A NETWORK FORMATION MODEL BASED ON SUBGRAPHS

ARUN G. CHANDRASEKHAR† AND MATTHEW O. JACKSON*  

ABSTRACT. We develop a new class of random-graph models for the statistical estimation of network formation that allow for substantial correlation in links. Various subgraphs (e.g., links, triangles, cliques, stars) are generated and their union results in a network. The challenge in estimating the frequencies with which subgraphs ‘truly’ form is that subgraphs can overlap and may also incidentally generate new subgraphs, and so the true rate of formation of the subgraphs cannot generally be inferred just by counting their presence in the resulting network. We provide estimation techniques for recovering the rates at which the underlying subgraphs were formed from the observation of a single (large) network. We provide results on identification of the true underlying rates of subgraph formation from various statistics, as well as a new Central Limit Theorem for correlated random variables that establishes asymptotic normality for our estimators. We also show that if the network is sparse enough then direct counts of subgraphs are consistent and asymptotically normal estimators. We illustrate the models with applications.  

JEL CLASSIFICATION CODES: D85, C51, C01, Z13.  

KEYWORDS: Subgraphs, Random Networks, Random Graphs, Exponential Random Graph Models, Exponential Family, Social Networks, Network Formation, Consistency, Sparse Networks

Date: October 2015.

This grew out of a paper: “Tractable and Consistent Random Graph Models,” (http://arxiv.org/abs/1210.7375), which we have now split into two pieces. This part contains the material on subgraph generation models and includes new results on identification, asymptotic normality, and estimation via generalized method of moments that were not part of the original paper. We thank Alberto Abadie, Isaiah Andrews, Aureo de Paula, Han Hong, Bryan Graham, Guido Imbens, Elena Manresa, Michael Leung, Paul Goldsmith-Pinkham, Joe Romano, and Elie Tamer for helpful discussions as well as various seminar participants. We thank Andres Drenik for valuable research assistance. Chandrasekhar is grateful for support from the NSF Graduate Research Fellowship Program and NSF grant SES-1156182. 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; member of the NBER.  

‡Department of Economics, Stanford University; external faculty member of the Santa Fe Institute; fellow of CIFAR.
1. Introduction

For a researcher interested in an economic or social interaction, understanding how networks form is essential for at least three reasons. First, evaluating welfare and efficiency questions require understanding why people chose particular relationships. Second, any analysis of peer effects or social learning must account for the endogeneity of the network, which implies that their positions in the networks may be driven by their characteristics in ways that can be estimated (e.g., see Goldsmith-Pinkham and Imbens (2013)). Third, policy evaluation has to account for potential changes in social structure. This enormous scope of settings in which such endogenous network issues arise, has driven a literature on network formation.\footnote{See Jackson (2008) for references and background.} Nonetheless, we still lack a general and flexible way of estimating network formation that is tractable, reliably consistent, and admits nontrivial correlations in link formation.

The challenges with statistical estimation of network formation are that (i) the relationships in a network are interdependent (often, this comes from externalities across relationships, which are what make networks interesting), and (ii) researchers usually observe just one network in a given study. Although a network consists of many links, those links fail to provide independent observations. Despite the lack of independence, it is possible that the many relationships in a network still provide rich enough information to consistently estimate a network model and test of hypotheses from a single observed network, at least hypothetically. Here we develop a class of models that admit substantial correlations in links and also provide practical techniques of estimating the models, showing that they are easily estimable even when a researcher only has one network.

Before discussing our approach, let us discuss some of the other approaches that are available.

A most basic approach is to use what are known in the random graph and social networks literature as ‘stochastic block models’, in which links may depend on node characteristics but are (conditionally) independent of each other. That approach assumes away the correlation between links that is present and a main question of interest in many social and economic networks, and so not well-suited for applications beyond community detection. Thus, stochastic block models do not offer a solution to the problem of correlated network relationships.\footnote{See Section 7.1 for an empirical analysis showing that a model allowing for correlated links substantially outperforms a stochastic block model in matching several key network characteristics.}

There are also a set of random formation network models where one explicitly models a meeting process and a link formation algorithm (e.g., Barabasi and Albert (1999); Jackson...}
and Watts (2001); Jackson and Rogers (2007); Currafini, Jackson, and Pin (2009, 2010); Christakis, Fowler, Imbens, and Kalyanaraman (2010); Bramoullé, Currafini, Jackson, Pin, and Rogers (2012)), which can be estimated. These approaches are useful in some contexts, but they are not designed nor intended for the general statistical testing of a wide variety of network formation models and hypotheses - such as doing statistical tests of fundamental hypotheses like, e.g., triadic closure (are links correlated across triples of nodes - so that if two people have a friend in common, are they more likely to be friends with each other than if link formation were independent).

Given this void, a literature spanning several disciplines (sociology, statistics, economics, and computer science) has turned to exponential random graph models (ERGMs) to meet these challenges. ERGMs incorporate such interdependencies and thus have become the workhorse models for estimating network formation. However, recently these models have come under fire as maximum likelihood and Bayesian estimators of the parameters may not be computationally feasible nor consistent, and so the software being used often provides inaccurate parameter estimates, except in cases in which the models are degenerate.

This lack of a flexible set of models that are computable and can be used across many applications for hypothesis testing and inference is what motivates our work here.

What we do is develop a new class of random-graph models for the statistical estimation of network formation that allow for substantial correlation in links. In these models, various subgraphs (e.g., links, triangles, cliques, stars) are generated directly. For instance, students may form friendships with their roommate(s), members of a study group, teammates, band members, etc.; researchers may form collaborations on writing papers in pairs, or triples, or quadruples, etc. This results in links, and those links are then naturally correlated since they are formed in combinations. The union of all these subgraphs results in a network. The challenge to the researcher is that often only the final network is observed: a survey may ask people to list their friends and acquaintances, or links may be observed on a social platform, or emails or phone calls are observed, and so forth; but the original formation process is often not observed. The challenge that then arises in estimating how the network formed is that subgraphs may overlap and may also incidentally generate new subgraphs, and so the true rate of formation of the subgraphs cannot generally be inferred just by counting their presence in the resulting network.

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 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)). That approach, however, is not designed for identifying the incidence of particular patterns of network relationships that may be implied by various social or economic theories of the type that we wish to address here.

