. It can never overshoot on links, and may slightly overshoot on triangles, only by the incidentals generated in the last steps.

There are many variations one could consider on the algorithm. For example, if one is conditioning on various covariates, then there might be more than one type of link, and since all types of links cannot be incidentally generated one can "top up" several types of subgraphs and not just k. Thus, in step 1 instead of using just k above, one might also use $k-1$, etc., for however many types of links there are, and similarly for the first part of step t.[39]

There are many other algorithms possible, and more generally a Method of Simulated Moments (MSM) approach could also be taken. For that, one simply searches on a grid of parameters, in each case simulating the SUGM and then picking $\widehat{p}$ as the parameter which minimizes

$$\widehat{p} := \underset{p}{\operatorname{argmin}} \left(S(g) - \mathrm{E}_p\left[S(g^{\mathrm{Sim}})\right]\right)' C \left(S(g) - \mathrm{E}_p\left[S(g^{\mathrm{Sim}})\right]\right).$$

5.4. **The Relation between SUGMs and SERGMs.**

We now show a relationship between SUGMs and SERGMs

Consider a model that is a variation on a SUGM where nature forms various subgraphs with a probability p_ℓ of a given subgraph $g_\ell \in G_\ell^n$ forming. The difference between this model and the SUGMs defined above is mainly in terms of the accounting convention: nature generates various subgraphs without worrying about whether they

[38] Add it uniformly at random out of candidate subgraphs that are not already a subgraph of some existing subgraph of g_{sim}^{t-1}. For instance, if adding a triangle, only consider triangles that are not already a subset of some clique of size 3 or more of the generated network through this step.

[39] To fix ideas, consider a SUGM in which there are two types of triangles and two types of links that are generated, accounting for covariates (as we will use in Section 7.1). For instance, links between pairs of nodes that are 'close' in terms of the characteristics and pairs of nodes that are 'far', and triangles involving nodes that are all 'close' and triangles that involve some nodes that are 'far' from each other. The statistics that we count for a network g are: $S_{T,C}(g)$, $S_{T,F}(g)$,$S_{U,C}(g)$,$S_{U,F}(g)$, where U is for unsupported links and T for triangle, and C is for 'close' and F is for 'far'.

overlap, so it could form a triangle and also form a link that already belongs to that triangle. For instance, in a nicely ordered SUGM, if nature first formed triangles and then links outside of triangles, if the triangle between nodes 1,2, and 3 was formed, then the links 12, 23, 13, would not be added later. In this variation of a SUGM, nature forms links and triangles without caring about overlap, so it might form the triangle 1,2,3 and then also the link 12.

Estimating such models is again not difficult in the sparse case, as there will be sufficient independent observations of different types of subgraphs that each parameter can still be accurately estimated, and the above results extend. This formulation makes it easier to relate to SERGMs.

It is useful to write the probability of a given subgraph $g_\ell \in G_\ell^n$ being generated as taking a logistic form:

$$p_\ell = \frac{\exp(\theta_\ell)}{\exp(\theta_\ell) + 1}, \tag{5.2}$$

where θ_ℓ is some function of ℓ.[40] The formation of a given subgraph is independent of other subgraphs. Again, let $\widetilde{S}_\ell$ denote the count of truly generated subgraphs $g_\ell \in G_\ell^n$.

THEOREM **3** (SERGM Representations of SUGMs)**.** *Suppose that the probability that subgraph of type ℓ forms is given by (5.2). This form of SUGM can be represented in a SERGM form:*

$$\mathrm{P}_\beta^n(\widetilde{S}) = \frac{K^n(\widetilde{S}) \exp\left(\widetilde{S} \cdot \theta\right)}{\sum_{s'} K^n(s') \exp\left(s' \cdot \theta\right)}, \tag{5.3}$$

where $K_\ell^n(s_\ell) = \binom{\widetilde{S}_\ell^n}{s_\ell}$ *and* $K^n(s) = (\prod_\ell K_\ell^n(s_\ell))$.

Theorem 3 provides a relationship between SUGMs and SERGMs. The two models are closely related, although the statistics counted here are all of the *actual subgraphs* (including overlaps) that nature generated, which can be estimated but not precisely known.

Note that this also provides a reason why in specifying SERGMs it is useful to have K's that differ from the N's that correspond to some ERGM model. Here specific K's are natural and yet differ from an ERGM formulation.

A direct implication of Theorem 3 is the following, which provides a general result on dynamic processes of network formation, where subgraphs are repeatedly considered and added and deleted over time.

COROLLARY **1.** *Consider any dynamic process such that with probability one, each subgraph is considered infinitely often, and when a subgraph is considered it is added with probability (5.2) if not already present and deleted with the complementary probability if it is already present. The resulting dynamic process has a steady state distribution given by (5.3).*

[40] We re-emphasize that through the indexing of ℓ we can *encode the covariates of the subgraph* $X(g_\ell)$.

6. Extensions

6.1. **Strategic/Preference-Based Random Network Models.**

As we have discussed above, SERGMs and SUGMs admit models where both choice and chance are important, and we describe a couple of examples to illustrate how preferences of individuals over networks can be incorporated.

6.1.1. *Mutual Consent Formation Models.*

Here we describe a strategic network model that harnesses some of the power of Theorem 2. The key aspect of the model is that decisions to link are not only bilateral but instead multilateral: sub-groups of individuals decide whether or not to form subgraphs. A pair of individuals may meet and decide (mutually) as to whether to add a single link, but also a group of three (or more) may meet and decide whether to form a some subgraph such as a clique or some other form (e.g., a ring, star, etc.).[41] Moreover, the probability that they form the subgraph could depend upon the characteristics of the individuals involved.

Consider subgraphs $g_\ell \in G^n_\ell$ with associated individuals $N(g_\ell)$.

The members of $N(g_\ell)$ meet according to a random process and have the opportunity to form g_ℓ. Both the probability with which the members meet and their preferences for forming g_ℓ can depend on their characteristics $X(g_\ell) = (X_i)_{i\in N(g_\ell)}$.

There are certain aspects of the members' characteristics, $H_i(X(g_\ell))$, that affect i's benefits from the subgraph.[42]

There is a probability π_ℓ that a subgraph g_ℓ of individuals with characteristics $X(g_\ell)$ meets and decides whether to form g_ℓ. So it might be more likely that individuals of similar ages meet than ones with different ages.

Individual i obtains a utility[43]

$$U_{i,\ell}(X(g_\ell)) = \gamma_{\ell,X_i} H_i(X(g_\ell)) - \epsilon_{i,\ell}$$

from the formation of a given subnetwork g_ℓ, where γ_{ℓ,X_i} depends on the subnetwork in question and possibly on the characteristics of i and $\epsilon_{i,\ell}$ is a random idiosyncratic term.

The subnetwork then forms conditional upon it having met if and only if $U_{i,\ell}(X(g_\ell)) \geq 0$ for all i (say with at least one strictly positive).[44] If the error term has an atomless distribution, then the strictness is inconsequential. Let $F_\ell(\cdot)$ be the distribution of

[41]For additional theoretical underpinnings of coalition-based network formation models see Jackson and van den Nouweland (2005); Caulier et al. (2013).

[42] For instance if $X(g_\ell)$ were a list of the individuals' ages, then it might be that i's benefit from the subgraph is a function of i's distance from the average characteristics: $h_i(X(g_\ell)) = \left\| X_i - \sum_{j\neq i} X_j/(m_\ell - 1) \right\|$. It could also be that i benefits from the maximum value of X_{-i}, or suffers from variation in the characteristics. h can be tailored to the specific application and list characteristics.

[43]Here we simplify notation by omitting the dependence of the utility on a given individual's position in the subnetwork. Everything stated here extends directly allowing utility to depend on position: for instance, getting higher utility from being the center of a star rather than on its periphery, but the notation becomes cumbersome.

[44] This then corresponds nests pairwise stability as defined by Jackson and Wolinsky (1996), subject to the meeting process. One can adjust this to take into account other rules for group formation, and this also easily handles directed networks.

error terms for the formation of subgraphs in G_ℓ^n. So, the probability that a subgraph $g_\ell \in G_\ell^n$ with characteristics $X(g_\ell) = (X_i)_{i \in N(g_\ell)}$ is formed is

$$p_\ell = \pi_\ell \prod_{i \in N(g_\ell)} F_\ell \left(\gamma_{\ell, X_i} H_i(X(g_\ell)) \right). \tag{6.1}$$

Let $S_\ell(g)$ denote the number of subnetworks in G_ℓ^n (consisting of individuals with characteristics $X(g_\ell)$) that form in a network g, counted excluding networks already counted as subset of some $\ell' < \ell$, as under the well-ordered condition. Under the conditions of Theorem 2, the SUGM in (6.1) is then easily estimated and consistent.[45]

It is important to note that such a formulation allows us to do welfare computations and changes in welfare due to changes in, say, the distribution of X if they include parameters that a policymaker may control - or it may be that a policymaker could change the π function in some well-defined way by say, subsidizing interactions among groups with certain sorts of characteristics who might tend to meet infrequently.

6.1.2. *Strategic Network Formation and Potential Functions.*

There is a nice connection between strategic network formation models and potential functions that spans a series of papers: Jackson and Watts (2001), Butts (2009), Mele (2011), Badev (2013). For example, Butts (2009) and Mele (2011) show that if links are recognized independently over time, and then added or deleted based on individual choices according to a logistic function, then the steady-state distribution can be represented as an ERGM. In those models, only one agent makes a decision at a time and links must be directed. Here we generalize the set of models that are covered, and also extend to allow for mutual consent. We also provide a directed version of the formation model which generalizes the results of Mele (2011).

Agent i's payoff from network g can depend on which subgraphs i is a member of, as well as things such as to whom his or her neighbors are connected.

Let $\mathcal{G}$ denote some set of subgraphs from which agents derive utility.

Consider some agent i and a subgraph $g_\ell \in \mathcal{G}$ that i is a member of, $i \in g_\ell$. Let the members of g_ℓ (including i) have a vector of characteristics described by X_ℓ. Agent i gets some marginal payoff,

$$v(g_\ell, X_\ell),$$

from having this subgraph in the network where this function can depend on the type of subgraph and the characteristics of all the agents involved in the subgraph.

Agent i's utility from a network g is then

$$u_i(g) = \sum_{g_\ell \in \mathcal{G}, i \in g_\ell} v(g_\ell, X_\ell). \tag{6.2}$$

[45] Although the probabilities of various subgraphs are directly estimable (and hence identified) under the conditions of Theorem 2, of course whether the various parts of $\pi_\ell \prod_{i \in N(g_\ell)} F_\ell \left(\gamma_\ell H_i(X(g_\ell)) \right)$ are well identified depends on the specifics of the the functional forms involved. Just as an example, consider a situation with two types. Set $X_i = 1$ if i is of type 1 and $X_i = 2$ if i is of type 2 and $H_i(X) = X_i - X_j$ and consider $m = 2$. Then γ_2 and $-\gamma_2$ both lead to the same value of $F_2 \left(\gamma_2 (X_i - X_j) \right) \times F_2 \left(\gamma_2 (X_j - X_i) \right)$. So here it could not be judged whether the type 2 has a greater expected utility (net of the random term) from the match than the type 1 or whether it is the reverse. There are some obvious simplified formulations that allow for identification, for example setting instead $h_i(X) = |X_i - X_j|$. It might also require specifying a (nonlinear) functional form for π_ℓ as in Currarini et al. (2010).

Note that this allows the agent's utility to depend on the presence of "friends of friends" by including subgraphs of the form $g_\ell = \{ij, jk\}$. Of course, it also allows agents' payoffs to depend on direct links, cliques, and so forth.

Next, consider a network formation process where agents can form links in pairs and they add the link whenever their *mutual* gain is positive. The idea is that they can bargain and make side payments (either in cash or by exchange of favors) to add links whenever those links are mutually beneficial.

In particular, a network g is to be *pairwise stable with transfers* if: [46]

- $ij \in g$ implies that $u_i(g) + u_j(g) \geq u_i(g - ij) + u_j(g - ij)$, and
- $ij \notin g$ implies that $u_i(g) + u_j(g) \geq u_i(g + ij) + u_j(g + ij)$.

For this setting, we can then define the following function f, which is a potential function for network formation under pairwise stability with transfers:

$$f(g) = \sum_{g_\ell \subset g} 2v(g_\ell, X_\ell). \tag{6.3}$$

It follows that $f(\cdot)$ is a potential function for a network formation game that follows pairwise stable with transfers. In particular, direct calculations show that for any g and $ij \in g$:

$$f(g) - f(g - ij) = (u_i(g) + u_j(g)) - (u_i(g - ij) + u_j(g - ij)). \tag{6.4}$$

Thus, the difference between the value that f assigns to g and what it assigns to $g - ij$ is exactly the sum of the differences that i and j assign to the two networks.

For any such setting, Theorem 1 in Jackson and Watts (2001) implies that there exists at least one network that is pairwise stable with transfers, and moreover that there are no cycles in the improving paths.[47]

Now let us describe a dynamic process of network formation. Let g^0 be some starting network, and let g^t denote the network in place at the end of time t. Let g^t_{-ij} denote the network of links other than ij. In each period, there exists some positive probability of each given link being recognized (the two agents in question "meet"). The recognition probabilities can depend on the pair in question and the network in place at the time and the probability that link ij is recognized conditional on the network in place g^t is denoted $p(ij, g^t_{-ij})$.[48]

We emphasize that the meeting process is quite general as it is allowed to depend on the attributes of the agents i and j as well as the network in question. Thus, for example, it allows their meeting probability to depend on whether or not the pair have friends in common, and can even depend on how many friends in common they

[46]This definition is from Bloch and Jackson (2006) and is related to the definition of pairwise stability allowing for side payments that appears in the conclusion of Jackson and Wolinsky (1996).

[47]An improving path is a sequences of networks that differ from each other by one link such that if a link is added or deleted then the pair of agents in the link see an increase in their summed utilities from the change.

[48]Since each link probability could depend on the network other than the link ij, if each link's recognition probability does not exceed $1/\binom{n}{2}$ then the sum of all link recognition probabilities does not exceed 1, and that leave the residual probability $1 - \sum_{ij} p(ij, g^t_{-ij})$ to be the probability that there is no link recognized in the current period and the period simply advances. The scaling of probabilities is irrelevant to the steady state of the process, and so it is fine to allow periods to pass without any recognition.

have, and can depend on any other aspect of the network, as well as the agents' characteristics.

Once recognized at some time t, i and j decide whether to add or delete the link, conditional upon the rest of the network in place at that time g^{t-1}_{-ij}. The probability that the link is added/kept is a logistic function of the mutual value of the link:[49]

$$\frac{\exp\left(u_i(g^t_{-ij}+ij)+u_j(g^t_{-ij}+ij)\right)}{\exp\left(u_i(g^t_{-ij}+ij)+u_j(g^t_{-ij}+ij)\right)+\exp\left(u_i(g^t_{-ij}-ij)+u_j(g^t_{-ij}-ij)\right)}. \tag{6.5}$$

This defines an aperiodic and irreducible Markov chain over the space of all networks, and so it has a unique steady-state distribution. Moreover, it is a reversible Markov chain and the unique steady-state distribution is given by[50]

$$\mathrm{P}(g)=\frac{\exp(f(g))}{\sum_{g'}\exp(f(g'))}.$$

Thus, this is an ERGM:

$$\mathrm{P}(g)=\frac{\exp\left(\sum_{g_\ell\subset g}2v(g_\ell,X_\ell)\right)}{\sum_{g'}\exp\left(\sum_{g_\ell\subset g'}2v(g_\ell,X_\ell)\right)}.$$

We can then rewrite $\sum_{g_\ell\subset g}2v(g_\ell,X_\ell)$ as a function that simply keeps track of statistics of how many subgraphs of a network g are of a given form, (ℓ,X_ℓ), denoted $S_{\ell,X_\ell}(g)$. This then has a SERGM representation:

$$\mathrm{P}(s)=\frac{\exp\left(\sum_\ell s_{\ell,X_\ell}2v_{\ell,X_\ell}\right)}{\sum_{s'}\exp\left(\sum_\ell s'_{\ell,X_\ell}2v_{\ell,X_\ell}\right)}.$$

We remark that link recognition probabilities do not enter the final steady-state distribution, which is only determined by the preferences as captured via the v functions.

6.1.3. *Directed Network Formation.*

Everything stated above has an analog for directed links ij where the decision to add the link is taken by agent i (with stability defined by Nash equilibrium), and where the subgraphs, g_ℓ's, are directed. The only change is to drop the '2' in the above formulas and require that each agent obtain utility from each subgraph in which they direct some link.[51] The directed version of the above generalizes a result in Mele (2011).

[49]With a slight abuse of notation, we allow $g^t_{-ij}+ij$ to denote the network where ij is present and the network of other links is described by g^t_{-ij}, and similarly $g^t_{-ij}-ij$ denotes the network where ij is not present and the network of other links is described by g^t_{-ij}.

[50]We omit the standard proof as, for instance, it is a direct extension of the proof of Theorem 1 in Mele (2011), noting that the link recognition probability can depend on g^t_{-ij} without affecting the steps of his proof.

[51]The model can be specified to allow agents to derive utility from subgraphs in which they have some "in-links" but no "out-links", but can also allow them not to.

6.1.4. *Search Intensity Models.*

Another interesting class of strategic/random network formation models that we can extend to the setting here are where agents face overall costs of forming relationships - not just costs associated with various subgraphs (as in the models above). To account for such overall tradeoffs in the network formation processes, we can also include search intensities as have been analyzed in various formation models such as Currarini et al. (2009, 2010); Borgs et al. (2010); Golub and Livne (2010). Those models are of bilateral link formation; but are easily extended to more general SUGMs as we briefly describe.