For details see Bhamidi, Bresler, and Sly (2008); Shalizi and Rinaldo (2012); Chandrasekhar and Jackson (2012).
Despite the fact that the formation can only be inferred, there are fairly simple conditions for identification, as different rates of generation for subgraphs lead to different observed network characteristics. Effectively, one can estimate the frequencies at which various subgraphs should appear in the final network based on their formation rate. So, we provide estimation techniques for recovering the frequencies at which the underlying subgraphs were formed from the observation of a single (large) network. We provide results on identification of the true underlying parameters governing subgraph formation from various statistics.

Beyond the identification issue, for the models to be useful in hypothesis testing and inference, we also need to provide results on the asymptotic distributions of the estimates. Existing central limit theorems from the spatial and time-series econometrics literatures do not apply to our setting, as we need to allow subgraphs to form on arbitrary groups of nodes, which then results in correlation patterns across all links in the network. In particular, the standard arguments exploiting strong mixing of random variables do not apply since there is no sense in which the random variables we are concerned with begin to become arbitrarily far from each other, and therefore essentially uncorrelated. Thus, we use a powerful lemma from Stein (1986) in order to prove a new central limit theorem for correlated random variables that provides for more general and permissive results than previously available for our setting. This establishes asymptotic normality for our estimators, and should be useful beyond our network setting. They may be of independent interest as they have a connection to the study of central limit theorems for random variables described by dependency graphs (Baldi and Rinott (1989); Goldstein and Rinott (1996); Chen and Shao (2004)).

Finally, we also show that if the network is sparse enough then direct counts of subgraphs are consistent and asymptotically normal estimators - providing a very easy estimation technique for many network applications, as many social and economic networks are relatively sparse.

We conclude the paper with some illustrative applications and extensions of the models, showing how one can apply the models to cases where the nodes have continuous-valued characteristics that influence the formation of subgraphs and showing that these models provide much better fits of some network data than alternative models that ignore the link correlations.

Since the first writing of this, the literature has expanded (beyond the specific dynamic models mentioned above) to include other models providing for practical, statistical estimation of network formation (e.g., Mele (2013); Boucher and Mourifié (2012); Badev (2013); Goldsmith-Pinkham and Imbens (2013); Sheng (2013); de Paula, Richards-Shubik, and Tamer (2014); Leung (2014); Graham (2014)). An excellent recent overview is presented by de Paula (2015). To deal with the complications faced when links in a network are correlated, this literature has taken several approaches: (i) view correlations as driven by
unobserved heterogeneity, but have links be uncorrelated conditional on all true (observed
and unobserved) characteristics (Graham (2014)); (ii) consider a model in which nodes only
link to other nodes that are close enough in some geographic or characteristic space, so that
links that involve distant enough pairs of nodes become independent (e.g., Boucher and Mou-
rifé (2012); Leung (2014)); (iii) take advantage of the relationship between certain classes of
strategic network formation models and potential games (Butts (2009); Mele (2013); Badev
(2013)); (iv) derive restrictions on parameters of an observed network under the presumption
that it is in equilibrium (pairwise stable) (de Paula, Richards-Shubik, and Tamer (2014)).

Each of these approaches comes with advantages and disadvantages, and our approach is
quite distinct from all of them, both in terms of the fundamentals of the approach (working
with subgraphs as the basic building block) and the technicalities of allowing nontrivial
conditional correlations (developing a new central limit theorem for non-trivially correlated
random variables). Our contribution is to develop models of network formation that admit
considerable interdependency, and have the presence of links be highly correlated - even across
distances, but still prove consistency and asymptotic normality of the parameter estimates.

2. An Example

We begin with a canonical example that illustrates our family of models of network for-
mation. The goal is to help the reader understand how the general class of models work
before we get into the notation. We carry the example through the paper to illustrate the
results, even though the models and results apply much more generally.

2.1. Links and Triangles.

In this leading example, nodes form not only bilateral relationships but may also form tri-
adic relationships (triangles) directly. This leads to obvious forms of correlations in links. For
example, suppose that links represent productive relationships (so, for instance, researchers
coadjoint papers either as a pair or in triples). An individual gets value from each ‘team’
that he or she joins based on the characteristics of the nodes in the group, and the size
of the group. Moreover, groups are allowed to overlap. For instance, i could co-author a
paper directly with j, and then also co-author a different paper with j and k. As another
example, two students i and j who express a friendship, might be part of a club and also a
study-group or be roommates in a double or triple room, etc. Similarly, two villagers might
be part of a self-help-group together but also engage in other activities bilaterally. Or, two
tennis partners may also be in a band in a trio (de Paula (2015)). Thus, people could be
involved in multiple and partly overlapping groups with the same other node(s).

In this example, some links form directly, and the set of such links, ij, that form is denoted
\( L \). A *triangle* is a set of three nodes, ijk, that form together as a group, and the set of such
triangles that form is denoted \( T \).
To emphasize, \( i \) may form multiple relationships with another node \( j \), for instance the link \( ij \), but also the triangles \( ijk, hij \), and so forth.

The simplest version of this example is one in which nodes have no characteristics, and links form independently with some probability \( \beta_{0,L} \in [0, 1] \) and triangles form independently with probability \( \beta_{0,T} \in [0, 1] \).

The overall network that forms is

\[
g = \{ ij : ij \in \mathcal{L} \text{ or } \exists k \text{ s.t. } ijk \in \mathcal{T} \},
\]

so that two agents are connected if they are part of a link or both are part of any triangle. The observed network \( g \) is a projection and the researcher only observes this single, network.

The existence of each link is correlated with many others - any others with whom it could be part of a triangle - and this invalidates estimation techniques that treat links as independent. Moreover, if we slightly enrich the example, to include cliques of size 4 instead of size 3 (triangles), then all links are correlated with each other - not just ones that have nodes in common. This complete lack of independence, together with the fact that the researcher often only observes one network, present nontrivial estimation challenges. These challenges can be overcome as we now discuss in detail.

### 2.2. Estimation Issues.

The researcher’s goal is to use the observed data to recover the parameters of interest, for example, the \((\beta_{0,L}, \beta_{0,T})\) in our leading example. If the researcher observed the sets \( \mathcal{L} \) and \( \mathcal{T} \) directly, then estimation would be straightforward. Indeed, in some instances a researcher might have information on all the various groups a given individual is involved in: for instance in the case of a co-authorship network, the researcher may observe all the papers a researcher has written. However, often researchers only have summary binary information concerning whether each two individuals in the society interact: for instance, the adjacency matrix of a network indicating which pairs of agents are ‘friends’ based a survey, or from observing that they are friends on a social platform, or from observing that they send emails to each other, text or phone each other, etc. The researcher might even have multigraph information regarding several types of interactions, but still in binary form - not observing how the relationships originally formed, or why they are maintained - which could also involve correlation.