Each agent i with characteristics X_i puts in a search effort $e(X_i, m, X) \in [0, 1]$ to form cliques of size m with characteristics X. $m = 2$ indicates links, and so $e(X_i, 2, (X_i, X_j))$ is the effort that agent i expends in trying to form links with agents who have characteristics X_j; and $e(X_i, 3, (X_i, X_j, X_h))$ is the effort that agent i expends in trying to form triangles where the other two agents have characteristics X_j and X_h, and so forth. "Effort" is simply a shorthand for either the time spent socializing in various ways, or else it could simply indicate a relative openness to forming relationships of various types.

An agent obtains a utility

$$u(X_i, m, X)$$

from being of type X_i in a clique of size m with characteristics X.

The probability that a given clique Cl_m forms depends on the vector of efforts for such cliques of those in the clique, $(e_j(, m, X))_{j \in Cl_m}$, according to a function $\pi_{m,X}((e_j(m, X))_{j \in Cl_m})$ that is nondecreasing in each of its arguments.

An agent also pays a cost of network formation: $c(X_i, (e_i(m, X))_{m,X})$ that depends on his or her characteristics X_i and the search efforts that he or she exerts in forming various links and cliques, $(e_i(, m, X))_{m,X}$.

Thus, an agent i's overall expected payoff as a function of the all of the agents' efforts is described by

$$\left(\sum_{m, Cl_m : i \in Cl_m} \pi_{m,X}((e_j(m, X))_{j \in Cl_m}) u(X_i, m, X) \right) - c(X_i, (e_i(m, X))_{m,X})$$

In a case where the u's are nonnegative, this defines a supermodular game: agent i's change in payoff from increasing any dimension of $(e_i(m, X))_{m,X}$ is nondecreasing in the vector of strategies $(e_j(m, X))_{j \neq i, m, X}$. In such games, pure strategy equilibria exist and form a complete lattice (e.g., see Topkis (2001)). Additional conditions on π, u, and c can ensure uniqueness of equilibrium, depending on the specific functional forms that are used to parameterize the model, or one can appeal to equilibrium selection.[52]

Models of this structure thus define SUGMs, where the relative frequencies $p_{m,X}$ of cliques of size m consisting of agents with characteristics described by the profile X. Specifying functional forms for π, u and c then allows for estimation of parameters of the model and of the equilibrium, provided the specification is tight enough to be well-identified.

Although the above formulation is described for cliques, it is easily adjusted for any subgraphs (for instance an agent may value being the center of a star with m agents).

[52]Here there are positive spillovers/externalities from strategies, and so generally the maximal equilibrium will Pareto dominate the others, and so a standard refinement would be to look at the Pareto efficient equilibrium which is then unique and pure (e.g., see Vives (2007) for some background).

In the obvious extension one needs to keep track of the positions of the various types of agents in the subgraph as there are then asymmetries in positions and, for instance, agents might care about the characteristics of the agent at the center of a star.

6.2. **Noise and Almost-Cliques.**

Although observed social networks often exhibit significant clustering and numbers of triangles significantly above what would be observed with independent link formation, typically full cliques of larger sizes are rarer. One reason for a failure to see completed large cliques is that small amounts of measurement error makes a clique exponentially (in the number of links in the clique) less likely to be observed. We discuss a correction for this.

Consider an example in which there is a probability $\varepsilon > 0$ that a data set fails to exhibit any given link (independently across links) that is truly present. A clique of ℓ nodes has $\ell(\ell-1)/2$ possible links. If it were present, then it would be fully observed only with a probability that all of its links are observed, or $(1-\varepsilon)^{\ell(\ell-1)/2}$. For example, suppose that $\varepsilon = .1$. Then the probability that a clique of size 3 is observed without any missing links due to measurement error is .73, while this drops to .53 for a clique of size 4, to .35 for a clique of size 5, and to .21 for a clique of size 6.

Why is this an important issue? Suppose that the true model generates cliques of size 3 and 4 in addition to links. The 27 percent of triangles that are missed due to measurement error, will end up classified as two or fewer links, thus biasing downwards the parameter on triangles and upwards that on links. The 47 percent of cliques of size 4 that are missed, will generally contribute to increased observations of triangles and links. For instance, deleting one link from a clique of size 4 makes it appear as two cliques of size three. Taking two links out makes it either lead to a triangle plus a link, or four extra links. This can substantially bias upwards the counts of triangles and links.

There are two ways to deal with this, one more precise and the other easier but less precise. The precise way to do this, is to model the error and include it in the specification of a SERGM or SUGM. For example, given a SUGM with links, triangles and cliques of size 4 together with probability ε of missing a link in the data, then the probability of seeing a clique of size 4 becomes $p_4(1-\varepsilon)^6$. The probability of seeing a clique of size 4 less one link (under the sparsity conditions) becomes $p_4 6(1-\varepsilon)^5\varepsilon$, and the probability of seeing a clique of size 4 less two links (again under the sparsity conditions) becomes $p_4 15(1-\varepsilon)^4\varepsilon^2$, and so forth. Thus, one can then consider the following count statistics: links, two-stars, triangles, cliques of size 4 less two links, cliques of size 4 less one link, and cliques of size 4, etc. Each of these has a well defined probability given the original specified p_2, p_3, p_4 and the ε. One can then estimate these parameters (including the ε) using the SUGM techniques from our theorem.[53]

A less precise way to do correct for incomplete cliques is to simply count any structure on ℓ nodes that has at least $x_\ell \ell(\ell-1)/2$ links as a clique of size ℓ, where $x_\ell \leq 1$ is some adjustment factor. While a crude adjustment, this can still improve over ignoring the issue altogether. For example, in a setting with links, triangles and 4-cliques, one could make an adjustment by counting any configuration on four nodes with at least

[53]One then has to count these in a well-ordered way, so that triangles that are part of a 4-clique less some links are not counted.

4 links as a 4-clique, but then stick with counting triangles only if all three links are present and they are not part of any modified 4 clique.

While measurement error is one possible way in which a clique could form with missing links, another possibility is that the formation process is such that agents choose which relationships to form, but have payoffs to cliques that allow them to benefit from fewer than all links being present. Then allowing some noise in the formation (either utility-based or trembles) would lead to some links in cliques not forming. While different in interpretation, the techniques that account for such missing links would be similar to that described above: simply directly accounting for the probabilities that various subgraphs form. As SERGMs and SUGMs can already admit arbitrary subgraphs as statistics, this would simply place additional restrictions on the relationships between the probabilities of various subgraphs, but would otherwise use similar estimation techniques as outlined in our results.

6.3. **Measurement Error and Sampling.** Suppose that a researcher only samples half of the nodes of a given network. In using a SUGM, the probabilities of forming various subgraphs project to the subset without any adjustment. That is, the probability of forming a triangle on three nodes in the subsample is the same as the probability of forming a triangle on three nodes in the overall population (accounting for characteristics appropriately). The same is true for a count-SERGM. Thus, parameter estimates of probabilities of subgraphs can be obtained by examining those on subgraphs without making any adjustments. From these estimates the researcher can easily estimate (directly) the expected statistics that should appear in the overall network, and so recreate missing data.[54]

6.4. **Continuous Covariates.**

For ease of exposition, we have focused on models in which covariates are captured by indexing subgraphs by covariates. This encompasses covariates that take on a finite set of values or are approximated by a finite set of values, and is a flexible approach, although it may not work as well with fully continuous data that take on a wide range of values.

Such continuous covariates can also easily be handled, as our models and results have natural extensions to continuous covariates. Let us briefly discuss these extensions here and refer the interested reader to Appendix C for full details.

We discuss the SUGM extension. Let node i be associated with a covariate vector X_i that lies in a compact subset of $\mathbb{R}^d$. Let the probability that a given subnetwork $g_\ell \in G_\ell$ forms be a function $p^n_\ell(X_\ell; \gamma)$ of the vector of node covariates, where γ is some vector of parameters.

Estimating the parameters γ depends on the functional form of $p^n_\ell(x_\ell; \gamma)$. It could take many forms, such as a linear probability model, a logistic form, etc. Consistency and asymptotic normality of the estimators depend on the rate at which γ tends to extremes – thereby affecting the probabilities of various subgraphs and their dependence on covariate values. We provide some sufficient conditions for consistency and asymptotic normality of the estimators in Appendix C.

[54]See Chandrasekhar and Lewis (2013) for further discussion on using this approach to deal with sampled network data.

7. Illustrative Empirical Applications

To illustrate the models we provide a few applications.

7.1. **Network properties generated by SUGMs.**

Our first illustration is to compare a simple model that estimates linking probabilities based on node characteristics (caste and geography) with a SUGM that also includes triangles. The idea is to compare how well these replicate various features of actual networks, including features of the networks such as clustering, the size of the giant component, average path length, degree distributions, and various eigenvalue properties of the adjacency matrices.

For this exercise we use the Banerjee et al. (2013) data consisting of networks in 75 Indian villages. Here we focus on "advice" networks: where an edge represents whether a household speaks to another household when having to make an important decision. This is a simple representation of the informational network structure within the sample of villages, and the networks are reasonably connected (with more than two-thirds of the nodes being in a giant component) and yet also reasonably sparse for small networks.

In addition to the average degree and clustering (which are at least partly captured by links and triangles), we are interested in other quantities motivated by theory. We look at the first eigenvalue of the adjacency matrix, which is a measure of diffusiveness of a network under a percolation process (e.g., Bollobás et al., Jackson (2008)). A related quantity is the spectral gap, which is the difference in the magnitudes of the first and second eigenvalues of the adjacency matrix. This is intimately related to the expansiveness of the network. We are also interested in the second eigenvalue of the stochasticized adjacency matrix. This is a quantity that is key in local average learning processes and modulates the time to consensus (DeMarzo et al. (2003), Diaconis and Freedman (1981), Golub and Jackson (2012)). Additionally, we look at the fraction of nodes that belong to the giant component of the network, as empirical networks are often not completely connected. Finally, we consider average path length (in the largest component).

Our procedure is as follows. For every village, we estimate each of two network formation models. The first network formation model is a link-based model where the probabilities can also depend on geographic and caste covariates. In particular, pairs of household are categorized as either being "close" or "far" and then separate probabilities of links are estimated for "close" and "far" pairs. "Close" refers to pairs of nodes that are of the same caste and are below the median geographic distance (the median GPS distance taken across all pairs of households), and "far" to those that either differ in caste or are further than the median distance. The second network formation model is a SUGM with the same structure except for the addition of triangles.[55] We estimate parameters for the village network for each model and then generate a random network from the model based on the estimated parameters. We do 100 such simulations

[55]Similarly, we categorize triangles as being "close" if all nodes are of the same caste and all pairs are below the median distance, and "far" otherwise.

for each of the 36[56] villages and for each of the two models. We then compare the aforementioned network characteristics from the simulations with the actual data.

Table 1 presents the results. We find that networks simulated from the SUGM better match the structural properties exhibited by the empirical Indian village networks than those simulated from a link-based model.

TABLE 1. Network Properties

		Data	Link-based model with covariates	SUGM with links and triangles	SUGM with isolates, links and triangles
		[1]	[2]	[3]	[4]
Models are fit to different combinations of these statistics.	Number of Unsupported Links	160.8	236.2	161.2	161.8
	Number of Triangles	39.2	3.1	39.7	39.5
	Average Degree	2.3243	2.3260	2.5916	2.5219
	Number of Isolates	54.9722	25.7222	31.4444	65.9167
None of the models are directly fit to any of these statistics.	Average Clustering	0.0895	0.0105	0.1268	0.0829
	Fraction in Giant Component	0.7061	0.8315	0.7982	0.6718
	First Eigenvalue	5.5446	3.8578	4.6762	5.3025
	Spectral Gap	0.9550	0.3354	0.6684	1.0617
	Second Eigenvalue of Stochastized Matrix	0.9573	0.9632	0.9559	0.9069
	Average Path Length	4.6921	5.6565	5.1215	4.1180

Notes: Column [1] presents the average value of various network characteristics across the 36 villages. Columns [2], [3] and [4] present simulation results. In a simulation we first estimate parameters of a given model for a given village and then randomly draw a graph from the model with the estimated parameters. We run 100 simulations for each of the villages for each of the models and average across the simulations, and the entries report these averaged across the villages.

Both the SUGM and the link-based model do quite well for average degree. As expected, the SUGM matches the triangle count and the unsupported link count (as these are the statistics on which the model is based) whereas the link-based model matches average degree quite closely (as this is the moment on which this model is based).

Neither model is based on the remaining statistics. The first and most obvious thing to note is that the link-based model does extremely poorly when it comes to matching clustering while the SUGM does much better, which is natural given that the SUGM explicitly includes triangles. More interestingly, conditioning on the triangles in the SUGM is enough to deliver better matches on all of the other dimensions. For instance, the link-based model considerably underestimates the first eigenvalue (3.86 as compared to 5.54), whereas the SUGM performs better (4.68). Similarly, the link-based model underestimates the expansiveness of the networks with a spectral gap of 0.34 instead of 0.96. The SUGM again performs considerably better (0.67). These sorts of results also hold true for the average path length, fraction of nodes in the giant component, and the second eigenvalue of the stochasticized matrix.

Beyond these two models, we also fit a SUGM that includes isolates, in addition to links and triangles. Not surprisingly, it fits isolates much better than either of the previous models. The more interesting aspects are in the other features to which none of the models are fit. Here we see that including isolates significantly improves, beyond the improvement from triangles, the fits on clustering, the size of the giant component, the first eigenvalue, and spectral gap. Accounting for isolated nodes changes the density among remaining nodes in ways that better match the overall structure of the network. The dimension on which it does not perform as well is that it worsens the fit on the second eigenvalue (the homophily measure). However, that is likely because the

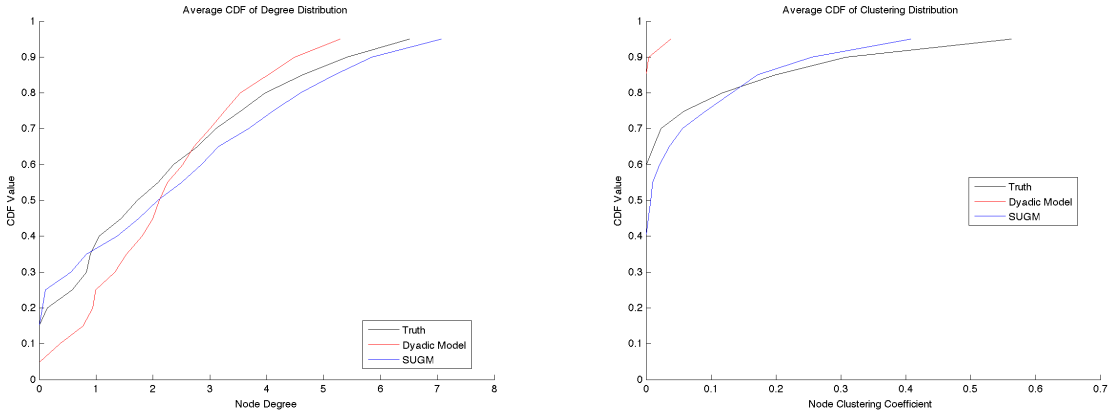
[56]Because we have both complete GPS and caste data for only 36 villages, we restrict attention to these in our analysis.

model is not sufficiently geared towards the covariates that affect segregation, and so densifying the remaining network reduces segregation. Including a richer set of covariates into the model would help counter-act that, but is beyond our illustrative purposes here.

We also examine distributional outcomes. In Figure 4, we show CDFs of node degrees and clustering. The CDFs from the empirical data are computed as follows. For every village, we compute the degree and clustering coefficient for each 5th percentile from 5 to 95. We then average these values across the villages in our sample. The simulated CDFs are computed by taking the analogous cross-village average from simulated data as described in Table 1. For parsimony, we compare only the isolates-links-triangles SUGM and the links-based model.

Figure 4a shows the degree distributions. The SUGM does considerably better than the links-based model in matching the entire degree distribution. Specifically, the links-based model undershoots both the lower and upper tails of the degree distribution, despite hitting the average correctly. The SUGM, though slightly overshooting the average degree, better matches the distribution overall.

Figure 4b shows the distribution of clustering coefficients. The link-based model is unable to generate any non-trivial clustering and essentially has a degenerate distribution (the short red curve in the upper left). The SUGM generates a distribution similar to the data, significantly outperforming the link-based model.



(A) Average CDF of Degree

(B) Average CDF of Clustering Coefficient

FIGURE 4. Distributions of degree and clustering coefficients - averaged across the 75 villages. The figure displays the CDFs from the data (grey), the isolates-links-triangles SUGM (blue), and the link-based model (red).

The results of the analysis in this section are not sensitive to the covariates included. That is, it is not simply that the SUGM allows for more parameters that enable it to better match the data. It is that it includes richer network structures. In Appendix F, we enrich the links-based model to include polynomials of a large set of demographic covariates including geographic distance, caste composition, quality of access to electricity, quality of latrines in the household, number of beds, number of rooms, etc. We show that the links-based model, even aided by a considerable amount of data and

more degrees of freedom, cannot replicate structural features of the network that are captured by very simple SUGMs that rely on minimal amounts of covariate data.

It is perhaps not surprising that SUGMs do a much better job at recreating network structures that standard link-based models, but nonetheless it is important. Moreover, the fact that the SUGMs do a better job than a link-based model of recreating not only local clustering and triangle patterns but also many other features of the real networks that it is not based upon suggests that there is substantial value added of modeling the formation of triangles and isolates.