Thus, the general problem is that the formation of the subgraphs is not directly observed, and so must be inferred in order to estimate the parameters of interest. From the perspective of the researcher, the observed network \( g \) is a projection of \( \mathcal{L} \) and \( \mathcal{T} \). For example, if

\[
g_{ij}g_{jk}g_{ik} = 1,
\]
is it the case that $ijk$ formed as a triangle, or that $ij$, $jk$ and $ik$ formed as links, or that $ij$ and $jk$ formed as links and $ik$ formed as part of a different triangle $ikm$, or some combination of these or other combinations?

Figure 1 provides an illustration.

![Diagram of network formation](image)

**Figure 1.** The network that is formed and eventually observed is shown in panel D. The process comes from forming triangles independently with probability $\beta_{0,T}$ as in (B) in green; and also forming links, in grey, independently with probability $\beta_{0,L}$ as in (C). New links are dashed while links that overlap with some link also formed in a triangle are in solid and bold. We see that there is both (i) overlap as some links coincide with links already in triangles, as well as (ii) extra triangles that were generated ‘incidentally’. Given that we only observe the resulting network in panel D, we need to infer the formation of the different subgraphs carefully and not simply by directly counting observed links and triangles.

This presents a challenge for estimating a parameter related to triangle formation since some of the observed triangles were directly generated in the formation process, and others were “incidentally generated;” and similarly, it presents a challenge to estimating a parameter for link formation since some truly generated links end up as parts of triangles. It could also
be that the link 12 formed directly or as part of the triangle 124 or both. This would not be observed either, and so we face several challenges in estimating the number of directly generated links.⁵

Despite this challenge, the model parameters can generally be identified, as we prove below. For instance, as the probabilities of links and triangles vary so do the properties of the expected networks, and by looking at a large enough network, one can consistently recover the parameters of interest. To understand how, note that as one varies \( \beta_{0,L}, \beta_{0,T} \), the relative rates of overall observed links and triangles change, as do the number of triangles that overlap with each other. One can calculate the relative rates at which incidental links and triangles are expected to be generated, and there is an invertible relationship between observed counts of links and triangles, and the underlying rates at which they were expected to be directly formed. Moreover, this can be done via estimators that are easily computed and have nice asymptotic normality properties despite the fact that the links are all correlated. Thus, beyond the identification problem, we also prove a new central limit theorem for correlated random variables that could be correlated in ways that do not satisfy standard mixing assumptions used in time series or spatial econometrics.

2.3. Outline.

The remainder of the paper proceeds as follows.

- We define subgraph generation models (Section 3).
- We then show how identification works in these models and discuss generalized methods of moments approaches to estimation, pointing out how these estimators are easily computed (Section 4).
- Next, we show that these estimators also satisfy an asymptotic normality result so that inference and hypothesis testing are possible (Section 5). This involves developing and proving a new central limit theorem, building on a lemma from Stein (1986), for correlated random variables (which should also be useful beyond the networks application).
- We then show that in the case of relatively sparse networks, identification and estimation come directly from observed counts of subgraphs.
- We conclude with some empirical illustrations and applications, including utility-based network formation.

3. Subgraph Generation Models

We now provide the general definition of subgraph generation models (SUGMs).

⁵One could view this as an issue of measurement error with correlation.
\( n \geq 3 \) is the number of nodes on which a network is formed. In some cases, nodes have characteristics that we denote by \( x_i \) for a generic \( i \in \{1, \ldots, n\} \).

We denote a network by \( g \), the collection of subsets of \( \{1, \ldots, n\} \) of size 2 that lists the edges or links that are present in its graph. So, \( g = \{\{1,3\},\{2,5\}\} \) indicates the network that has links between nodes 1 and 3 and between nodes 2 and 5. For notational ease, we simply write \( g = \{13,25\} \), and write \( ij \in g \) to denote that link \( ij \) is present in network \( g \).

In general our model easily accommodates directed graphs, and all of the definitions below extend directly, in which case instead of pairs of nodes, these would be ordered pairs so that \( ij \) and \( ji \) would differ. However, for ease of exposition, most of the examples and discussion below refer to the undirected case.

Let \( \mathcal{G}^n \) denote the set of all networks on \( n \) nodes.

In a SUGM, subgraphs are directly generated, and then the resulting network is the union of all of the links in all of the subgraphs. Degenerate examples of this are Erdos-Renyi random networks, and the generalization of that model, stochastic-block models, in which links are formed with probabilities based on nodes’ attributes. The more interesting classes of SUGMs include richer subgraphs, and hence involve dependencies in link formation. It might be that people of the same caste meet more frequently or are more likely to form a relationship when they do meet, as in a stochastic block model, but it could also be that groups of three (or more) 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, where subgraphs can be based on the subgraph shape as well as the nodes’ characteristics.

SUGMs are formally defined as follows.

There are finitely many types of nonempty subgraphs, indexed by \( \ell \in \{1, \ldots, k\} \), on which the model is based - for instance in the links and triangles case \( \ell \in \{L, T\} \). The \( k \) subgraph types are denoted by \((G_\ell)_{\ell \in \{1, \ldots, k\}}\), where each \( G_\ell \subset \mathcal{G}^n \) is a set of possible subgraphs on \( m_\ell \leq n \) nodes. Each subgraph in \( g' \in G_\ell \) is homomorphic to (a relabeling of) every other one in \( g'' \in G_\ell \) (so for any \( g', g'' \in G_\ell^n \), there exists a bijection \( \pi \) on \( \{1, \ldots, n\} \) such that \( ij \in g' \) if and only if \( \pi(i)\pi(j) \in g'' \)). One can also place additional restrictions based on node characteristics, for instance, requiring that the characteristics \( X_i \) and \( X_{\pi(i)} \) be the same – which would be the case if one wanted to defined subgraphs as “triangles that involve one child and two adult nodes”. As an example, the set \( G_\ell \) for some \( \ell \) could be all triangles, and for another \( \ell \) could be all stars with one central node and four other nodes, and another \( \ell \) could be all of the links that involve people of different castes, and so forth. These could

---

\[6\text{This definition does not admit isolates since we define subgraphs to be nonempty and connected, but isolates are easily admitted with notational complications, and are illustrated in some of our supplementary material and examples.}\]
also be directed subgraphs in the case of a directed network. A few examples are pictured in Figure 2.