7.2. **Links across social boundaries.**

Individuals are associated with groups and identities that can lead to strong social norms about interactions across groups. For instance, in much of India there are still strong forces that influence if and when individuals form relationships across castes.

Here we further illustrate our models in answering the question posed at the beginning of this paper: Are people significantly more likely to form cross-caste relationships when those links are unsupported (without any friends in common) compared to when those links are supported with at least one friend in common? The SERGM/SUGM statistical framework allows us to look at whether individuals have significantly higher ratios of cross caste relationships over within-caste relationships when those relationships are unsupported compared to when they are supported. The idea is that cliques of three more more may dictate greater adherence to a group norm which individuals are able to avoid in isolated bilateral relationships.

To analyze this, we examine data from the 75 Indian villages mentioned above. We work with two caste categories: the first consists of people in scheduled castes and scheduled tribes and the second consists of those people in any other caste. Scheduled castes and scheduled tribes are those defined by the Indian government as being disadvantaged. This is a fundamental caste distinction over which the strongest cultural forces are likely to focus. Additional norms are at work with finer caste distinctions, but those norms are more varied depending on the particular castes in question while this provides for a clear caste barrier.

The SUGM that we analyze is defined as follows.[57] Individuals may meet in pairs or triples and then decide whether to form a given link or triangle. The link is formed if and only if both individuals wish to form the link, and a triangle is formed if and only if all three individuals wish to form it.

In particular, there are probabilities, denoted $\pi_L(diff), \pi_L(same)$, that a given link has an *opportunity* to form (that the pair of people involved meet and can choose to form the relationship) that depend on the pair of individuals being of different castes or of the same caste, respectively. Similarly, there are probabilities, denoted $\pi_T(diff), \pi_T(same)$, that a given triangle has an *opportunity* to form (that the three people involved meet and can choose to form the relationship) that depend on the triple of individuals being of all the same castes or two of the same and one of a different caste.

Preferences are similarly described. Let $p_L(same)$ be the probability that an individual will desire to form a link with an individual of the same caste group, and $p_L(diff)$ be the probability that an individual will desire to form a link with an individual of a

[57]We could use either SERGMs or SUGMs, here.

different caste group. Correspondingly, let $p_L(same)$ be the probability that an individual will desire to form a triangle when all individuals are of the same caste group, and $p_T(diff)$ be the probability that an individual will desire to form a triangle when it consists of people from both caste groups.

The hypothesis that we explore is that $p_T(diff)/p_T(same) < p_L(diff)/p_L(same)$ so that people are more reluctant to involve themselves in cross-caste relationships when those are "public" in the sense that other individuals observe those relationships; with a null hypothesis that they are equal $p_T(diff)/p_T(same) = p_L(diff)/p_L(same)$.

Note that the probability that a "same" link forms is

$$P_L(same) = p_L(same)^2 \pi_L(same)$$

as it requires both agents to agree, and the probability that a "different" link forms is

$$P_L(diff) = p_L(diff)^2 \pi_L(diff).$$

Similarly, the probability that a "same" triangle forms is

$$P_T(same) = p_T(same)^3 \pi_T(same)$$

and a "different" triangle forms is

$$P_T(diff) = p_T(diff)^3 \pi_T(diff),$$

where the cubic captures the fact that it takes three agreements to form the triangle. The difference in the exponents reflects that it is more difficult to get a triangle to form than a link. Hence, to perform a careful test, we have to adjust for the exponents as otherwise we would just uncover a natural bias due to the exponent that would end up favoring cross-caste links.

Another challenge in identifying a preference bias is that it could be confounded by the meeting bias. Thus, we first model the meeting process more explicitly and show that it should produce an bias towards making triangles relatively more likely to be cross-caste than links. Thus, our test is conservative in the sense that if we find cross-caste links relatively more likely, that is evidence for a (strong) preference bias.

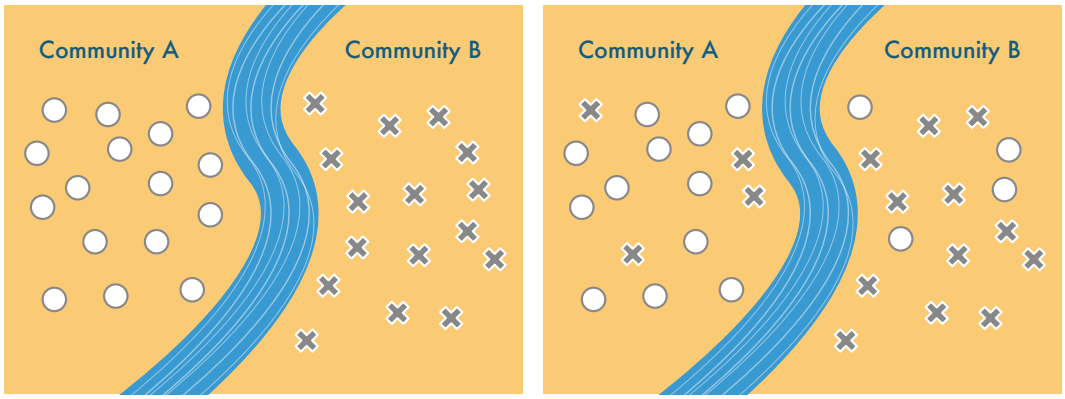
Consider a meeting process where people spend a fraction f of their time mixing in the community that is predominantly of their own types and a fraction $1 - f$ of their time mixing in the other caste's community. Then at any given snapshot in time, a community would have f of its own types present and $1 - f$ of the other type present, as depicted in Figure 5.

Having two randomly picked nodes bump into each other within a community, there is a $f^2+(1-f)^2$ probability of the nodes being of the same type, and a $1-(f^2+(1-f)^2)$ probability of them being of different types.[58] Thus, the relative meeting frequency of different type links compared same type links is

$$\frac{\pi_L(diff)}{\pi_L(same)} = \frac{1 - (f^2 + (1-f)^2)}{f^2 + (1-f)^2}.$$

For triangles, picking three individuals out of the community at any point in time would lead to a $f^3 + (1-f)^3$ probability that all three are of the same type, and

[58] To keep things simple, we consider equal-sized groups, but the argument extends with some adjustments to asymmetric sizes.



(A) Individuals all on own-community side of river

(B) Fraction $f = \frac{1}{4}$ mixed across communities

FIGURE 5. A geographically driven meeting process such that individuals spend 3/4 of their time in their own community are thus more likely to meet their own kind.

$1-(f^2+(1-f)^2)$ of them being of mixed types, and so

$$\frac{\pi_T(diff)}{\pi_T(same)} = \frac{1-(f^3+(1-f)^3)}{f^3+(1-f)^3}.$$

It follows directly that for $f \in (0,1)$:

$$\frac{\pi_T(same)}{\pi_T(diff)} < \frac{\pi_L(same)}{\pi_L(diff)}. \tag{7.1}$$

So different type triangles are more likely to have opportunities to form under this random mixing model than different type links. In particular, note that

$$\frac{p_T(diff)}{p_T(same)} < \frac{p_L(diff)}{p_L(same)} \text{ if and only if } \left(\frac{P_T(diff)}{P_T(same)}\frac{\pi_T(same)}{\pi_T(diff)}\right)^{1/3} < \left(\frac{P_L(diff)}{P_L(same)}\frac{\pi_L(same)}{\pi_L(diff)}\right)^{1/2}.$$

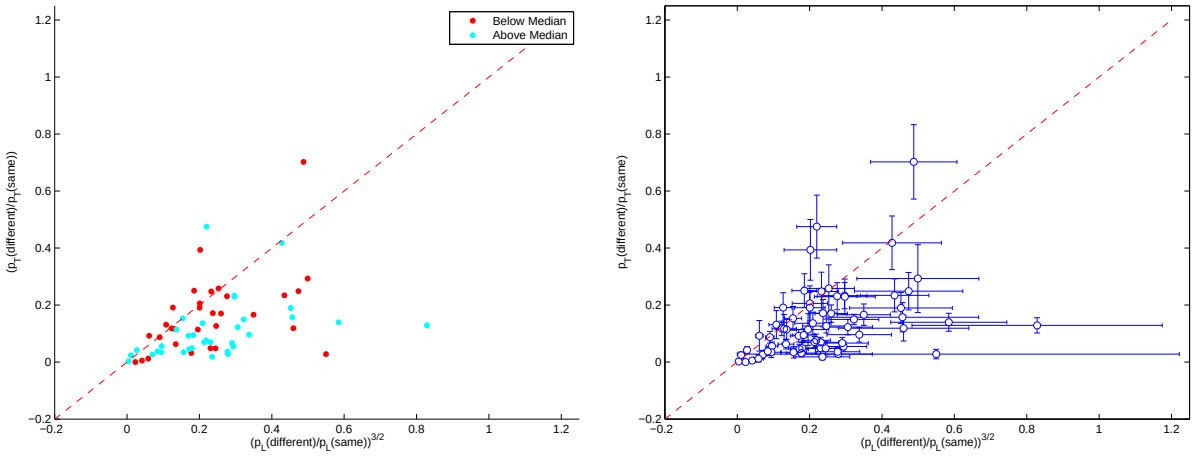
In summary, given (7.1), a sufficient condition for $\frac{p_T(diff)}{p_T(same)} < \frac{p_L(diff)}{p_L(same)}$ is that

$$(P_T(diff)/P_T(same)) < (P_L(diff)/P_L(same))^{3/2}.$$

Figure 6 shows the results. For the bulk of villages, cross-caste relationships relative to within-caste relationships are more frequent as isolated links as opposed to being embedded in triangles, *even when adjusting for the fact that triangles take more consent*. The difference is significant well beyond a 99.9 percent confidence level.[59]

In the left panel of Figure 6 villages are color coded by the relative sizes of the two caste-based groups. The red villages are such that one of the two caste designations dominates the village and the other group is relatively small, while the blue villages are ones in which the two caste designations are more balanced in terms of sizes. In other contexts, homophily has been found to be strongest when groups are evenly

[59]This is from doing a conservative nonparametric test: under the null hypothesis the number of villages for which the ratio is less should be 1/2 with a binomial distribution on the number above or below.



(A) Colored by fragmentation (B) Pointwise standard errors

FIGURE 6. Comparison of the relative propensity to form cross-caste versus same-caste relationships for triangles (vertical axis) compared to links (horizontal axis). The propensity is lower for triangles than links in a significant number of villages, even when adjusting link propensities downwards by raising them to the 3/2 power to adjust for the number of consents needed to form the subgraphs. The color coding on the left panel distinguishes those villages that have above/below the median size minority group.

balanced (e.g., see McPherson et al. (2001); Currarini et al. (2009, 2010)). Here we see that the social pressures against mixed-caste triangles are stronger when the two caste designations are more evenly balanced.

7.3. **Multigraphs and motives for linking.**

Does the fact that two agents borrow and lend money to each other make it more likely that they will also seek advice from each other, or engage in other sorts of relationships? Is it more likely that pairs of agents who have multiple relationships with each other will be embedded in cliques?

These questions involve multigraphs, and asks whether different sorts of relationships are dependent upon each other and on network features. The observation that different types of relationships that individuals have with each other may be interdependent dates back to Simmel (1908). Our SERGM and SUGM models extend directly to multigraph settings in obvious ways, as we illustrate in the application below.

We begin by presenting three basic theories for multiplexing in social relationships and then employ our techniques to investigate them in data.

First, there may be a fixed cost of forming relationships. Conditional on having established one relationship with another person, it is cheaper to construct the second relationship. That is, if i is willing to link to j to borrow kerosene/rice, then it might

be cheaper to for i to also establish a social link with j. Decreasing marginal costs after a first relationship would lead individuals to tend to form multiple relationships with other individuals rather than many single relationships with different individuals.

Second, decisions of individuals to form relationships depend on their compatibility and characteristics. To the extent that compatibility is driven by characteristics beyond those quantified in a data set, pairs of individuals' decisions to form relationships could be correlated even after controlling for all observable characteristics. Thus, if i and j lend each other kerosene and rice, then they are likely to be "compatible," which then suggests that they are more likely to have social and other links together.

Third, is what we call a *support theory* based on incentives for informal favor exchange. In such a theory there are two ways to support the exchange of costly favors. One way is to have frequent enough favor exchange so that it is in each individual's interest to provide favors whenever his or her friend asks for one. In this way, an isolated pair of individuals who have multiple relationships (exchanging various types of favors, etc.) can sustain more favor exchange. This theory is thus based on complementarities between relationships: the value of one type of relationships provides additional incentives to perform in other types of relationships thus enhancing the value of those other relationships. Another source of incentives to exchange favors is to embed that exchange in a clique. If some individual in a clique fails to provide a favor when asked, then all of the other individuals in the clique collectively punish the agent by not providing favors future favors to that individual. The embedding in a clique provides the additional incentives necessary to motivate costly favor provision.[60] Thus, the leverage needed to incentivize the provision of favors comes either through building valuable multi-level bilateral relationships, or else by embedding the exchange within cliques with more individuals, or both. Under this theory, it should be rare to see favor exchange between two individuals who do not share multiple relationships and do not have friends in common.

To examine these theories, we use the framework outlined in Section 3.3. We again use the data collected across 75 Indian villages collected by Banerjee et al. (2013). The specific usefulness of the data is that it includes multigraphs including various forms of favor exchange links (whether households borrow or lend kerosene, rice, money,...) and social links (whether household members visit each other socially, go to temple together, ...). Favor exchange links indicate whether households borrow/lend kerosene, rice from/to each other and social links indicate whether members of a household visit members of another household or receive these members as guests.[61]

Let us first distinguish between the first two theories: the fixed cost theory and compatibility theories. To do this we examine multiplexing of supported links (those in cliques) to those that are not supported. Under a theory of fixed costs of linking, multiplexing would have no reason to depend on whether links are supported or not, while under a theory of compatibility (based on unobserved characteristics) multiplexing would be more likely when links are supported then when they are not. In particular, under compatibility, if i and j both have links to k, then that indicates that

[60]For more on this, see Jackson et al. (2012).

[61]For the purposes of our analysis, we use all 75 Indian villages and construct networks at the household level. We build undirected, unweighted multigraphs with two link types (where a link of a given type is present if either household claimed that type of relationship with the other).

they are both compatible with k, for instance sharing common characteristics. In that case, i and j are more likely to share common characteristics with each other than if they were not both linked to some common k, and thus are more likely to have high multiplexing. Unsupported links are more likely to have arisen spuriously and are thus less likely to be multiplexed.

TABLE 2. Support and Multiplexing

	Fraction with Designated Property			Fraction that are Multiplexed		
	Supported	Multiplexed	Supported or Multiplexed	Supported	Unsupported	Difference
Favor	0.7303	0.7688	0.9210	0.7945	0.6991	0.0954***
Social	0.6828	0.5594	0.8327	0.6132	0.4436	0.1696***
Difference	0.0376***	0.2094***	0.0883***	0.1813***	0.2555***	-0.0742***

Notes: Two households have a favor link if they borrow/lend material goods such as kerosene or rice to/from each other. Two households have a social link if they have members that visit each other's households.

The data are consistent with the compatibility theory but not with fixed cost theory, as seen in the last column of Table 2: the difference between the fraction of supported links that are multiplexed and the unsupported links that are multiplexed is significant and positive for both favor and social links.[62] This does not necessarily mean that fixed costs are not at play, but that fixed cost theory cannot be the sole explanation for the high levels of multiplexing.

Next, we consider the support theory. Under that theory, favor exchange links should be more likely to either be embedded in triangles or multiplexed than social links.[63] Indeed, as seen in Table 2, more than 92 percent of favor links are either supported or multiplexed, while the same is true of only 83 percent of the social links.[64] Moreover, when looking at unsupported links, favor exchange links are almost 26 percent more likely to be multiplexed than social links, which is consistent with the theory. This difference is not predicted by either the fixed cost or compatibility theories.

The analysis shows that neither fixed cost nor compatibility theories could account for the data alone or in combination, while support theory could.[65] It is possible that all three theories have some role in what is going on, and so what we learn from the analysis is that neither a fixed cost nor a compatibility theory can account for all aspects of the data, and that there are significant differences between the support

[62]All of the differences in the table are significant beyond the 99 percent level.

[63]Social links might still involve some favor exchange, so the prediction is valid only to the extent that there are some social relationships that do not involve some sort of favor exchange.

[64]Of course, social links could include things like sharing of information and other sorts of favors, and so some of those links might not be purely hedonic and might require some support or multiplexing to function as well.

[65]Support theory would suggest that the percentage of favor links that are supported or multiplexed should be 100 percent. There are three possible reasons why the number is only 92 percent. One is measurement error. A second is that there are other types of relationships that are not included here. A third is that some favor exchange might be frequent enough between a pair that it is self-sustaining without any support or multiplexing.

and multiplexing of favor exchange versus social links and so incentives seem to be an integral part of the story consistent with support theory.[66]

This exercise of counting supported links and unsupported links separately and directly is legitimized under our SUGMs, in which all of the three theories are embedded.

8. Conclusion

We presented two new classes of models of network formation, SERGMs and SUGMs, based on the idea that network formation is driven by the properties of a network and/or by the formation of various subgraphs. This turns the focus away from the network as the unit of analysis, and instead focuses on its properties, subgraphs and statistics. This perspective allows us to develop direct estimation techniques and to derive results concerning consistency and asymptotic distributions of the estimators. Given the growing literature on estimation of network formation, such results are essential.