![Figure 2. Examples of subgraphs. Links could be directed or undirected.](image)

Another ingredient is a vector of parameters, $\beta_0 \in \mathcal{B}$, where $\mathcal{B}$ is (unless otherwise noted) a compact subset of some finite dimensional Euclidean space. For instance, $\beta_0 = (\beta_{0L}, \beta_{0T}) \in [0, 1]^2$ in the links and triangles example. In some applications, the parameters have the same dimension as the number of types of subgraphs, although this is not necessary. For example, the $\beta_0$’s may be preferences parameters of agents who choose to form subgraphs based on their own and neighbors characteristics and the shape of the subgraph (e.g., see Sections 7.2 and 8.1); and some preference parameters (e.g., how much agents like interacting with other agents who have similar characteristics) may influence the formation of more than one type of subgraph.

A network $g$ on $n$ nodes is randomly formed as follows:
- Each of the possible subnetworks $g_{\ell} \in G_{\ell}$ is independently formed with probability a probability $p_{\ell}(\beta_0)$ for each $\ell \in \{1, \ldots, k\}$. In some examples, each $p_{\ell}$ is synonymous with parameter $\beta_{0\ell}$.
- The resulting network, $g$, is the union of all the links that appear in any of the generated subgraphs.

3.1. An Example with Node Characteristics.

To make things very simple, let us consider the case of discrete characteristics.

Suppose that nodes come in two colors: blue and red. In our example of links and triangles, there are now three types of links: (blue, blue), (blue, red), (red, red); and four types of triangles (blue,blue,blue), (blue,blue,red), (blue,red,red), (red,red,red).

In this case, the set of possible $\ell$’s would be: \{(blue, blue), (blue, red), (red, red), (blue,blue,blue), (blue,blue,red), (blue,red,red), (red,red,red)\}.

So,

$$G_{\text{(blue,blue)}} = \{ij : x_i = \text{blue}, x_j = \text{blue}\}$$
and
\[ G_{\text{blue,blue,red}} = \{ijk : x_i = \text{blue}, x_j = \text{blue}, x_k = \text{red}\}, \]
and so forth.

The parameters
\{β₀_{\text{blue,blue}}, β₀_{\text{blue,red}}, β₀_{\text{red,red}}, β₀_{\text{blue,blue,blue}}, β₀_{\text{blue,blue,red}}, β₀_{\text{blue,red,red}}, β₀_{\text{red,red,red}}\};
could be the probabilities that the subgraphs in question form.

One could restrict or enrich the model by having simpler or more complex sets of parameters – for instance requiring that β₀_{\text{blue,blue}} = β₀_{\text{red,red}}, or by having preference parameters that govern the probabilities of various subgraphs forming, as we discuss in Section 7.1.

This example is one with a finite set of possible node characteristics. It is also easy to accommodate continuous node characteristics. For instance, we can let \( p_T(x_i, x_j, x_k; β₀) \) be the probability of a triangle but be governed by the characteristics of the nodes in a logistic or probit formulation. We discuss this in Section 9.

4. IDENTIFICATION AND ESTIMATION OF SUGMS

We now discuss identification and estimation of SUGM parameters, given that the researcher observes a single network \( g \).

4.1. Arrays.

While identification does not require this, when we discuss the consistency and asymptotic normality of parameter estimates, we consider what happens as the number of nodes becomes ‘large’,\(^7\) so that we can say something about convergence of the estimates to the true parameters. So, to describe how parameter estimates behave as a function of the number of nodes \( n \), is useful to consider a sequence of distributions governed by parameters indexed by \( n \) and study the asymptotic behavior of estimators of parameters along the sequence. This approach is standard in the random graphs literature (e.g., see the classic reference Bollobas (2001)).

Research on social networks has long observed that parameters need to adjust with the number of nodes. For example, friendship networks among a small set of agents (say 50 or 100) and large set of agents (thousands or much more) often have comparable average degrees (number of links per node).\(^8\) As a concrete example, consider friendships among high school students in the U.S. based on the Add Health data set (e.g., see (Currrarini,\(^8\)

\(^7\)Generally, rates of convergence will be relatively fast in \( n \) as this leads to an order of \( n^2 \) link observations, and even more observations of larger subgraphs. Rates of convergence are given explicitly in the results below.

\(^8\)See Chandrasekhar (2015) for examples networks of varying size ranging from village network data in sub-saharan Africa or India to university dorm friendship network data which all exhibit somewhat comparable number of links per node.
Jackson, and Pin, 2009, 2010)). There are some high schools with only 30 students and others with around 3000 students. The average degree is ranges between 6 and 8 over the high schools, the link probability shrinks dramatically with \( n \): from roughly \( \frac{6}{30} \) to roughly \( \frac{8}{3000} \). Thus, irrespective of the size of their school, students have numbers of friends of roughly the same order of magnitude, and so the frequency of friendship formation must decrease with \( n \).

Thus, the network has a distribution of the form: \( g^n \sim P ( \cdot | \beta^n_0 ) \).

As a specific example, in order to match the high school friendship data, it could be the case that the probability that a given link forms \( \beta^n_{0,L} \) would satisfy \( \beta^n_{0,L} = \frac{h_{0,L}}{n^{h_L}} \) for some \( h_L \) and the probability that a given triangle forms \( \beta^n_{0,T} \) would satisfy \( \beta^n_{0,T} = \frac{h_{0,T}}{n^{h_T}} \) and we are interested in estimating the parameter vector \( \beta^n_0 = (\beta^n_{0,L}, \beta^n_{0,T})' \).

With this in mind, the goal is to find an estimator \( \hat{\beta}^n \) for \( \beta^n_0 \) for which

\[
\delta (\beta^n, \beta^n_0) \overset{P}{\rightarrow} 0
\]

for a metric \( \delta \). We discuss the choice of a metric in Section 4.3 and the asymptotic distribution of the normalized estimator in Section 5.