Models of network formation, such as those described here, can be useful beyond simply studying network patterns, such as in situations where network features are drivers of economic behaviors. For example, farmers may be significantly more likely to learn to use a new farming technology in one network than another, and so influencing the network of communication among farmers could be useful. SERGMs and SUGMs, can help us understand the drivers of network formation and thus which sorts of interventions might lead to improvements in network features of interest. Moreover, it is apparent that accounting for endogeneity of networks is important in studying peer influence (e.g., Aral et al. (2009); Goldsmith-Pinkham and Imbens (2013); Jackson (2013); Lindquist and Zenou (2013)) and so having practical models of network formation is useful beyond direct estimation.

It is important to emphasize that our models can easily simulate networks. This can be difficult (impossible) for standard ERGMs. Here, our work suggests several avenues. First, in the case of sparse networks, the SUGMs are perfectly suited for easily generating networks: the model directly translates into an algorithm for generating networks by generating various subnetworks. Second, in the case of SERGMs, the models are well-adapted to generating statistics of networks, even though in some cases it might be difficult to generate the networks themselves. For example, if some profile of statistics S is generated, then randomly picking a network g that exhibits statistics exactly S can be a hard problem. This suggests a third avenue in line with our interpretation of SUGMs. Instead of viewing S as the realized statistic of the network, nature forms a network by forming subnetworks, even when they are dense. So, the S profile of generated subnetworks is picked by nature based on the SERGM. This will generate some incidental subnetworks, and so a different observed S' from S, but is still a perfectly well-defined model for generative purposes.[67]

[66]There are other patterns in the data that the three theories yield no predictions about, which present interesting questions for future research.

[67]In fact, by simulating such processes one can estimate the relationship between S' and S which can then be used for estimation purposes when networks are not sparse, similar to the algorithm we provide for SUGMs. We leave the general treatment of such algorithms and estimations for future research.

Also, given our results, a researcher can use variations of standard approaches of model selection to deduce which statistics to incorporate in a SERGM or SUGM. Consider the following example. Suppose that a researcher is interested in developing a model that captures the density, homophily, and clustering in an observed network. An objective function can be built where a model's score is based on the total difference between its predictions of the relevant statistics (under best-fit parameters) and the observed statistics; and, as is commonly done, a penalty can be included for the number of parameters in the model. Then one can examine SUGMs that incorporate various subsets of characteristic-based links, triangles, larger cliques, isolated nodes, and so forth, and check to find which model minimizes the objective function and is thereby selected.

Finally, we note that the approach we have taken can be further extended. In fact, once one adopts a SERGM formulation, many other sorts of applications beyond networks, such as matching problems, partitioning problems, club membership and others can also be incorporated.

References

Aral, S., L. Muchnik, and A. Sundararajan (2009): "Distinguishing Influence Based Contagions from Homophily Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences*. 8

Badev, A. (2013): "Discrete Games in Endogenous Networks: Theory and Policy," *mimeo: University of Pennsylvania*. 6.1.2

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. Jackson (2013): "Diffusion of Microfinance," *Science*, 341, DOI: 10.1126/science.1236498, July 26 2013. 7.1, 7.3

Bhamidi, S., G. Bresler, and A. Sly (2008): "Mixing time of exponential random graphs," *Arxiv preprint arXiv:0812.2265*. 1, 1, 2.1.1

Bickel, P., A. Chen, and E. Levina (2011): "The method of moments and degree distributions for network models," *Annals of Statistics*, 39, 2280–2301. 5

Bloch, F. and M. Jackson (2006): "Definitions of equilibrium in network formation games," *International Journal of Game Theory*, 34, 305âĂŞ318. 46

Bollobás, B., S. Janson, and O. Riordan (2011): "Sparse random graphs with clustering," *Random Structures & Algorithms*, 38, 269–323. 13

Borgs, C., J. Chayes, J. Ding, and B. Lucier (2010): "The Hitchhiker's Guide to Affiliation Networks: A Game-Theoretic Approach," *ArXiv:1008.1516v1*. 6.1.4

Bramoullé, Y. and R. Kranton (2007): "Risk-sharing networks," *Journal of Economic Behavior & Organization*, 64, 275–294. 1

Butts, C. (2009): "Using Potential Games to Parameterize ERG Models," *working paper, UCI*. 6.1.2

Calvo-Armengol, A. (2004): "Job contact networks," *Journal of Economic Theory*, 115, 191–206. 1

Calvo-Armengol, A. and Y. Zenou (2005): "Job matching, social network and word-of-mouth communication," *Journal of Urban Economics*, 57, 500–522. 1

Caulier, J.-F., A. Mauleon, and V. Vannetelbosch (2013): "Contractually stable networks," *International Journal of Game Theory*, 1–17. 41

Chandrasekhar, A. and R. Lewis (2013): "Econometrics of sampled networks," Stanford working paper. 5, 54

Chatterjee, S. and P. Diaconis (2011): "Estimating and Understanding Exponential Random Graph Models," *Arxiv preprint arXiv:1102.2650*. 4

Chatterjee, S., P. Diaconis, and A. Sly (2010): "Random graphs with a given degree sequence," *Arxiv preprint arXiv:1005.1136*. 1, 1

Christakis, N., J. Fowler, G. Imbens, and K. Kalyanaraman (2010): "An Empirical Model for Strategic Network Formation," *NBER Working Paper*. 7

Currarini, S., M. Jackson, and P. Pin (2009): "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 77, 1003–1045. 6.1.4, 7.2

——— (2010): "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107, 4857–4861. 45, 6.1.4, 7.2

DeMarzo, P., D. Vayanos, and J. Zwiebel (2003): "Persuasion Bias, Social Influence, and Unidimensional Opinions*," *Quarterly journal of economics*, 118, 909–968. 7.1

Diaconis, P. and D. Freedman (1981): "On the Statistics of Vision: The Julesz Conjecture," *Journal of Mathematical Psychology*, 24, 112âĂŞ–138. 7.1

Frank, O. and D. Strauss (1986): "Markov graphs," *Journal of the American Statistical Association*, 832–842. 1

Furusawa and H. Konishi (2007): "Free trade networks," *Journal of International Economics*, 7, 310–335. 1

Goldsmith-Pinkham, P. and G. Imbens (2013): "Social Networks and the Identification of Peer Effects," *Journal of Business and Economic Statistics*, 31:3, 253–264. 8

Golub, B. and M. Jackson (2012): "How Homophily Affects the Speed of Learning and Best-Response Dynamics," *Quarterly Journal of Economics*, 127, 1287–1338. 7.1

Golub, B. and Y. Livne (2010): "Strategic Random Networks: Why Social Networking Technology Matters," *SSRN working paper*. 6.1.4

Hammersley, J. and P. Clifford (1971): "Markov fields on finite graphs and lattices," . 1

Handcock, M. (2003): "Assessing degeneracy in statistical models of social networks," . 1, 16

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, and M. Morris (2003): *statnet: Software tools for the Statistical Modeling of Network Data*, Seattle, WA. 2.1.1

Jackson, M. (2008): *Social and economic networks*, Princeton: Princeton University Press. 2, 7.1

——— (2011): *Handbook of Social Economics*, San Diego: North Holland, chap. An Overview of Social Networks and Economic Applications. 16

——— (2013): "Unraveling Peers and Peer Effects: Comments on Goldsmith-Pinkham and Imbens' "Social Networks and the Identification of Peer Effects"," *Journal of Business and Economic Statistics*, 31:3, 270–273, DOI: 10.1080/07350015.2013.794095. 8

Jackson, M., T. Barraquer, and X. Tan (2012): "Social Capital and Social Quilts: Network Patterns of Favor Exchange," *American Economic Review*, 102,

1857–1897. 1, 60
Jackson, M. and A. van den Nouweland (2005): "Strongly Stable Networks," *Games and Economic Behavior*, 51, 420–444. 41
Jackson, M. and A. Watts (2001): "The Existence of Pairwise Stable Networks," *Seoul Journal of Economics*, 14(3), 299–321. 6.1.2, 6.1.2
Jackson, M. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71, 44–74. 44, 46
Krackhardt, D. (1988): "Predicting with Networks: Nonparameteric Multiple Regression Analysis of Dyadic Data," *Social Networks*, 10, 359–381. 1
Lindquist, M. and Y. Zenou (2013): "Key Players in Co-offending Networks," *working paper Stockholm University*. 8
McPherson, M., L. Smith-Lovin, and J. Cook (2001): "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444. 7.2
Mele, A. (2011): "A Structural Model of Segregation in Social Networks," *working paper*. 7, 6.1.2, 6.1.3, 50
Moreno, J. and H. Jennings (1938): "Statistics of Social Configurations," *Sociometry*, 1, 342–374. 1
Rinaldo, A., S. Petrovic, and S. Fienberg (2011): "Maximum likelihood estimation in network models," *arXiv preprint arXiv:1105.6145*. 23
Schutz, B. (2003): *Gravity from the ground up*, Cambridge Univ Pr. 15
Shalizi, C. and A. Rinaldo (2012): "Consistency under Sampling of Exponential Random Graph Models," *ArXiv 1111.3054v3*. 5
Simmel, G. (1908): *Sociology: Investigations on the Forms of Sociation*, Leipzig: Duncker and Humblot. 7.3
Snijders, T. (2002): "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, 3, 240. 1, 16
Snijders, T., P. Pattison, G. Robins, and M. Handcock (2006): "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153. 10, 16, 17
Topkis, D. (2001): *Supermodularity and Complementarity*, Princeton University Press. 6.1.4
Vives, X. (2007): "Supermodularity and Supermodular Games," *IESE Occasional Paper 07/18*. 52
Wasserman, S. and P. Pattison (1996): "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp," *Psychometrika*, 61, 401–425. 1

Appendix A. Proofs

Proof of Theorem 1.

Consider a sequence of count statistics $S^n = (S_1^n, \ldots, S_k^n)$ whose ℓ-th entry takes on nonnegative integer values with some maximum value $\overline{S}_\ell^n \to \infty$, and the count SERGMs specified with $K^n(s) = \prod_\ell \binom{\overline{S}_\ell^n}{s_\ell}$.

$$\mathrm{P}_\beta\left(S^n = s\right) = \frac{\prod_\ell \binom{\overline{S}_\ell^n}{s_\ell} \exp\left(\beta^n \cdot s\right)}{\sum_{s' \in A^n} \prod_\ell \binom{\overline{S}_\ell^n}{s'_\ell} \exp\left(\beta^n \cdot s'\right)}. \tag{A.1}$$

We rewrite (A.1) as

$$\mathrm{P}_\beta\left(S^n = s\right) = \frac{\prod_\ell \left[\binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)\right]}{\sum_{s' \in A^n} \prod_\ell \left[\binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)\right]}. \tag{A.2}$$

If we consider the distribution conditional on S^n lying in

$$B^n = \prod_\ell \{\lfloor \mathrm{E}_{\beta^n_\ell}[S^n_\ell](1-\varepsilon)\rfloor, \ldots, \lfloor \mathrm{E}_{\beta^n_\ell}[S^n_\ell](1+\varepsilon)\rfloor\} \subset A^n$$

(for large enough n under the not conflicted condition), then we can write

$$\mathrm{P}_\beta\left(S^n = s | s \in B^n\right) = \frac{\prod_\ell \left[\binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)\right]}{\sum_{s' \in B^n_\ell} \prod_\ell \left[\binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)\right]} = \frac{\prod_\ell \left[\binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)\right]}{\prod_\ell \left[\sum_{s'_\ell \in B^n_\ell} \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)\right]},$$

or

$$\mathrm{P}_\beta\left(S^n = s | s \in B^n\right) = \prod_\ell \frac{\binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)}{\sum_{s'_\ell \in B^n_\ell} \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)}. \tag{A.3}$$

Next, we consider binomial distributions which we will relate back to the above expressions for the SERGM.

For each ℓ, consider a binomial distribution that has $p^n_\ell = \frac{\mathrm{E}_{\beta^n}[S^n_\ell]}{\overline{S}^n_\ell}$, with range from 0 to $\overline{S}^n_\ell$. Taking a product of independent binomial distributions

$$\mathrm{P}^{Bin}\left(\widetilde{S}^n = s\right) = \prod_\ell \frac{\binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)}{\sum_{s'_\ell \in [0,\overline{S}^n_\ell]} \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)} = \frac{\prod_\ell \binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)}{\sum_{s'_\ell \in [0,\overline{S}^n_\ell]} \left[\prod_\ell \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)\right]}. \tag{A.4}$$

The corresponding probability that $\widetilde{S}^n = s$ given $s \in B^n$ can be written as

$$\mathrm{P}^{Bin}\left(\widetilde{S}^n = s | s \in B^n\right) = \prod_\ell \frac{\binom{\overline{S}^n_\ell}{s_\ell} \exp\left(\beta^n_\ell s_\ell\right)}{\sum_{s'_\ell \in B^n_\ell} \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)}. \tag{A.5}$$

For a binomial distribution, the probability that $\widetilde{S}^n_\ell \in B^n_\ell \to 1$. Thus, under independent binomial distributions, the probability that $\widetilde{S}^n \in B^n \to 1$, and so it follows from (A.4) and (A.5) that

$$\frac{\mathrm{P}^{Bin}\left(\widetilde{S}^n = s | s \in B^n\right)}{\mathrm{P}^{Bin}\left(\widetilde{S}^n = s\right)} = \frac{\sum_{s'_\ell \in [0,\overline{S}^n_\ell]} \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)}{\sum_{s'_\ell \in B^n_\ell} \binom{\overline{S}^n_\ell}{s'_\ell} \exp\left(\beta^n_\ell s'_\ell\right)} \to 1, \tag{A.6}$$

uniformly for $s \in B^n$.

Collecting from (A.2) and (A.4) it follows that for $s \in B^n$

$$\mathrm{P}_\beta\left(\widetilde{S}^n = s\right) \geq \mathrm{P}^{Bin}\left(\widetilde{S}^n = s\right). \tag{A.7}$$

Collecting from (A.3) and (A.5) it follows that for $s \in B^n$

$$\mathrm{P}^{Bin}\left(\widetilde{S}^n = s | s \in B^n\right) = \mathrm{P}_\beta\left(\widetilde{S}^n = s | s \in B^n\right). \tag{A.8}$$

Then (A.7) and (A.8) and the fact that $\mathrm{P}_\beta\left(\widetilde{S}^n = s | s \in B^n\right) \geq \mathrm{P}_\beta\left(\widetilde{S}^n = s\right)$, together with (A.6), imply that

$$\frac{\mathrm{P}_\beta\left(\widetilde{S}^n = s\right)}{\mathrm{P}^{Bin}\left(\widetilde{S}^n = s\right)} \to 1, \tag{A.9}$$

uniformly for $s \in B^n$.

The remainder of the claimed results then follows easily from standard properties of the binomial distribution (see also, Lemma E.1).

In particular, the variance terms are computed as follows. First consider a binomial $\mathrm{Bin}(p_\ell^n; n)$ with $p_\ell^n n \to \infty$. By the Lindeberg-Feller Central Limit Theorem $\sqrt{n}\left(\widehat{p}_\ell^n - p_\ell^n\right) \rightsquigarrow \mathcal{N}\left(0, p_\ell^n\left(1 - p_\ell^n\right)\right)$. Letting $\beta_\ell^n = g\left(p_\ell^n\right) = \log \frac{p_\ell^n}{1-p_\ell^n}$, note

$$g'(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.$$

Finally, by the delta method,

$$\sqrt{n}\left(\widehat{\beta_\ell^n} - \beta_\ell^n\right) \rightsquigarrow \mathcal{N}\left(0, p_\ell^n\left(1 - p_\ell^n\right)\left[g'(p_\ell^n)\right]^2\right)$$

and therefore observing that $p_\ell^n = \frac{\exp \beta_\ell^n}{1+\exp \beta_\ell^n}$, $p_\ell^n(1-p_\ell^n)\left[\frac{1}{p_\ell^n(1-p_\ell^n)}\right]^2 = \frac{1}{p_\ell^n(1-p_\ell^n)}$ and substituting for β_ℓ^n, we have

$$\sqrt{n}\left(\widehat{\beta_\ell^n} - \beta_\ell^n\right) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{\frac{\exp \beta_\ell^n}{1+\exp \beta_\ell^n}\left(1 - \frac{\exp \beta_\ell^n}{1+\exp \beta_\ell^n}\right)}\right).$$

This argument replacing n with $\bar{S}_\ell^n$ shows the result. ■

Proof of Theorem 2. We provide the proof without covariates to save on notation, but it extends directly. We begin the proof by showing the following. For any ℓ, the fraction of counts of subnetworks ℓ generated incidentally by some other subnetworks goes to 0. That is, consider some ℓ and $g_\ell \in G_\ell^n$ on m_ℓ nodes. Let

$$\widetilde{p}_\ell^n = \frac{\mathrm{E}(\widetilde{S}_\ell^n)}{y_\ell^n \binom{n}{m_\ell}}.$$

This is no more than p_ℓ^n, as the denominator includes all possible subgraphs of size ℓ (where y_ℓ^n is the number of subgraphs of type ℓ that can be formed on any given m_ℓ nodes).[68] Let us consider the probability z_ℓ^n that g_ℓ is incidentally generated by other subnetworks.

We show that $z_\ell^n / \widetilde{p}_\ell^n \to 0$, which implies that $z_\ell^n / p_\ell^n \to 0$.

Consider $g_\ell \in G_\ell^n$ and an incidentally generating subclass $(\ell_j, h_j)_{j \in J}$.

[68]Here we provide the proof that applies without subgraphs being characteristic dependent. The extension to characteristic dependent subgraphs is straightforward simply by adjusting all numbers to reflect possible networks with given node characteristics as dependent on n and the sets of nodes that have particular characteristics as a function of n, but it is notationally much more intensive.

We show that the probability $z_\ell^n(J)$ that it is generated by this subclass goes to zero relative to $\widetilde{p}_\ell^n$, so that $z_\ell^n(J)/\widetilde{p}_\ell^n \to 0$ for each J, and since there are at most $M_\ell \leq k^{m_\ell}$ such generating classes, this implies that $z_\ell^n/\widetilde{p}_\ell^n \to 0$.

For a subnetwork in $G_{\ell_j}^n$, the probability of getting *at least one* such network that has the h_j nodes out of the m_ℓ in g_ℓ is no more than[69]

$$y_{\ell_j}^n \binom{n}{m_{\ell_j} - h_j} \widetilde{p}_{\ell_j}^n \leq y_{\ell_j}^n n^{m_{\ell_j} - h_j} \widetilde{p}_{\ell_j}^n.$$

Thus,

$$\frac{z_\ell^n(J)}{\widetilde{p}_\ell^n} \leq \frac{\Pi_{j \in J} n^{m_{\ell_j} - h_j} y_j^n \widetilde{p}_{\ell_j}^n}{\widetilde{p}_\ell^n}.$$

Therefore

$$\frac{z_\ell^n(J)}{\widetilde{p}_\ell^n} \leq \frac{\Pi_{j \in J} y_j^n n^{m_{\ell_j}} \widetilde{p}_{\ell_j}^n}{n^{\sum_j h_j} \widetilde{p}_\ell^n}.$$

Recall that $M_J = \sum_{j \in J} h_j - m_\ell$ and that $M \geq 1$ (since $|J| \geq 2$ and each set of h_j intersects with at least one other set of $h_{j'}$ for some $j' \neq j$).

Therefore

$$\frac{z_\ell^n(J)}{\widetilde{p}_\ell^n} \leq \frac{\Pi_{j \in J} y_{\ell_j}^n n^{m_{\ell_j}} \widetilde{p}_{\ell_j}^n}{n^{M_J} n^{m_\ell} \widetilde{p}_\ell^n}.$$

The numerator is of the order $\Pi_{j \in J} \mathrm{E}(S_j^n)$ while the denominator is of the order $n^{M_J} \mathrm{E}(S_\ell^n)$.

Under the sparseness condition,

$$\frac{\Pi_{j \in J} \mathrm{E}(S_{\ell_j}^n)}{n^{M_J} \mathrm{E}(S_\ell^n)} \to 0,$$

and so we have verified the claim.

To finish the ratio consistency proof, note that the claim then implies that $\frac{\widehat{S}_\ell^n(g)}{S_\ell^n} \to 1$. Thus, dividing top and bottom by $\overline{S}_\ell^n(g)$, it follows that $\frac{\widehat{p}_\ell^n(g)}{S_\ell^n/\overline{S}_\ell^n(g)} \to 1$. Given the growing condition and properties of the binomial distribution, it also follows that $\frac{S_\ell^n/\overline{S}_\ell^n(g)}{p_\ell^n} \to 1$, and so $\frac{\widehat{p}_\ell^n(g)}{p_\ell^n} \to 1$.

Next, note that the above also implies that the distribution $D_n^{1/2} \left(\widehat{p}_1^n(g), ..., \widehat{p}_k^n(g)\right)'$ converges to the distribution of $D_n^{1/2} \left(\widetilde{p}_1^n(g), ..., \widetilde{p}_k^n(g)\right)'$, where $\widetilde{p}_\ell^n(g) = S_\ell^n/\overline{S}_\ell^n(g)$.

The asymptotic normality of the (joint) distribution then follows from the usual Linderberg-Feller central limit theorem applied to triangular arrays of binomial random variables. This applies under the growing conditions of the theorem.[70] ■

[69] This is a loose upper bound as it simply adds the probability that each possible one forms - but becomes more accurate as the probability of any one occurring vanishes.

[70] Let $X_{1T}, ..., X_{TT}$ be a triangular array of $\mathrm{Ber}(p_T)$ random variables and $p_T T \to \infty$ (as under the growing condition). To apply the Lindeberg-Feller central limit theorem for triangular arrays, we check the Lindeberg condition: for any $\epsilon > 0$,

$$\frac{1}{T p_T (1 - p_T)} \sum_{t \leq T} \mathrm{E}\left[X_t^2 \mathbf{1}\left\{|X_t| \geq \epsilon \sqrt{T p_T (1 - p_T)}\right\}\right] = o(1).$$

The condition is implied by the fact that $T p_T \to \infty$, and from the sparsity conditions which imply that p_T is bounded away from 1.

Proof of Theorem 3.

Let $\mathcal{G}^n$ denote the set of all possible subgraphs in the model:

$$\mathcal{G}^n = \cup_\ell G_\ell^n. \tag{A.10}$$

Letting G be the set of subgraphs that are truly formed, we can write

$$\mathrm{P}(G) = \prod_{g_\ell \in G} \frac{\exp(\theta_\ell)}{\exp(\theta_\ell)+1} \prod_{g_\ell \notin G} \frac{1}{\exp(\theta_\ell)+1}.$$

This can be rewritten as

$$\mathrm{P}(G) = \frac{\exp\left(\sum_{g_\ell \in G} \theta_\ell\right)}{\prod_{g_\ell \in \mathcal{G}^n} (\exp(\theta_\ell)+1)}.$$

Note that

$$\prod_{g_\ell \in \mathcal{G}^n} (\exp(\theta_\ell)+1) = \sum_{A \subset \mathcal{G}^n} \left(\exp(\sum_\ell |A \cap G_\ell^n| \theta_\ell) \right).$$

Given that the number of A's that have counts $s' \in \prod_\ell \{0, \ldots, \overline{S}_\ell^n\}$ is exactly

$$\prod_\ell \binom{\overline{S}_\ell^n}{s'_\ell},$$

it follows that

$$\prod_{g_\ell \in \mathcal{G}^n} (\exp(\theta_\ell)+1) = \sum_{s' \in \prod_\ell \{0,\ldots,\overline{S}_\ell^n\}} \left(\prod_\ell K_\ell^n(s'_\ell) \right) \exp(\sum_\ell s'_\ell \theta_\ell),$$

where $K_\ell^n(s'_\ell) = \binom{\overline{S}_\ell^n}{s'_\ell}$. This means that we can write

$$\mathrm{P}(G) = \frac{\exp\left(\sum_\ell |G_\ell^n \cap G| \cdot \theta_\ell\right)}{\sum_{s' \in \prod_\ell \{0,\ldots,\overline{S}_\ell^n\}} \left(\prod_\ell K_\ell^n(s'_\ell)\right) \exp\left(\sum_\ell s'_\ell \theta_\ell\right)}. \tag{A.11}$$

Note that there are $K_\ell^n(\widetilde{S})$ different collections of truly generated subgraphs G that have the same value $\widetilde{S}$ and that each is equally likely. Thus

$$\mathrm{P}(\widetilde{S}) \frac{K_\ell^n(\widetilde{S}) \exp\left(\sum_\ell \widetilde{S}_\ell \cdot \theta_\ell\right)}{\sum_{s' \in \prod_\ell \{0,\ldots,\overline{S}_\ell^n\}} \left(\prod_\ell K_\ell^n(s'_\ell)\right) \exp\left(\sum_\ell s'_\ell \theta_\ell\right)}, \tag{A.12}$$

which is the same as (5.3), which completes the proof. ■

Appendix B. A Useful Lemma on SERGM Statistic Domains

Lemma B.1. *Consider a SERGM with associated* $\mathrm{P}_\beta(s)$ *as described in (3.1) and consider any* $\varepsilon > 0$. *Suppose that for each* β *in some set* B *there exists* A_β *such that* $\mathrm{P}_\beta(A_\beta) \geq 1 - \varepsilon$. *For each* s *let* $A_s = \cup_{\beta \in B : s \in A_\beta} A_\beta$. *Letting*

$$\overline{\mathrm{P}}_\beta(s) = \frac{K_S(s) \exp(\beta s)}{\sum_{s' \in A_s} K_S(s') \exp(\beta s')},$$

it follows that for any $\beta \in B$ *and* $s \in A_\beta$:

$$\frac{1}{1-\varepsilon} \geq \frac{\overline{\mathrm{P}}_\beta(s)}{\mathrm{P}_\beta(s)} \geq 1.$$

Proof of Lemma B.1.

Let

$$\widehat{\mathrm{P}}_\beta(s) = \frac{K_S(s)\exp(\beta s)}{\sum_{s'\in A_\beta} K_S(s')\exp(\beta s')}.$$

Since $\mathrm{P}_\beta(A_\beta) \geq 1-\varepsilon$, it follows that

$$\frac{1}{1-\varepsilon} \geq \frac{\sum_{s'} K_S(s')\exp(\beta s')}{\sum_{s'\in A_\beta} K_S(s')\exp(\beta s')} \geq 1.$$

this implies that for any β and $s \in A_\beta$:

$$\frac{1}{1-\varepsilon} \geq \frac{\widehat{\mathrm{P}}_\beta(s)}{\mathrm{P}_\beta(s)} \geq 1.$$

Note also that for any β and $s \in A_\beta$

$$\widehat{\mathrm{P}}_\beta(s) \geq \overline{\mathrm{P}}_\beta(s) \geq \mathrm{P}_\beta(s).$$

The claimed result follows from the last two sets of inequalities. ■

Appendix C. Online Appendix: Extension to Continuous and Interdependent Covariates

We consider an environment in which nodes draw covariates that can be continuous and even interdependent. Then, based on their characteristics, they form a graph via the SUGM process. We are interested in estimating both probability functions as well as possible parameters which may correspond to random utility foundations (e.g., coefficients in a logistic regression term).

Environment. Every node $i \in \{1, ..., n\}$ draws a d-dimensional covariate vector $x_i^n \in \mathcal{X}$. For simplicity we let $\mathcal{X} = \prod_{k=1}^{d} [x_{L,k}, x_{H,k}]$ be a d-dimensional product of intervals of $\mathbb{R}$.[71] We assume $x_\ell x_\ell'$ has full rank along the sequence. For expositional simplicity in our proofs we considering a sequence of fixed-regressors, $x_{\ell,n}$ where n indexes the sequence. Clearly stochastic regressors can be accomodated.

Example C.1. Let $x_i^n = (1, u_i^n)$, where $u_i^n \in [0, 1]$ such that the design matrix carries full rank. In the simulation exercise corresponding to this example, we will draw them as independent $U[0, 1]$ random variables.

SUGM Formation. Given characteristics, the n nodes engage in a SUGM graph formation process. The realized data sequence consists of a triangular array of random graphs and covariate vectors drawn from a random field $\{(g^n, (x_1^n, ..., x_n^n)) : \ n \in \mathbb{N}\}$. The researcher observes this for a given n and a given realization.

Specifically, consider a set of nicely ordered statistics (S_ℓ^n) again with each statistic counting subgraphs H_ℓ with m_ℓ nodes, where the statistics S_ℓ do not condition on covariates. We are therefore counting, for instance, 4-cliques, triangles (not in 4-cliques), and unsupported links.

A group of size m_ℓ forms with a probability $p_\ell^n(x_{\ell,j}; \gamma_\ell)$ which depends on some function of the m_ℓ individuals' characteristics and a parameter γ_ℓ, whose value in theory may depend on n.[72]

To make things concrete, examples of $p_\ell^n(x_\ell; \gamma)$ include:

(1) a linear probability model with uniform link function $p_\ell^n(x_{\ell,j}; \gamma_\ell) = \gamma_{\ell,n}' x_{j,\ell}$,

(2) a logistic regression model $p_\ell^n(x_{\ell,j}; \gamma_\ell) = \frac{\exp(\gamma_{\ell,n}' x_{j,\ell})}{1+\exp(\gamma_{\ell,n} x_{j,\ell})}$,

for $j \in \{1, ..., \overline{S}_\ell(g)\}$. It should be clear that there are any number of examples here that could be used and the choice is up to the modeler's discretion as to what best describes the nature of the problem at hand.

A truly generated object is a subgraph on m_ℓ nodes that is generated in the ℓth phase independently with probability $p_\ell^n(x_{\ell,j}; \gamma_\ell)$. Incidental generation may occur and the union is the graph g^n.

The group-level characteristic, x_ℓ, is of course a function of individual level characteristics: $x_{\ell,i_1,...,i_{m_\ell}} = f_\ell(x_{i_1}, ..., x_{i_\ell})$. For example, $f_\ell(x_i, x_j) = |x_i - x_j|$.

[71] We will allow these covariates to be interdependent. The substantive assumption we need to make is that the sequences of design matrices and have full rank.

[72] It is easy to modify this such that $f_\ell = f_{\ell,i}$ so that every node makes its own decision to be in the group or not, and its covariates are not treated symmetrically with the other m_ℓ nodes.

Example C.1. [Continued] The sequence of graphs g^n are triangles and links-based. A triangle forms with probability defined by log-odds

$$\log \frac{p_T^n(x_T;\gamma_T)}{1-p_T^n(x_T;\gamma_T)} = \gamma'_{0,n,T}x_T = (\alpha_{0,n,T},\beta_{0,T})x_T$$

where $x_T = (1,u_T)$ and $u_T = (|u_i-u_j|+|u_j-u_k|+|u_k-u_i|)/3$.

A link forms with probability

$$\log \frac{p_L^n(x_L;\gamma_L)}{1-p_L^n(x_L;\gamma_L)} = \gamma'_{0,n,L}x_L = (\alpha_{0,n,L},\beta_{0,L})x_L$$

where $x_L = (1,u_L)$ and $u_L = |u_i-u_j|/2$.

Pairs and triples that are further in covariate space are less likely to link.

Estimation. The above defines a well-defined network-generation process. As before, we need a relative sparsity condition to hold so that when we count a structure, with probability approaching one it was not incidentally generated. Here we provide a sufficient condition for relative sparsity hold as the continuous covariates vary. The condition is that given m_ℓ nodes, no matter what the value of each covariate is among these nodes, the probability of forming the subgraph isomorphic to H_ℓ shrinks at the same as n grows to infinity. This will ensure the relative rate of incidentally generated objects is unaffected by the particular values of the covariates.[73]

LEMMA **C.1.** *Given a growing sequence of graphs with associated covariates and covariate space $\mathcal{X}$, and probability functions $p_\ell^n(x_\ell;\gamma_\ell)$ smooth in both arguments,*

$$\min_{x_1,\ldots,x_{m_\ell}\in\mathcal{X}^{m_\ell}} p_\ell^n\left(x_\ell;\gamma_\ell\right) = O\left(\max_{x_1,\ldots,x_{m_\ell}\in\mathcal{X}^{m_\ell}} p_\ell^n\left(x_\ell;\gamma_\ell\right)\right).$$

If relative sparsity is satisfied at $x_i = 1$ for all i, then relative sparsity is satisfied for any sequence of covariates.

Proof. We can always replace incidental generation probabilities with their maximal values over the covariates, the truly generating probability with its minimal probability. These are all of the same order as when evaluated with $x_i = 1$ by hypothesis. ■

Of course this isn't the only condition to maintain relative sparsity, but it may often be a natural condition to assume.

We now show properties of estimators from the two examples of $p_\ell^n(x_\ell;\gamma_\ell)$ we have discussed.

[73]Such an assumption excludes the possibility that individuals who are close in wealth are more likely to form pairs than triads for wealth levels below some threshold but beyond this threshold it is when individuals are far from others in wealth that pairs are more likely to form than triads. (More specifically, in this example a wealth covariate should not be used, but rather, a wealth covariate with an indicator for whether individuals are below or above the threshold must be used.)

Linear Probability Model. Consider the linear probability model discussed above:

$$p_\ell^n(x_\ell;\gamma_\ell) = \sum_k \gamma_\ell^k x_{k,\ell},\ k = 1, ..., d$$

where $\gamma_{0,n,\ell}^k \to 0$ as $n \to \infty$. It is straightforward to check that the following is true.

THEOREM **C.1.** *Assume* $\|\gamma_{0,n,\ell}\|_1 = \Theta(1/n^{m_\ell - h_\ell})$ *with* $0 < h_\ell < m_\ell$ *and the* h_ℓ *are such that relative sparsity condition is satisfied. Then*

$$\sqrt{n^{m_\ell + h_\ell}}\,(\widehat{\gamma} - \gamma_{0,n,\ell}) \rightsquigarrow \mathcal{N}(0, V)$$

where $V = \text{plim}\, \frac{1}{n^{m_\ell}}(x_\ell' x_\ell)^{-1}(\frac{n^{h_\ell}}{n^{m_\ell}} x_\ell' \epsilon_\ell \epsilon_\ell' x_\ell)\frac{1}{n^{m_\ell}}(x_\ell' x_\ell)^{-1}$.

We omit the proof, which is entirely standard. We get super-consistent rates as the parameters are going to zero rapidly, but not too rapidly so that a central limit theorem still applies. Because relative sparsity applies, only a vanishing proportion of ℓ-objects are incidentally generated.

Logistic Regression. We turn to our main example where $p_\ell^n(x_{\ell,j};\gamma_\ell)$ is given by a logistic link function. In all that follows $\gamma_{0,n}$ consists of elements that are either order constant or tending to $-\infty$. The rates will be set in the assumptions.