### 4.2. Generalized Method of Moments.

Having observed the network, the parameters, \( (\beta^n_{0,1}, \ldots, \beta^n_{0,L}, \ldots, \beta^n_{0,k}) \in \mathcal{B} \), are estimated based on some vector of statistics of the network, \( S^n(g) = (S^n_1(g), \ldots, S^n_k(g)) \). We set the dimension of the statistics equal to the number of parameters for notational convenience, but it is not needed in any of the results that follow, except in illustrative examples in which we are explicit that the \( \beta^n_{0,\ell} \) are the rates of formation of a particular subgraph. This vector might keep track of, for instance, the number of links that exist, the number of all triangles that are present, or the number of links that do not lie in any triangles, or the number of links that are involved in more than one triangle, and so forth. It may be that the statistics that are most useful to track, or are available, do not directly correspond to the counts of the subgraphs in question. Indeed, as we shall see below, sometimes identification has to be achieved via statistics that are not direct counts of the subgraphs in question.

The GMM (Generalized Method of Moments) estimators associated with some vector of statistics \( S^n(g) \) are denoted by \( \hat{\beta}^n \) and are defined as follows.

For each \( \ell \in \{1, \ldots, k\} \), we allow for some normalizing rate \( r^n_\ell (g) \) (e.g., normalizing the count statistics by some function of their variance), and we let

\[
\overline{S}^n_\ell (g) := \frac{S^n_\ell (g)}{r^n_\ell} \quad \text{and} \quad \overline{S}^n (g) := (\overline{S}^n_1(g), \ldots, \overline{S}^n_k(g)).
\]

The objective function to be minimized is

\[
Q^n(\beta, g) := \left( \overline{S}^n (g) - \mathbb{E}_\beta[\overline{S}^n (g)] \right)' W^n \left( \overline{S}^n (g) - \mathbb{E}_\beta[\overline{S}^n (g)] \right),
\]
for some positive definite weighting matrix $W^n$.

The GMM estimators are defined by

$$\hat{\beta}^n(g) := \arg\min_{\beta \in B} Q^n(\beta, g).$$

As in Pötscher and Prucha (1997) and Jenish and Prucha (2009), we also define an expected version of the objective function:

$$Q^n(\cdot) := E_{\beta_0} \left( S^n(g) - E_{\beta} [S^n(g)] \right)^T W^n E_{\beta_0} \left( S^n(g) - E_{\beta} [S^n(g)] \right).$$

Note that $Q^n(\beta_0^n) = 0$ and so $Q^n(\cdot)$ is minimized at $\beta_0^n$.

The following lemma is a variation on a standard consistency theorem for GMM estimators, here allowing parameters to vary with $n$.

**Lemma 1** (Pötscher and Prucha (1997); Jenish and Prucha (2009)). If

1. the weighting matrices $W^n \xrightarrow{p} W$ for a positive semi-definite $W$,
2. the functions $E_{\beta} [S^n(g)]$ are continuous in $\beta \in B$,
3. the parameters $\beta^n_0$ are identifiably unique in the sense that for any $\epsilon > 0$
   $$\liminf_{n \to 0} \left[ \inf_{\beta \in B, \, \delta(\beta, \beta^n_0) > \epsilon} \left| Q^n(\beta) - Q^n(\beta^n_0) \right| \right] > 0,$$
4. and, there is uniform convergence of the objective functions:
   $$\sup_{\beta \in B} \left| Q^n(\beta, g) - Q^n(\beta) \right| \xrightarrow{p} 0,$$

then $\delta(\hat{\beta}^n, \beta^n_0) \xrightarrow{p} 0$.

This follows from Pötscher and Prucha (1997), Lemma 3.1.

Our moment functions will generally be measurable and continuous in the parameters, so (1) will easily be satisfied with SUGMs. With care in picking the rate functions, $r^n$, a uniform law of large numbers applies to the objective function, and so (3) can generally be satisfied. Thus, if identifiable uniqueness (2) holds, then consistency follows. The core challenge, then, is making sure that our SUGMs satisfy an identifiable uniqueness condition.

In particular, by looking at some features of the resulting network – the share of links that are present, the share of triangles that are present, etc. – is it possible to recover that parameter values that likely generated the network? For pedagogical reasons, we consider several examples that illustrate how to proceed in prominent cases of interest and are easily extended: (a) links and triangles, and (b) links and stars (in Appendix C).

**4.3. A Choice of Metric.** The choice of metric $\delta$ is important in ensuring that we really recover the parameters of interest in a useful way. To see why, consider the simplest possible model that just has links and no characteristics (an Erdos-Renyi random network), and consider an application like the Add Health example mentioned above, so that the average
degree is steady as \( n \) grows. In that case, the probability that any given link forms, \( \beta_{0,L}^n \), would satisfy \( \beta_{0,L}^n = \frac{b_{0,L}}{n-1} \). Thus, a very uninteresting estimator for which \( \left\| \hat{\beta}_L^n - \beta_{0,L}^n \right\|_{\ell_1} \xrightarrow{P} 0 \), would be to set \( \hat{\beta}_L^n = 0 \). We want to rule out such a degenerate estimator as being consistent. Thus, we adjust the metric to normalize the difference require that the estimate really approximates \( \beta_{0}^n \) in a strong sense - not only in distance, but in rate too. For example, if we set

\[
(4.2) \quad \delta(x, y) := \max_{\ell} \left[ \frac{|x_\ell - y_\ell|}{\max(|x_\ell|, |y_\ell|)} \right],
\]

then the requirement becomes

\[
\max_{\ell} \frac{\left| \hat{\beta}_\ell^n - \beta_{0,\ell}^n \right|}{\max(\beta_\ell^n, |\beta_{0,\ell}^n|)} \xrightarrow{P} 0,
\]

which requires that \( \hat{\beta}_\ell^n \) and \( \beta_{0,\ell}^n \) be proportional to each other far enough along the sequence. Thus, if \( \beta_{0}^n \) approaches 0, saying that \( \hat{\beta}_\ell^n \) is a good estimate of it also requires that \( \hat{\beta}_\ell^n \) approach 0 at the same rate, which is a much stronger conclusion.\(^9\)

Hence, we take (4.2) as the metric, unless otherwise specified.

4.4. Identification and Consistent Estimation with Links and Triangles.

We show how we can derive the identification results needed to apply Lemma 1 to our leading example of SUGM based on links and triangles, where links are formed with probability \( \beta_{0,L} \) and triangles with probability \( \beta_{0,T} \).

Let

\[
S_L(g) = \frac{\sum_{i<j} g_{ij}}{\binom{n}{2}}
\]

be the fraction of links (whether isolated or in a triangle) in the network \( g \), and

\[
S_T(g) = \frac{\sum_{i<j<k} g_{ij}g_{jk}g_{ik}}{\binom{n}{3}}
\]

be the fraction of possible triangles in the network \( g \).

Clearly, to prove results on how estimation works as a function of the number of nodes requires that the sequence have some logic to it, and which is captured as follows.\(^11\) We

\(^9\)We take 0/0 = 0.

\(^10\)To see why this is better, consider the degenerate estimator of \( \hat{\beta}_\ell^n = 0 \). In this case \( \delta(0, \beta_{0,\ell}^n) = \frac{n_{\ell}^n \left| 0 - \beta_{0,\ell}^n \right|}{b_{0,\ell}^n} = \frac{|0 - b_{0,\ell}^n|}{b_{0,\ell}^n} = 1 \) which does not tend to zero - so the \( \delta \) metric tells us that this is not a good estimator, but if we just worked with standard distance as our metric, then \( \left| 0 - \beta_{0,\ell}^n \right| \rightarrow 0 \).

\(^11\)Formulating things in this manner makes proofs more transparent, and it could be done in many ways - just remapping the function relative to the \( n \) of interest. Note that we can also allow the constants \( b_{0,L}^n, b_{0,T}^n \) be indexed by \( n \) and have \( b_{0,L}^n, b_{0,T}^n \in [D, \overline{D}]^2 \) for some compact set, as we show in the appendix, where the proof is given for this more general formulation.
say that an array of parameters is well-sequenced if \( \beta^n_0 = (\beta^n_0, L, \beta^n_0, T) = (\frac{b_0}{n^L}, \frac{b_0}{n^T}) \) where the \( h_L, h_T \) parameters allow for degrees to grow or shrink at some rate relative to \( n \).

**Proposition 1.**

- A SUGM of links and triangles is identified for any \( \beta_L, \beta_T \). That is, if \( \beta'_L, \beta'_T \neq \beta_L, \beta_T \) then \( E_{\beta'_L, \beta'_T}[S_L(g), S_T(g)] \neq E_{\beta_L, \beta_T}[S_L(g), S_T(g)] \).
- If a sequence of SUGMs is well-sequenced, with \( h_L > 1/2, h_T \in [h_L + 1, 3h_L) \), then \( \beta^n_{0,L}, \beta^n_{0,T} \) are identifiably unique.

Identification points out that changing the parameter vector, for a fixed \( n \), changes the expected counts of links and triangles. Indentifiable uniqueness is more demanding, as it requires that the expected counts change sufficiently that one can statistically distinguish parameter vectors with high probability. It is guaranteed under some weak requirements, as \( h_L > 1/2, h_T \in [h_L + 1, 3h_L) \) admit most social applications. The restrictions on \( h_T \) are needed since if \( h_T < h_L + 1 \) then almost all links are formed by triangles rather than directly and so the link parameter becomes hard to accurately estimate, and if \( h_T > 3 \) then triangles become so rare that the triangle parameter becomes hard to estimate. It is possible to also get identifiable uniqueness for the very dense network case in which \( h_L \leq 1/2 \), but that relies on a different identification strategy (using statistics other than just counts of links and triangles - also counting two-stars) that we describe in Section D in the Supplementary Appendix.

To further understand the identification, consider Figure 3. Each configuration involves two triangles, but the bottom one with only five links is relatively more easily incidentally formed than the upper one. Thus, by looking at the combination of how many triangles and how likely links there are, we can begin to sort out relative rates of the two parameters.

![Figure 3.](image-url)
We provide more details behind the argument. First, in this case one can show via some direct probability calculations that (see the proof of Proposition 1 for this derivation, as the link expression may not be obvious):

\[ E_{\beta_L, \beta_T} [S_L(g), S_T(g)] = \left[ \beta_T + (1 - \beta_T)q^n_L, \beta_T + (1 - \beta_T)(q^n_L)^3 \right], \]

where

\[ q^n_L = \beta_L + (1 - \beta_L)(1 - (1 - \beta_T)^{n-3}) \]

is the probability that a link forms when some particular triangle that it could be part of does not form directly. Note that the terms \( \beta_T + (1 - \beta_T)(q^n_L)^3 \) is the probability that a triangle forms, either directly \( \beta_T \) or does not form directly, \( 1 - \beta_T \) but then each of the links then forms on its own \( (q^n_L)^3 \).\(^{12}\) The term for the estimation of links has to be written carefully to account for the possibility of a link forming as part of a triangle, and there are various ways to write the expression. This particular expression is useful in the proof of identifiable uniqueness, since it is easy to compare it to and this distinguish it from the triangle expression, showing how different parameters thus lead to different rates of formation of links and triangles.

In this case, GMM is straightforward\(^{13}\) and just involves finding the \( \hat{\beta}_L \) and \( \hat{\beta}_T \) that solve

\[ E_{\hat{\beta}_L, \hat{\beta}_T} [S_L, S_T] = (S_L(g), S_T(g)) \]

for the observed \( S_L, S_T \), where the left hand side is given by (4.3).

More generally, beyond links and triangles, for more complicated SUGMs for which we may not have closed form for the GMM equation, one can use simulated method of moments, or various standard iterative techniques to obtain the estimates.

### 4.4.1. Simulations

We further illustrate the consistency of the estimators of links and triangles example via some simulations.

We set \( n = 350 \) and run 500 simulations of generating a network under the SUGM and then calculating the estimates for each of 20 parameter values for \( (\beta^n_{0,L}, \beta^n_{0,T}) \):

\[ \beta^n_{0,L} = \frac{b_0}{n} \quad \text{and} \quad \beta^n_{0,T} = \frac{3b_0}{n^2} \]

with \( b_0 \in \{.5, 1, ..., 9.5, 10\} \). This generates networks with expected degrees ranging between 2 and 50 and for parsimony guarantees that on average the number of links generated directly and by triangles are similar.

In Figure 4 we plot three different things:

---

\(^{12}\)Conditional upon the triangle not forming directly, the links are then independent.