THEOREM **C.2.** *Assume that* $\|\gamma_{0,n}\|_1 \cdot \sup_{x\in\mathcal{X}} \|x\|_\infty \lesssim h_\ell \cdot \log n^{m_\ell}$ *for* $0 \leq h_\ell < m_\ell$. *Additionally, assume that relative sparsity holds. Then*

$$J_n^{1/2}\,(\widehat{\gamma}_\ell - \gamma_{0,\ell,n}) \rightsquigarrow \mathcal{N}\,(0, I_d)\,.$$

Proof. Follows from Lemma C.3. The first hypothesis of the lemma is the same as that in Lemma C.2 and is assumed here for each ℓ. Additionally, assumption (2) of Lemma C.3 follows from relative sparsity. Relative sparsity implies that the h_ℓ are ordered such that for every ℓ share of incidentally generated ℓ-th objects goes to zero, corresponding to the number of incidentals being on the order of $O_p(z_{n,\ell} \cdot n^{m_\ell})$ in Lemma C.3. ■

This means that the rate of convergence of the parameters governing the probability is given by $\sqrt{n^{m_\ell - k_\ell}}$ where $0 < h_\ell < m_\ell$ tunes the sparsity of the model.

Example C.1. [Continued] Consider $\alpha_{0,L}^n = \log(1/n^{0.7})$ and $\alpha_{0,T}^n = \log(1/n^{1.75})$, $\beta_{0,L} = -2$ and $\beta_{0,T} = -3$. Then triangles form at order $1/n^{1.75}$ and links at order $1/n^{0.7}$. The theorem shows that all parameters have estimators that are consistent and, in the case of links, are asymptotically normally distributed at $\sqrt{n^{1.3}}$-rate and $\sqrt{n^{1.25}}$-rates (for links and triangles, respectively).

For some intuition as to why this works, first consider the case of a triangular array of n i.i.d. Bernoulli random variables distributed with probability $p_n \downarrow 0$ at a rate $\Theta(1/n^h)$ for $0 < h < 1$. Then the log odds is given by $\log \frac{p}{1-p} = \alpha_n$ where $\alpha_n = -h \log(C \cdot n)$ for some constant $C > 0$. It is easy to show by the Lindeberg-Feller central limit theorem for triangular arrays that in this case

$$\sqrt{n}\left(\frac{\widehat{p}_n - p_n}{\sqrt{p_n}}\right) \rightsquigarrow \mathcal{N}\,(0, 1)$$

provided $p_n n \to \infty$. This implies that $\sqrt{np_n}\,(\widehat{\alpha}-\alpha_n) = \sqrt{n^{1-h}}\,(\widehat{\alpha}-\alpha_n) \rightsquigarrow \mathcal{N}(0,1)$. This follows from observing that α_n will be consistent[74] and by the delta method

$$\sqrt{\frac{n}{p_n}}\,(\widehat{\alpha}-\alpha) \rightsquigarrow \mathcal{N}\left(0, \left[\partial_p \left\{\log\frac{p_n}{1-p_n}\right\}\right]^2\right) = \mathcal{N}\left(0, p_n^{-2}\right)$$

which implies $\sqrt{np_n}\,(\widehat{\alpha}-\alpha) \rightsquigarrow \mathcal{N}(0,1)$.

Next we offer an intuition for why this works with a finite set of discrete covariates. Let $\log\frac{p(x)}{1-p(x)} = \alpha_n + \beta x$ for x in some finite discrete set. It is clear that repeating the above argument delivers the same rate of convergence at every covariate value.

We now consider the general case. The data consists of a triangular array $\{(y_{i,n}, x_{i,n}) : n \in \mathbb{N}\}$ where $y_{i,n}$ is a binomial outcome governed by $p^n(x_{i,n};\gamma_{0,n})$. To conserve on notation let $q_{in} = p\left(x_{in}'\gamma_{0n}\right)$ and put $J_n = \sum_{i\leq n} q_{in}\left(1-q_{in}\right)x_{in}x_{in}'$. Under the maintained assumptions it will be the case that $\frac{n^h}{n}J_n \xrightarrow{\mathrm{P}} J$.

LEMMA **C.2.** *Assume that* $\|\gamma_{0,n}\|_1 \cdot \sup_{x\in\mathcal{X}}\|x\|_\infty \lesssim h\cdot\log n$ *for* $0 \leq h < 1$. *Then,*

$$J_n^{1/2}\left(\widehat{\gamma}_n - \gamma_{0n}\right) \rightsquigarrow \mathcal{N}(0, I_d).$$

Equivalently, the result implies that $\sqrt{n^{1-h}}\,(\widehat{\gamma}_n - \gamma_{0n}) \rightsquigarrow \mathcal{N}(0, J^{-1})$. This shows the sub-$\sqrt{n}$ rate of convergence.

Observe that in the example where $q_{in} \propto \exp(\alpha_{0n} + \beta_0 w_{in})$, then this corresponds to $\alpha_{0n} = \log(C\cdot n^{-h})$ where $0 \leq h < 1$ and some constant $C > 0$. More generally, the requirement ensures that the parameter (times covariate value) does not diverge too rapidly so that a central limit theorem can be applied.

Proof of Lemma C.2. The result is an extension of/corollary to Theorem 5.2 of Hjort and Pollard (1993). The convexity-based argument allows consistency and asymptotic normality to be argued in one step. Consider the random convex function

$$A_n(s) = \sum_{i\leq n} \log f_i\left(y_{in}, \gamma_{0n} + J_n^{-1/2}s\right) - \log f_i\left(y_{in}, \gamma_{0n}\right)$$

This is minimized by $s = J_n^{1/2}\left(\widehat{\gamma}_n - \gamma_{0n}\right)$.

This can be expressed as

$$A_n(s) = U_n's - \frac{1}{2}s's - r_n(s)$$

where[75]

$$U_n = J_n^{-1/2}\sum_{i\leq n}\left(y_{in}-q_{in}\right)x_{in} \rightsquigarrow \mathcal{N}(0, I),$$

which applies by a Lindeberg-Feller central limit theorem for triangular arrays, as $\min_x q_i(x) = \Theta(\max_x q_i(x)) = \omega(1/n)$ by hypothesis on $\gamma_{0,n}$, x_ℓ, and the Bernoulli

[74] $|\widehat{\alpha}-\alpha| = \left|\log\frac{\widehat{p}_n}{1-\widehat{p}_n} - \log\frac{p_n}{1-p_n}\right| \leq \left\{\left(\frac{1-\bar{p}_n}{\bar{p}_n}\right)\left(\frac{1}{(1-\bar{p}_n)^2}\right)\right\}|\widehat{p}_n - p_n| \lesssim_p \frac{|\widehat{p}_n - p_n|}{\bar{p}_n} = O_p\left(\sqrt{\frac{1}{np_n}}\right) \to 0.$

[75] Observe that $J_n^{-1/2} = \sqrt{n^{1-h}}\left(\frac{n^h}{n}\sum_{i\leq n} q_{in}\left(1-q_{in}\right)x_{in}x_{in}'\right)^{-1/2}$.

probability. Meanwhile

$$r_n(s) = \sum_{i \leq n} \frac{1}{6} q_i (1 - q_{in}) \cdot \eta_i \left(s' J_n^{-1/2} x_{in} \right) \cdot \left(s' J_n^{-1/2} x_{in} \right)^3.$$

The proof of Theorem 5.2 of Hjort and Pollard (1993) shows $r_n(s) \to 0$. This exploits that $\lambda_n := \max_{i \leq n} \left| J_n^{-1/2} x_{in} \right| \to 0$, which holds by the fact that the covariates live in a compact set (making clear that this isn't a tight assumption). ■

Observe that because of relative sparsity, incidental generation is small. Therefore, for most of the data the preceding result directly applies. However, for a vanishing proportion of m_ℓ-tuples, the structures are present due to incidental generation. This can be written in the notation of the preceding lemma by saying that some of our n data points are "invalid", but the probability that an observation is invalid is bounded by $z_n \downarrow 0$ at a fast enough rate. Note that relative sparsity directly implies this.

LEMMA **C.3.** *Assume the hypotheses of Lemma C.2. Assume either*

1. *every observation that is zero become independently invalid with probability at most $z_n \downarrow 0$, or*
2. *an $O_p(z_n \cdot n)$ share of observations become invalid, with $z_n \downarrow 0$.*

Then the conclusion of Lemma C.2 holds.

Proof. Clearly the second condition is implied by the first, so we only prove the former. Without loss of generality let $1, .., n^*$ denote the set of valid observations and $n^*+1, ..., n$ the valid observations. Note that n^* is random and is $O_p(z_n n)$.

$$U_n = J_n^{-1/2} \sum_{i \leq n} (y_{in} - q_{in}) x_{in} = J_n^{-1/2} \left[\sum_{i \leq n^*} (y_{in} - q_{in}) x_{in} + \sum_{n^* < i \leq n} (y_{in} - q_{in}) x_{in} \right].$$

Observe that

$$\frac{n^h}{n} \sum_{n^* < i \leq n} q_{in} (1 - q_{in}) x_{in} x_{in}' = \frac{n^h}{n} z_n n = o_p(1).$$

This implies

$$\begin{aligned} J_n^{-1/2} \sum_{i \leq n} (y_{in} - q_{in}) x_{in} &= \left[\frac{n^h}{n} \sum_{i \leq n^*} q_{in} (1 - q_{in}) x_{in} x_{in}' + \frac{n^h}{n} \sum_{n^* < i \leq n} q_{in} (1 - q_{in}) x_{in} x_{in}' \right]^{-1/2} \\ &\quad \times \sqrt{\frac{n^h}{n}} \left[\sum_{i \leq n^*} (y_{in} - q_{in}) x_{in} + \sum_{n^* < i \leq n} (y_{in} - q_{in}) x_{in} \right]. \end{aligned}$$

Thus

$$\left[\frac{n^h}{n} \sum_{i \leq n^*} q_{in} (1 - q_{in}) x_{in} x_{in}' + \frac{n^h}{n} \sum_{n^* < i \leq n} q_{in} (1 - q_{in}) x_{in} x_{in}' \right]^{-1/2} \xrightarrow{\mathrm{P}} J^{-1/2}.$$

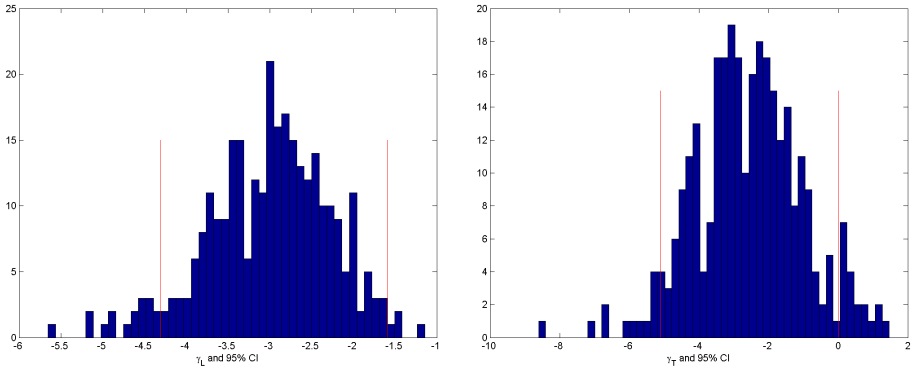
Meanwhile, we have $\sum_{i\leq n^*}(y_{in}-q_{in})\,x_{in} = O_p\left(\frac{1}{\sqrt{n^{1-h}}}\right)$ and to complete the argument

$$\begin{aligned}
\frac{1}{\sqrt{z_n n}}\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} &= \frac{1}{\sqrt{n^{1-k}}}\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} = O_p(1), \text{ where } k=h+\delta \\
\implies O\left(\frac{1}{\sqrt{n^{1-h}}}\right)\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} &= O\left(\frac{1}{\sqrt{n^{1-k+\delta}}}\right)\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} \\
&= O\left(\frac{1}{n^{\delta/2}}\cdot\frac{1}{\sqrt{n^{1-k}}}\right)\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} \\
&= O_p\left(n^{-\delta/2}\right) = o_p(1)
\end{aligned}$$

showing the result. ■

Example C.1. [Continued] Recall we have set $\alpha_L^n = \log(1/n^{0.7})$ and $\alpha_T^n = \log(1/n^{1.75})$, $\beta_L = -2$ and $\beta_T = -3$. Let $n = 100$. Then the average degree is 3.75, the average clustering is 0.14, the fraction of nodes in the giant component is 92% and the maximal eigenvalue of the adjacency matrix is 5.5. Thus, the resulting graph is comparable in structure to the empirical data.

We then run 200 simulations of this process where we generate a graph and then estimate the model parameters via sequential logistic regressions. First we regress whether a triple exists on a constant and the triad-level covariate over all $\binom{n}{3}$ observations to get $(\widehat{\alpha}_T^b, \widehat{\beta}_T^b)$, for simulation $b = 1, ..., 100$. Second, on the unused ij pairs not in triangles we regress whether a link exists on a constant and the pair-level covariate which is a logit on all $\binom{n}{2}$ observations less used pairs. From this we get $(\widehat{\alpha}_L^b, \widehat{\beta}_L^b)$ for $b = 1, ..., 100$. The results are displayed in Figure 7.



(A) Links parameter (B) Triads parameter

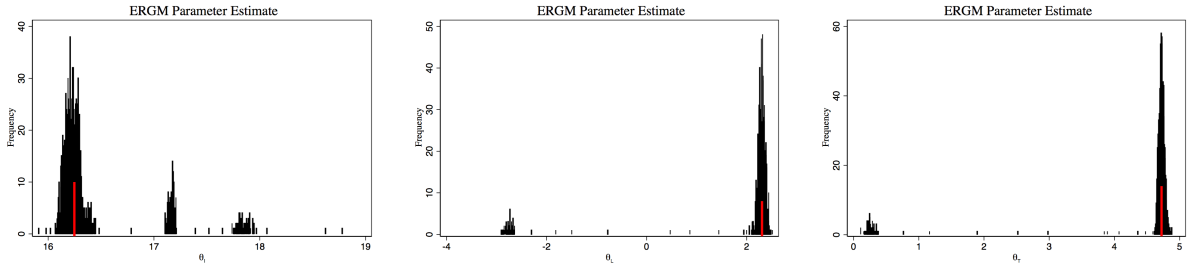
FIGURE 7. Displays the distribution of estimated parameter value as well as the median 95% confidence interval from a simple logistic regression.

We show that the parameters are correctly centered and exhibit good coverage properties.

Appendix D. Online Appendix: Isolates, Links, Triangles Example

Here we perform some additional diagnostic exercises around the Statnet ERGM estimation from Section 2.1.1.

First, we randomly generate networks that have *exactly* 20 isolates, 45 links and 10 triangles on 50 nodes. Thus, the statistics of the networks are identical, and only the location of the links and triangles changes. Any two networks with exactly the same statistics should lead to exactly the same parameter estimates as they have exactly the same likelihood under all parameter values. Thus, the only variation comes from imperfections in the software and estimation procedure. As illustrated in (Figure 8), although there is slightly less noise in the parameter estimates, they still cover similar ranges and exhibit similar features as those in Figure 1, and have similar difficulties in the standard error calculations.



(A) Isolate Parameter Estimates

(B) Link Parameter Estimates

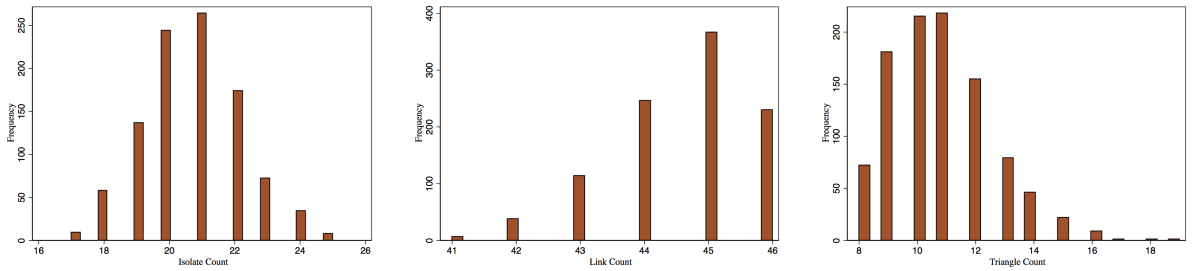
(C) Triangle Parameter Estimates

Figure 8. Standard ERGM estimation software (Statnet) output for 1000 draws of networks on 50 nodes, each having *exactly* 20 isolated nodes, 45 links, and 10 triangles. The red lines (on top of each other) are the median left and right 95 percent confidence interval lines (not capturing 95 percent of the estimates). Networks with identical statistics should lead to identical parameter estimates: all of the variation comes from imprecisions in the estimation procedure.

Second, we report the distribution of the statistics from the simulated networks (Figure 9) - they are fairly tightly clustered about the mean values.

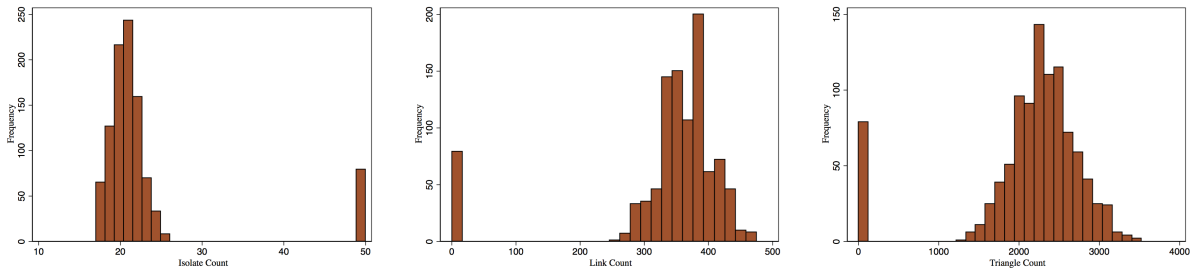
Next, for each of the 1000 simulated networks, using the parameter estimates we simulate a network using Statnet's simulation command. We then check whether the simulated networks come anywhere close to matching the original networks. Although most of the networks turn out to have nearly 20 isolates, they generally have thousands of links and triangles. Figure 10 looks nothing like the counts from the original networks (Figure 9).