\(^{13}\)Note that in this case the model is ‘just identified’, and so the GMM estimator can be obtained for any positive definite weighting matrix, and so we can work with the S’s for the S’s via a reweighting.
(1) the true parameters,
(2) estimates \((\beta_L, \beta_T)\) based on the GMM estimation outlined above, and
(3) estimates \((\tilde{\beta}_L, \tilde{\beta}_T)\) simply directly set equal to the share of subgraphs.

While (1) is the truth and (2) is the GMM estimator above, (3) plots the value of a direct estimator counting the frequency of triangles and links not in triangles. If incidental generation of subgraphs is sufficiently low, the naive estimator also does well.

![Figure 4](image_url)

**Figure 4.** The true parameters are the diamonds, the GMM estimates are the circles, and the ‘naive’ estimates from direct frequency counts of triangles and links (not in triangles) are the stars. The GMM estimators are so accurate that it is difficult to even distinguish the circles from the diamonds since they lie on top of each other. The direct counts are accurate for the sparse case, but not for the dense case, while the GMM estimators are accurate for all cases.

In Figure 4 it is difficult even to distinguish the GMM estimates from the true parameters since they lie directly on top of each other. The diamonds are the true parameters and the circles are the GMM estimates. In contrast, the estimates from just using direct counts of triangles and links (not in triangles) is highly accurate for relatively low degree, but then not accurate once the degree becomes large. In the sparse domain both estimators do well. However, as we increase the density of links and triangles, there is an increase rate of incidental triangles and fewer links that form outside of triangles, we can clearly see that the direct estimator over-estimates triangle formation, \(\beta_{0,T}^n\), as it does not account for incidental triangle formation, and underestimates \(\beta_{0,L}^n\). The GMM estimator resolves these issues and recovers the true parameters in the non-sparse case.

4.5. **An Example with Links and \(K\)-stars.**
A similar identification result applies to an example with links and \( K \)-stars (a subgraph where there is one ‘center’ node that is linked to \( K \) other ‘peripheral’ nodes - the top right graph in Figure 2 is a 4-star). \( K \)-stars become useful in estimating networks that have core-periphery structures or fat-tailed degree distributions. This appears in the Supplementary Appendix, Section C.

5. Asymptotic Normality

While results on identification and consistency establish that parameter estimates are consistent in large enough networks, we also need to determine the asymptotic distributions of the estimates so that we can do inference and hypothesis testing. Thus, we now show that our parameter estimates, appropriately normalized, are asymptotically normally distributed. The hurdle is that the observations (e.g., links, and other subgraphs) are correlated.

To establish the asymptotic distribution we first show that various subgraph counts are asymptotically normally distributed, and then a standard argument on GMM estimation (e.g., Newey and McFadden (1994)) implies that the same holds for the parameter estimates. For instance, for our example with links and triangles, to establish asymptotic normality of the parameter estimates the key is to first show that

\[
\frac{SL(g) - E_{g0}[SL(g)]}{\sigma_L^n} \rightsquigarrow N(0, 1),
\]

and

\[
\frac{ST(g) - E_{g0}[ST(g)]}{\sigma_T^n} \rightsquigarrow N(0, 1),
\]

where \((\sigma_L^n)^2 := \text{var}(SL(g))\) and \((\sigma_T^n)^2 := \text{var}(ST(g))\).

The challenge is that the subgraphs are correlated: links \( ij \) and \( jk \) are correlated since triangle \( ijk \) could have formed. Similarly, triangle \( ijk \) is correlated with \( ijl \) and even with \( ist \) (as they all could be partly generated by \( ijs \)). Going even further, in a model with links and four cliques, any two links, \( ij \) and \( kl \), are correlated for any \( i, j, k, l \), since a four clique including both them could have been generated. Similarly, any four cliques on nodes \( i, j, k, l \) and \( u, v, w, z \) are correlated for arbitrary indices.

Existing central limit theorems that allow for correlated random variables do not apply for our setting as they require require a spatial/ordered lattice structure (e.g., Bolthausen (1982)). In the typical logic of central limit theorems based on strong mixing arguments in the spatial and time series literature, random variables are embedded in some space where there are “close” and “far” random variables and the further they are, the less correlated they are. Other researchers (e.g., Boucher and Mourifié (2012); Leung (2014)) working on network formation exploit these spatial techniques by embedding nodes in some space so that only “nearby” nodes can link and “distant” nodes cannot link in order to satisfy mixing conditions and apply a central limit theorem like Bolthausen (1982). As \( n \to \infty \) most nodes get further
and further apart and therefore essentially never link. The reason this is unsatisfying for our purposes is that such a strategy imposes a specific structure on the adjacency matrix: it has to be nearly block-diagonal. To see this, consider the simple case where nodes live on a line. Then in the adjacency matrix, only nodes within some limited distance to the left or right of any given node tend to be linked. While this may be fine for certain contexts, it is not an adequate description of a village network where there is no natural space on which some households in a village should be considered, ex ante, to be infinitely far apart (or students in a dorm should be considered, ex ante, to be infinitely unlikely to link to each other).

This is why the existing central limit theorems do not apply for our setting. Our modeling approach employs no such “nearby” versus ”distant” approach, and is novel to the literature. In order to manage this we prove a central limit theorem for correlated random variables which cannot be embedded in a spatial setting. The closest work is the literature on dependency graphs, discussed below.

5.1. A Central Limit Theorem for Correlated Random Variables.

We require some new notation, which we later apply to SUGMs.

Let \( \{X^N_\alpha : \alpha \in \Lambda^N\} \) be an array of random variables taking on values in \([0, 1]\). Here \( \alpha \in \Lambda^N \) is the set of labels, and the index is such that \(|\Lambda^N| = N\). For instance, in our SUGM settings the \( X_\alpha \) may be an indicator of the appearance of some particular subgraph, such as a link or triangle, the \( \alpha \) would track the pairs of nodes involved in a potential link \((ij)\) or triples of nodes in a triangle \((ijk)\), and \(N\) captures the \( \binom{n}{2}\) possible links or \( \binom{n}{3}\) possible triangles.

Let

\[
S^N := \sum_{\alpha \in \Lambda^N} \left( X^N_\alpha - \mathbb{E}\left[X^N_\alpha\right] \right).
\]

We provide conditions under which a normalized statistic

\[
\frac{S^N}{a_N^{1/2}} \sim \mathcal{N}(0, 1),
\]

where the normalizer, \(a_N\), is a measure of the variance of \(S^N\).