Simulating a network from an ERGM is even a more daunting task than estimating parameters from one, as there is no obvious network from which to seed the simulation procedure, and again there are far too many from which to calculate likelihoods. This is another advantage of SUGMs and count SERGMs, which are easily simulated.



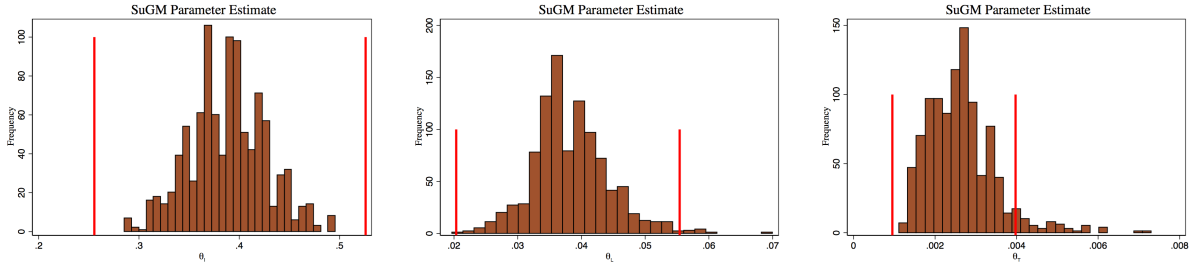
(A) Number of Isolates (B) Number of Links (C) Number of Triangles

FIGURE 9. For the 1000 simulated networks we report the distribution of the number of isolates, links and triangles.



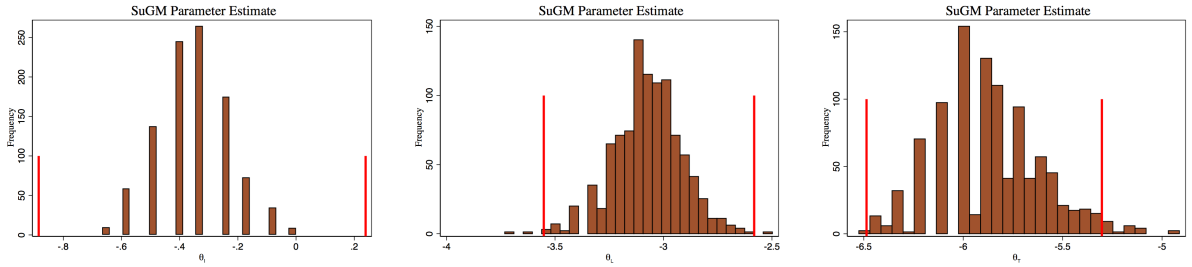
(A) Number of Isolates (B) Number of Links (C) Number of Triangles

FIGURE 10. For each of the 1000 simulated networks, using the parameter estimates from Statnet we simulate a network using Statnet's simulation command. The resulting distribution of the number of isolates, links and triangles are pictured here. They do not match the networks that generated them, which are pictured in Figure 9.



(A) Estimated p_I: $\widehat{p}_I$ (B) Estimated p_L: $\widehat{p}_L$ (C) Estimated p_T: $\widehat{p}_T$

FIGURE 11. For each of the 1000 simulated networks, we solve for estimated probabilities of isolates, links and triangles for a SUGM as defined in (2.5).



(A) Estimated θ_I: $\widehat{\theta}_I$ (B) Estimated θ_L: $\widehat{\theta}_L$ (C) Estimated θ_T: $\widehat{\theta}_T$

FIGURE 12. For each of the 1000 simulated networks, we solve for estimated SERGM parameters for isolates, links and triangles for as defined in (2.6).

APPENDIX E. ONLINE APPENDIX: ADDITIONAL CONSISTENCY RESULTS

General Consistency Results. Consider a sequence of SERGMs $(S^n, K^n_{S^n}, A_n, \beta^n)$ as defined in (3.1).

A sequence of SERGMs is *expectations-identified* with respect to a sequence of diagonal matrices $C_n > 0$ with positive diagonal entries if there exists $\gamma > 0$ such that

$$|C_n \mathrm{E}_\beta[S^n] - C_n \mathrm{E}_{\beta^n}[S^n]| > \gamma \, |\beta - \beta^n|$$

for all n.[76]

Expectations identification is an intuitive condition that requires that different parameters distinguish themselves with different means. It is a sort of minimal condition since if two different parameter values generate very similar expected statistics, then observing the realized statistic will not allow us to distinguish the parameters.

A sequence of SERGMs is *concentrated* with respect to a sequence of diagonal matrices $C_n > 0$ with positive diagonal entries if

$$C_n(S^n - \mathrm{E}_{\beta^n}[S^n]) \xrightarrow{\mathrm{P}} 0 \text{ for } \beta^n \in \mathcal{B},$$

where $\mathcal{B}$ is a set of admissible parameters.

The concentration condition requires that there is some normalization (C_n) for which the statistics will concentrate around their means. As we have not scaled statistics we have to allow for some renormalizations.[77]

Note that the choice of C_n in the following theorem links them across the two conditions. To guarantee concentration there has to exist a sequence of C_n that goes to 0 fast enough, while to guarantee expectations identification they cannot go to 0 too quickly. So, the C_n's identify the rate at which the statistics approach their means. The key to verifying consistency is then seeing whether there exists such sequences for which both conditions hold simultaneously.

[76]Here the subscript notation $\mathrm{E}_\beta[S^n]$ indicates that the expectation takes the probability to be specified by (3.1) with parameters $(S^n, K^n_{S^n}, A_n, \beta)$.

[77]Notice that the exponential term in the associated likelihood can be written $\exp(S'\beta) = \exp\left(S'C_nC_n^{-1}\beta\right)$ and we are interested in the associated parameters β, not $C_n^{-1}\beta$ which will typically trail off to infinity at polynomial rates in n.

PROPOSITION **E.1.** *If a sequence of SERGMs* $(S^n, K^n_{S^n}, A_n, \beta^n)$ *is expectations-identified and concentrated with respect to some* C_n*, then the random vectors are consistent so that*

$$|\widehat{\beta}^n(S^n) - \beta^n| \xrightarrow{\text{P}} 0.$$

The proof of Proposition E.1 is relatively routine.

It is also useful to state a ratio version of Proposition E.1 to address cases where parameters are, for instance, close to 0. Thus, we wish to have a stronger notion of consistency, not only requiring that the estimator approaches the true parameter, but that it does so in terms of a ratio. This requires corresponding definitions of concentration and identification that are ratio based. We do this in Proposition E.4.

The result is fairly tight in that a version of a converse holds as well. In particular, the following result holds.[78]

A sequence of SERGMs $(S^n, K^n_{S^n}, A_n, \beta^n)$, is *rate-expectations-identified* with rates given by a sequence of diagonal matrices $C_n > 0$ with positive diagonal entries if there exist $\gamma_H > \gamma_L > 0$ such that

$$\gamma_H|\beta - \beta^n| > |C_n \text{E}_\beta[S^n] - C_n \text{E}_{\beta^n}[S^n]| > \gamma_L|\beta - \beta^n|$$

for all n.

Rate-expectations-identification is a condition that says that the sequence of diagonal matrices C_n accurately captures the rate at which the expected statistics $\text{E}_\beta[S^n]$ nears $\text{E}_{\beta^n}[S^n]|$ as we let the vector of parameters β approach β^n.

PROPOSITION **E.2.** *If a sequence of SERGMs* $(S^n, K^n_{S^n}, A_n, \beta^n)$ *is rate-expectations-identified with rates* C_n*, then the random vectors are consistent (*$|\widehat{\beta}^n(S^n) - \beta^n| \xrightarrow{\text{P}} 0$*)* if and only if *the sequence is concentrated with respect to* β^n *and the same sequence* C_n*.*

Below, we also discuss a characterization of consistency in terms of variance of the statistics in greater detail. The argument is similar to those for standard estimators, e.g. Amemiya (1973): For consistency, we need enough variation so that the system accumulates information and concentrates around its mean. This corresponds to needing the norm of the variance matrix tending to infinity, just as we would in typical regression-like applications.

Proof of Proposition E.1. Recall that the MLE $\widehat{\beta}^n(s)$ is the β that solves

$$s = \text{E}_\beta[S^n].$$

Thus, since concentration implies that

$$C^n(S^n - \text{E}_{\beta^n}[S^n]) \xrightarrow{\text{P}} 0,$$

it follows that

$$C^n(\text{E}_{\widehat{\beta}^n(S^n)}[S^n] - \text{E}_{\beta^n}[S^n]) \xrightarrow{\text{P}} 0.$$

Given that expectations identification implies that

$$|C^n(\text{E}_\beta[S^n] - \text{E}_{\beta^n}[S^n])| > \gamma|\beta - \beta^n|$$

[78] We state a version for Proposition E.1, and the analog for ratio-consistency of the type in Proposition E.4 in the appendix is left to the reader.

for all n, it follows that

$$|\widehat{\beta}^n(S^n) - \beta^n| \xrightarrow{\mathrm{P}} 0$$

as claimed. ■

These results can be rephrased in terms of standard properties of extremum estimators and identifiable uniqueness.[79]

PROPOSITION **E.3.** *If a sequence of SERGMs $(S^n, K^n_{S^n}, A_n, \beta^n)$ satisfies concentration, rate-expectations identification, and the β^n all lie in a compact set, then $|\widehat{\beta}^n(S^n) - \beta^n| \xrightarrow{\mathrm{P}} 0$.*

Proof of Proposition E.3. Let

$$m_n(\beta) := C_n\left(S^n - \mathrm{E}_\beta[S]\right).$$

The objective function is $Q_n(\beta) := m_n(\beta)' m_n(\beta)$.

First, we want to show that the moment function satisfies a uniform law of large numbers: $\sup_{\beta\in\mathcal{B}} \|m_n(\beta)\| = o_p(1)$. By concentration, we have pointwise convergence of $m_n(\beta)$ to zero in probability. Therefore, we need to only check stochastic equicontinuity: that for every $\eta > 0$ there is a $\delta > 0$ with

$$\mathrm{P}\left(\sup_{\|\beta-\beta'\|<\delta} |m_n(\beta) - m_n(\beta')| > \eta\right) < \eta.$$

A sufficient condition (Andrews, 1994) is if a Hölder condition is satisfied:

$$|m_n(\beta) - m_n(\beta')| \leq X_n \cdot \|\beta - \beta'\|$$

where X_n is some $O_p(1)$ random variable. This is directly guaranteed by rate-expectations identification with $X_n = \gamma_H$.

Second, one can check that expectations identifiability guarantees identifiable uniqueness. Together with compactness of $\mathcal{B}$, this implies the above implies that $\widehat{\beta}$ is consistent for β^n. ■

Proof of Proposition E.2. Recall that the MLE $\widehat{\beta}^n(s)$ is the β that solves

$$s = \mathrm{E}_\beta[S^n].$$

Given Proposition E.1, we need only show that if consistency holds then concentration must also hold.

Given that rate-expectations identification implies that

$$a^n|\mathrm{E}_\beta[S^n] - \mathrm{E}_{\beta^n}[S^n]| < \gamma_H|\beta - \beta_0|$$

for all n, it follows that if if consistency holds so that

$$|\widehat{\beta}^n(S^n) - \beta^n| \xrightarrow{\mathrm{P}} 0,$$

then it must be that

$$a^n|\mathrm{E}_{\widehat{\beta}^n}[S^n] - \mathrm{E}_{\beta^n}[S^n]| \xrightarrow{\mathrm{P}} 0.$$

[79] Following, e.g., Gallant and White (1988), we say the sequence β^n is identifiably unique on $\mathcal{B}$ if

$$\liminf_{n\to\infty} \inf_{\beta\in B_{\beta^n}(\epsilon)} Q_{0,n}(\beta^n) - Q_{0,n}(\beta) > 0.$$

This implies that

$$a^n|S^n - \mathrm{E}_{\beta^n}[S^n]| \xrightarrow{\mathrm{P}} 0,$$

which implies concentration. ■

Ratio Convergence Results. A sequence of SERGMs $(S^n, K^n_{S^n}, A_n, \beta^n)$ is *ratio-expectations-identified* with respect to a sequence of diagonal C^n with positive diagonal entries if there exists $\gamma > 0$ such that

$$\left|\frac{C^n_{hh}\mathrm{E}_{\beta}[S^n_h]}{C^n_{hh}\mathrm{E}_{\beta^n}[S^n_h]} - 1\right| > \gamma \left|\frac{\beta_h}{(\beta^n)_h} - 1\right|$$

for all n and h.

A sequence of SERGMs $(S^n, K^n_{S^n}, A_n, \beta^n)$ is *ratio-concentrated* with respect to a sequence of diagonal C^n with positive diagonal entries if

$$\frac{C^n_{hh}S^n_h}{C^n_{hh}\mathrm{E}_{\beta_0}[S^n_h]} - 1 \xrightarrow{\mathrm{P}} 0$$

for each h.

PROPOSITION **E.4.** *If a sequence of SERGMs $(S^n, K^n_{S^n}, A_n, \beta^n)$ is ratio-expectations-identified and ratio-concentrated with respect to a sequence of diagonal C^n with positive diagonal entries, then $\frac{\widehat{\beta}^n(S^n)_h}{(\beta^n)_h} \xrightarrow{\mathrm{P}} 1$ for each h.*

Proof of Proposition E.4. Again, recalling that the MLE $\widehat{\beta}^n(s)$ is the β that solves

$$s = \mathrm{E}_{\beta}[S^n].$$

Thus, since ratio-concentration implies that

$$\frac{C^n_{hh}S^n_h}{\mathrm{E}_{C^n_{hh}\beta^n}[S^n_h]} - 1 \xrightarrow{\mathrm{P}} 0$$

it follows that

$$\frac{C^n_{hh}\mathrm{E}_{\widehat{\beta}^n(S^n)}[S^n_h]}{C^n_{hh}\mathrm{E}_{\beta^n}[S^n_h]} - 1 \xrightarrow{\mathrm{P}} 0.$$

Given that ratio expectations identification implies that

$$|\frac{C^n_{hh}\mathrm{E}_{\beta}[S^n_h]}{C^n_{hh}\mathrm{E}_{\beta^n}[S^n_h]} - 1| > \gamma|\frac{\beta_h}{(\beta^n)_h} - 1|$$

for all n, h, it follows that

$$\frac{\widehat{\beta}^n(S^n)_h}{(\beta^n)_h} - 1 \xrightarrow{\mathrm{P}} 0$$

or

$$\frac{\widehat{\beta}^n(S^n)_h}{(\beta^n)_h} \xrightarrow{\mathrm{P}} 1,$$

as claimed. ■

Relationship with Binomial Models.

We now discuss some conditions for consistency from the perspective of a binomial distribution. Consider the probability distribution defined over $\left\{0,...,\bar{X}_n\right\}$ by

$$\mathrm{P}_\beta\left(X_n=x\right)=\frac{K_n\left(x\right)\cdot\exp\left(\beta\cdot x\right)}{\sum_{x'}K_n\left(x'\right)\cdot\exp\left(\beta\cdot x'\right)}.$$

We can ask under what assumptions on $\left\{K_n\left(\cdot\right)\right\}_{n\in\mathbb{N}}$ does $\widehat{\beta}\xrightarrow{\mathrm{P}}\beta$? It may seem odd first that a single draw along a sequence of distributions can generate a consistent estimation. However, we observe that we can transform this distribution into one that resembles a reweighted binomial distribution:

$$\mathrm{P}_\beta\left(X_n=x\right)=\frac{\binom{\bar{X}_n}{x}\cdot\exp\left(\beta\cdot x+\lambda_n\left(x\right)\right)}{\sum_{x'}\binom{\bar{X}_n}{x'}\cdot\exp\left(\beta\cdot x'+\lambda_n\left(x'\right)\right)}, \tag{E.1}$$

where $\lambda_n(x):=\log\left(\omega_n(x)\right):=\log\left(K_n(x)/\binom{\bar{X}}{x}\right)$.

This comes from the fact that $\mathrm{P}_\beta\left(X_n=x\right)=\frac{\binom{\bar{X}_n}{x}\cdot\exp(\beta\cdot x)}{\sum_{x'}\binom{\bar{X}_n}{x'}\cdot\exp(\beta\cdot x')}$ is the distribution corresponding to the probability distribution of a binomial, $\mathrm{Bin}\left(\bar{X}_n;p=\frac{\exp\beta}{1+\exp\beta}\right)$. In turn, assumptions on $\lambda_n\left(x\right)$ will allow for consistency of the MLE of β.

LEMMA **E.1.** *$\widehat{\beta}:=\frac{\exp\frac{s_n}{\bar{S}_n}}{1+\exp\frac{s_n}{\bar{S}_n}}$ be the MLE in the above model. Assume $\omega_n\left(\cdot\right)$ is such that*

$$\mathrm{cov}^{\mathrm{Bin}\left(\bar{X}_n;p\right)}\left(x,\omega_n\left(x\right)\right)=o\left(\mathrm{E}^{\mathrm{Bin}\left(\bar{X}_n;p\right)}\left[\omega_n\left(x\right)\right]\right)$$

uniformly in a compact neighborhood of p. Then $\widehat{\beta}\xrightarrow{\mathrm{P}}\beta$.

Proof. First the normalizing constant can be written as

$$\phi\left(\beta\right)=\frac{\mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]}{\left(1-p\right)^{\bar{S}_n}}.$$

This follows from

$$\binom{\bar{S}_n}{s_n}\cdot\exp\left(\log\left(\frac{p}{1-p}\right)\cdot s_n+\lambda_n\left(s_n\right)\right)=\binom{\bar{S}_n}{s_n}\cdot\left(\frac{p}{1-p}\right)^{s_n}\times\omega_n\left(s_n\right)$$

and therefore

$$\sum_{s_n'}\binom{\bar{S}_n}{s_n'}\cdot\exp\left(\log\left(\frac{p}{1-p}\right)\cdot s_n'+\lambda_n\left(s_n'\right)\right)=\sum_{s_n'}\binom{\bar{S}_n}{s_n'}\cdot\left(\frac{p}{1-p}\right)^{s_n'}\times\omega_n\left(s_n'\right)=\frac{\mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]}{\left(1-p\right)^{\bar{S}_n}}.$$

Second, the maximum likelihood estimator solves the FOC of

$$s_n\log\left(\frac{p}{1-p}\right)+\log\left(1-p\right)^{\bar{S}_n}-\log\left(\mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]\right),$$

given by

$$0=s_n\frac{p}{p\left(1-p\right)}+s_n\frac{1-p}{p\left(1-p\right)}-\frac{\bar{S}_n}{1-p}-\frac{\partial_p\mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]}{\mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]}.$$

The parameter $\widehat{p}$ is therefore given implicitly by

$$\widehat{p} = \frac{s_n}{\bar{S}_n} - \frac{\partial_p \mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]}{\mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right]} \times \frac{1}{1-\widehat{p}}.$$

Third, observe that

$$\partial_p \mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right] = \frac{1}{p\left(1-p\right)} \mathrm{E}_p^{Bin}\left[\left\{s_n' - p\bar{S}_n\right\}\omega_n\left(s_n'\right)\right],$$

which follows from

$$\begin{aligned}
\partial_p \mathrm{E}_p^{Bin}\left[\omega_n\left(s_n'\right)\right] &= \sum_{s_n'} \binom{\bar{S}_n}{s_n'} \cdot \partial_p \left\{p^{s_n'}\left(1-p\right)^{\bar{S}_n - s_n'}\right\} \times \omega_n\left(s_n'\right) \\
&= \sum_{s_n'} \binom{\bar{S}_n}{s_n'} \cdot p^{s_n'}\left(1-p\right)^{\bar{S}_n - s_n'} \times \omega_n\left(s_n'\right) \times \left\{\frac{s_n'}{p} - \frac{\left(\bar{S}_n - s_n'\right)}{1-p}\right\} \\
&= \frac{1}{p\left(1-p\right)} \mathrm{E}_p^{Bin}\left[\left\{s_n' - p\bar{S}_n\right\}\omega_n\left(s_n'\right)\right].
\end{aligned}$$

Thus

$$\widehat{p} = \frac{s_n}{\bar{S}_n} - \frac{\mathrm{E}_{\widehat{p}}^{Bin}\left[\left\{s_n' - p\bar{S}_n\right\}\omega_n\left(s_n'\right)\right]}{\mathrm{E}_{\widehat{p}}^{Bin}\left[\omega_n\left(s_n'\right)\right]} \times \frac{1}{1-\widehat{p}}.$$

Under the assumed condition that $\sup_{p' \in B_\delta(p)} \frac{\mathrm{cov}^{Bin(\bar{S}_n;p')}(s,\omega_n(s))}{\mathrm{E}^{Bin(\bar{S}_n;p')}[\omega_n(s)]} = o\left(1\right)$, the result follows. ■

A Characterization of Consistency in terms of Variance.

We provide an alternative characterization of consistency of SERGMs in terms of the variance of the statistics.

We say that a SERGM is *expectations-identified* if $\beta' \neq \beta$ implies that $\mathrm{E}_\beta[S] \neq \mathrm{E}_{\beta'}[S]$.

For simplicity, we present the univariate case, though the multivariate case follows simply by controlling the norm of the covariance matrix instead.

PROPOSITION **E.5.** *Consider an expectations-identified SERGM.*

(1) *For any* $\varepsilon > 0$ *such that* $\varepsilon \leq \frac{\mathrm{var}_\beta[S]}{\max_{\beta' \in [\beta-\varepsilon,\beta+\varepsilon]} |k^3_{\beta'}(S)|}$,

$$\mathrm{P}_\beta\left[|\widehat{\beta}(S) - \beta| > \varepsilon\right] \leq \frac{4}{\varepsilon^2 \mathrm{var}_\beta(S)}.$$

(2) *Conversely, let* $B > 0$ *be such that* $\mathrm{var}_\beta[S] \leq B$ *and* $|S - \mathrm{E}_{\beta'}[S]| \leq B\mathrm{E}_{\beta'}[|S - \mathrm{E}_{\beta'}[S]|]$ *for all* β' *within* $\frac{1}{(B+1)B}$ *of* β. *Then*

$$\mathrm{P}_\beta^n\left[|\widehat{\beta}^n(S) - \beta| > \frac{1}{3(B+1)B}\right] > \frac{1}{2B-1}.$$

Consider a sequence of expectations-identified SERGMs indexed by the number of nodes n. We say that the sequence is consistent at β if there are sequences $\varepsilon^n \to 0$ and $r^n \to \infty$ such that

$$\mathrm{P}^n_\beta \left[|\widehat{\beta}^n(S) - \beta| > \varepsilon^n \right] < \frac{1}{r^n}.$$

COROLLARY **E.1.** *Consider a sequence of expectations-identified SERGMs indexed by the number of nodes n.*

1. *If $\mathrm{var}^n_\beta[S] \to \infty$ and $\mathrm{Skew}^n_\beta[S] \to 0$, then the sequence is consistent at β.*
2. *Conversely, if there is some $B > 0$ such that $\mathrm{var}^n_\beta[S] \leq B$ and $|S - \mathrm{E}^n_{\beta'}[S]| \leq B\mathrm{E}^n_{\beta'}\left[\left|S - \mathrm{E}^n_{\beta'}[S]\right|\right]$ for all β' within $\frac{1}{(B+1)B}$ of β for large enough n, then the sequence is not consistent.*

To see an implication of this result, consider a SERGM defined on a count of "large" subgraphs, for instance the number of components that contain some given percentage of the nodes. The variance of such a statistic is necessarily bounded, and so such a SERGM cannot be consistent.

Proof of Proposition E.5.

First, note that since $\beta' \neq \beta$ implies that $\mathrm{E}_\beta[S] \neq \mathrm{E}_{\beta'}[S]$, and also since $\mathrm{E}_\beta[S]$ is continuous in β, it must be that $\mathrm{E}_\beta[S]$ is monotone in β.

Let $\widehat{\beta}(S)$ be the MLE of β as a function of the statistic S. We know that $\widehat{\beta}(s)$ is the β such that $E_\beta[S] = s$, which varies continuously in β.

Let us consider

$$\mathrm{P}_\beta \left[|\widehat{\beta}(S) - \beta| > \varepsilon \right]$$

Note that $s'(\beta) = \mathrm{var}_\beta(S)$, and also that $s''(\beta) = \mathrm{E}_\beta\left[(S - \mathrm{E}_\beta[S])^3\right]$.

Therefore,

$$s(\beta + \varepsilon) = s(\beta) + \int_0^\varepsilon \left[\mathrm{var}_\beta(S) + \int_0^x k^3_{\beta+y}(S) dy \right] dx$$

and

$$s(\beta - \varepsilon) = s(\beta) - \int_0^\varepsilon \left[\mathrm{var}_\beta(S) + \int_0^x k^3_{\beta-y}(S) dy \right] dx.$$

Thus,

$$\min\left[|s(\beta+\varepsilon) - s(\beta)|, |s(\beta-\varepsilon) - s(\beta)|\right] \geq \varepsilon \mathrm{var}_\beta(S) - \frac{\varepsilon^2}{2}\left(\max_{\beta' \in [\beta-\varepsilon, \beta+\varepsilon]} |k^3_{\beta'}(S)| \right),$$

and by the choice of $\varepsilon \leq \frac{\mathrm{var}_\beta[S]}{\max_{\beta' \in [\beta-\varepsilon,\beta+\varepsilon]} |k^3_{\beta'}(S)|}$,

$$\min\left[|s(\beta+\varepsilon) - s(\beta)|, |s(\beta-\varepsilon) - s(\beta)|\right] \geq \varepsilon \mathrm{var}_\beta(S)/2.$$

Given the monotonicity of $E_\beta(S)$ in β, it then follows that

$$|\widehat{\beta}(S) - \beta| > \varepsilon \text{ implies } |S - \mathrm{E}_\beta(S)| > \varepsilon \mathrm{var}_\beta(S)/2.$$

Thus,

$$\mathrm{P}_\beta \left[|\widehat{\beta}(S) - \beta| > \varepsilon \right] \leq \mathrm{P}_\beta \left[|S - \mathrm{E}^n_\beta(S)| > \varepsilon \mathrm{var}_\beta(S)/2 \right].$$

By Chebychev's inequality it then follows that

$$\mathrm{P}_\beta\left[|S-\mathrm{E}^n_\beta(S)|>\varepsilon \mathrm{var}_\beta(S)/2\right]\leq \frac{\mathrm{var}_\beta(S)}{(\varepsilon \mathrm{Var}_\beta(\mathrm{S})/2)^2}=\frac{4}{\varepsilon^2\mathrm{Var}_\beta(\mathrm{S})}.$$

Therefore,

$$\mathrm{P}_\beta\left[|\widehat{\beta}(S)-\beta|>\varepsilon\right]\leq \frac{4}{\varepsilon^2\mathrm{Var}_\beta(\mathrm{S})},$$

as claimed.

Now let us examine the converse statement in the proposition.

Note that $\mathrm{var}_\beta[S]\leq B$, it follows that $\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]\leq B+1$, and therefore by assumption it also follows that $|S-\mathrm{E}^n_\beta(S)|\leq (B+1)B$.[80] It then follows that

$$\mathrm{var}^n_\beta(S)\leq |S-\mathrm{E}^n_\beta(S)|\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]\leq (B+1)B\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|],$$

and by similar reasoning

$$k^3_\beta[S]\leq |S-\mathrm{E}^n_\beta(S)|\mathrm{var}^n_\beta(S)\leq (B+1)^2B^2\mathrm{E}[|S-E^n_\beta(S)|].$$

By similar reasoning to the proof of the first part of the proposition, we can deduce that

$$\mathrm{P}_\beta\left[|\widehat{\beta}(S)-\beta|>\varepsilon\right]\geq \mathrm{P}_\beta\left[|S-E_\beta(S)|>\varepsilon\mathrm{Var}_\beta(\mathrm{S})+\varepsilon^2\left(\max_{\beta'\in[\beta-\varepsilon,\beta+\varepsilon]}|\mathrm{k}^3_{\beta'}(\mathrm{S})|\right)\right].$$

Therefore,

$$\mathrm{P}_\beta\left[|\widehat{\beta}(S)-\beta|>\varepsilon\right]\geq \mathrm{P}_\beta\left[|S-E^n_\beta(S)|>\left(\varepsilon(B+1)B+\varepsilon^2(B+1)^2B^2\right)E[|S-\mathrm{E}^n_\beta(S)|]\right].$$

For $\varepsilon=\frac{1}{3(B+1)B}$, it follows that

$$\mathrm{P}_\beta\left[|\widehat{\beta}(S)-\beta|>\varepsilon\right]\geq \mathrm{P}_\beta\left[|S-\mathrm{E}^n_\beta(S)|>\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]/2\right].$$

Given that $|S-\mathrm{E}^n_\beta(S)|<B\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]$, it follows that there exists $\delta=\frac{1}{2B-1}$ such that

$$\mathrm{P}_\beta\left[|S-\mathrm{E}^n_\beta(S)|>\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]/2\right]\geq\delta.$$

establishing the second part of the proposition.

To see this last claim, let $p=\mathrm{P}_\beta\left[|S-\mathrm{E}^n_\beta(S)|>\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]/2\right]$ and note that

$$\begin{aligned}(1-p)\mathrm{E}\left[|S-\mathrm{E}^n_\beta(S)|\;\middle|\;|S-E^n_\beta(S)|\leq \mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]/2\right]\\ +p\mathrm{E}\left[|S-\mathrm{E}^n_\beta(S)|\;\middle|\;\|S-\mathrm{E}^n_\beta(S)|>\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]/2\right]\\ =\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]\end{aligned}$$

and so, since $|S-\mathrm{E}^n_\beta(S)|<B\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]$, it follows that

$$(1-p)\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]/2+pB\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]\geq \mathrm{E}[|S-\mathrm{E}^n_\beta(S)|].$$

Therefore, $p\geq\delta=\frac{1}{2B-1}$, as claimed. ■

[80] To see that $\mathrm{E}[|S-\mathrm{E}^n_\beta(S)|]\leq B+1$, note that for a nonnegative random variable X,

$$\mathrm{E}[X]\leq \mathrm{P}[X\leq 1]1+\mathrm{P}[X>1]\mathrm{E}[X|X>1]$$

and

$$\mathrm{P}[X>1]E[X|X>1]<\mathrm{P}[X>1]\mathrm{E}[X^2|X>1]\leq \mathrm{E}[X^2],$$

and the claim follows.

Proof of Corollary E.1.

The second claim follows directly from Proposition E.5.

To see the first claim, let us consider two cases. First consider a case such that $Var^n_\beta(S)/|k^3_\beta(S)^n|$ is bounded away from 0. In that case apply the first part of Proposition E.5, with $\varepsilon^n = (Var^n_\beta(S))^{-a}$ for any $a < 1/2$. Next consider a case such that $\mathrm{var}^n_\beta(S)/|k^3_\beta(S)^n| \to 0$. In that case, set $\varepsilon^n = \frac{\mathrm{var}^n_\beta(S)}{|k^3_\beta(S)^n|}$. In that case, we need to check that $\frac{(\mathrm{var}^n_\beta(S))^3}{|k^3_\beta(S)^n|^2} \to \infty$, which is equivalent to $\mathrm{Skew}^{\mathrm{n}}_\beta(\mathrm{S}) \to 0$. ■

Appendix F. Online Appendix: Extension of Table 1

Here we present an extension of the analysis in Table 1. Instead of simply controlling for "close" versus "far" links on the dimensions of caste and GPS, we allow for a considerably richer specification. The goal here is to show that even when we control, flexibly, for a rich set of covariates, a link-based model exploiting the observable homophily is unable to replicate key features of observed networks. To do this, we estimate a link-based model within each village using the following vector of controls:

- Geographic distance between households,
- Square of geographic distance between households,
- Households are of different caste,
- Difference in number of rooms household has,
- Square of difference in number of rooms,
- Difference in number of beds,
- Square of difference in number of beds,
- Difference in quality of electricity,
- Square of difference in quality of electricity,
- Difference in latrine quality,
- Square of difference in latrine quality,
- Whether or not both households have the same status in terms of owning or renting their house.

We use a logistic regression for this estimation.

The estimated a vector of regression coefficients for each village capture how characteristics of a dyad correspond to linking probabilities. This gives a predicted probability that each household is linked to each of the other households in the village. We use these predicted probabilities to generate 100 simulated networks per village and study the characteristics of the resulting networks. These are presented in column [3] of Table 3.

Table 3. Estimation of Additional Models: Extension of Table 1

		Data	Link-based model with covariates	Link-based model with extended covariates	SUGM with links and triangles	SUGM with isolates, links and triangles
		[1]	[2]	[3]	[4]	[5]
Models are fit to different combinations of these statistics.	Number of Unsupported Links	160.8	236.2	236.2	161.2	161.8
	Number of Triangles	39.2	3.1	3.1	39.7	39.5
	Average Degree	2.3243	2.3260	2.3234	2.5916	2.5219
	Number of Isolates	54.9722	25.7222	27.3750	31.4444	65.9167
None of the models are directly fit to any of these statistics.	Average Clustering	0.0895	0.0105	0.0134	0.1268	0.0829
	Fraction in Giant Component	0.7061	0.8315	0.8082	0.7982	0.6718
	First Eigenvalue	5.5446	3.8578	4.0746	4.6762	5.3025
	Spectral Gap	0.9550	0.3354	0.3728	0.6684	1.0617
	Second Eigenvalue of Stochastized Matrix	0.9573	0.9632	0.9642	0.9559	0.9069
	Average Path Length	4.6921	5.6565	5.5407	5.1215	4.1180

Notes: Column [1] presents the average value of various network characteristics across the 36 villages. Columns [2], [3], [4] and [5] present simulation results. In a simulation we first estimate parameters of a given model for a given village and then randomly draw a graph from the model with the estimated parameters. We run 100 simulations for each of the villages for each of the models and average across the simulations, and the entries report these averaged across the villages.

Column [3] contains the statistics from the enriched link-based model, while the remainder of the table is exactly the same as what is presented in the body of the paper. Adding over 12 parameters to flexibly control for demographic attributes makes almost no difference in generating network characteristics that match the observed data, providing very small improvements, and still not coming close to doing as well

as the simple SUGMs. Moreover, since the specification developed here makes use of considerably richer data than those used in the two candidate SUGM models, it suggests that by decomposing a network into a tapestry of random structures (triangles, links and even isolates), considerable value is added in modeling higher order features of networks in a parsimonious way.