5.1.1. Stein’s Lemma.

Our proof uses a lemma from Stein (1986). We review it here, both to be self-contained and also to explain why this approach to proving asymptotic normality is useful and distinct from other approaches in the networks literature.

The key observation of Stein (1986) is that if a random variable satisfies

\[
\mathbb{E}[f'(Y) - Yf(Y)] = 0
\]

for every \(f(\cdot)\) that is continuous and continuously differentiable, then it must have a standard normal distribution.
This observation leads to a useful lemma, that allows one to characterize the Kolmogorov distance between a random variable $Y$ and a standard normally distributed $Z$, denoted $d_K(Y, Z)$. We can bound this from above by (a constant times) the Wasserstein distance, $d_W(Y, Z)$, which is the supremum of the above expression taken over a class of bounded functions with bounded first and second derivatives.

**Lemma 5.1 (Stein (1986); Ross (2011)).** If $Y$ is a random variable and $Z$ has the standard normal distribution, then

$$d_W(Y, Z) \leq \sup_{\{f: \|f\|, \|f''\| \leq 2, \|f'\| \leq \sqrt{2/\pi}\}} |E[f'(Y) - Y f(Y)]|.$$ 

Further

$$d_K(Y, Z) \leq (2/\pi)^{1/4} (d_W(Y, Z))^{1/2}.$$ 

By this lemma, if we show that a normalized sum of random variables satisfies

$$d_W(S^N, Z) = \sup_{\{f: \|f\|, \|f''\| \leq 2, \|f'\| \leq \sqrt{2/\pi}\}} |E[f'(S^N) - S^N f(S^N)]| \to 0,$$

then it must be asymptotically normally distributed.

Arguments based on Stein (1986), and his precursor work, Stein (1972), have been used to derive central limit theorems in two literatures: spatial statistics and dependency graphs. For example, for sequences of $\alpha$-mixing random variables in the spatial statistics literature, the oft-used Bolthausen (1982) central limit theorem uses a lemma from Stein (1972) to show asymptotic normality. In the spatial statistics literature, a standard structure would have $\Lambda^N \subset Z^d$ and this set is thought of as growing outwards with $N$. Then as $\Lambda^N$ expands, the $\alpha$’s that are added along the sequence are increasingly far apart, and correlations vanish with distance so that $X_\alpha$ is only correlated with $X_\eta$ that are close to $\alpha$ in the lattice. This does not work for our purposes because our setting has no such spatial structure, and we wish to allow any nodes to link to each other in our networks.

The dependency graph approach is closer to our approach and eliminates such spatial structure. Instead, collections of random variables are represented on a graph, where a link between two indices mean that they are correlated and no link means they are independent. The normalized sum is then asymptotically normally distributed provided that the dependency graph is sufficiently sparse (Baldi and Rinott (1989); Goldstein and Rinott (1996); Chen and Shao (2004)). However, previous results in that literature place overly-restrictive conditions on how various $X_\alpha$’s can be correlated across $\alpha$. For instance, we want models where in principle all links can be ex ante correlated, and in overlapping ways. Even previous results allowing for high- and low-correlation dependency sets are far too stringent to apply to our setting (Ross, 2011; Goldstein and Rinott, 1996; Chen and Shao, 2004).
Our approach can be thought of as extending the dependency neighborhoods approach. In principle every \( X_\alpha \) and \( X_\eta \) can be correlated, but we separate those into more highly and less highly correlated sets, and we still obtain a central limit theorem under fairly weak conditions that bound the total and relative correlations in these sets. For instance, in our context, the highly correlated sets for our example with triangles would be triangles that share a node or an edge, and the less correlated sets would be triangles that have no nodes in common.

5.1.2. Dependency neighborhoods.

For each \( \alpha, N \), we partition the index set of other random variables, \( \Lambda^N \), into two pieces. In particular, we define a set, called a dependency neighborhood, for each \( \alpha, N \):

\[
\Delta(\alpha, N) \subset \Lambda^N \text{ such that } \alpha \in \Delta(\alpha, N).
\]

The conditions for \( \eta \in \Delta(\alpha, N) \) are precisely defined below. We need \( \Lambda^N \) to be partitioned into \( \Delta(\alpha, N) \) sets for each \( \alpha \) in a specific manner to satisfy a few sufficient conditions.

Intuitively, this set includes the \( X_\eta \)'s that have relatively “high” correlation with \( X_\alpha \), and generally its complement includes the \( X_\eta \)'s that have relatively “low” correlation with \( X_\alpha \). There is substantial freedom in defining these sets, but an easy rule to applying them to (non-sparse) SUGMs is to set the \( \Delta(\alpha, N) \) sets to include the other subgraphs with which the subgraph \( \alpha \) shares some edges and could have potentially been incidentally generated.\(^{14}\)

We show that under conditions on the relative correlations inside and outside of the dependency neighborhoods, a central limit theorem applies.

5.1.3. The Central Limit Theorem.

We consider sequences such that the distribution over the \( X_\alpha \)'s is symmetric (although not necessarily exchangeable).\(^{15}\)

Let

\[
a_N := \sum_{\alpha, \eta \in \Delta(\alpha, N)} \text{cov}(X_\alpha, X_\eta),
\]

be the total sum of covariances across all the pairs of variables in each other's dependency neighborhoods, and let

\[
S^N := \frac{S^N}{a_N^{1/2}}
\]

be the normalized statistic.

\(^{14}\)In the sparse case, one can set \( \Delta(\alpha, N) = \alpha \), as in Corollary 1.

\(^{15}\)Symmetry is the requirement that for any \( \alpha \) and \( \alpha' \), there exists a permutation of labels that maps \( \alpha \) to \( \alpha' \) and leaves the distribution unchanged. For instance, the marginal distribution of any link \( ij \) is similar to the distribution of any other link \( kl \), and we can find a permutation of labels for which the joint distributions over this link and all other links are the same. This does not, however, mean that exchangeability holds, as the joint distribution of \( ij \) and \( jk \), is not the same as the distribution over \( ij \) and \( rs \). Note also that this does not preclude allowing for node characteristics, as those will be encoded into the specification of \( \alpha \).
In what follows, we maintain that $a_N \to \infty$, as otherwise there is insufficient variation to obtain a central limit theorem.

The following are the key conditions for the theorem:

\begin{align}
(5.1) \quad & \sum_{\alpha, \eta, \gamma \in \Delta(\alpha, N)} \mathbb{E}[X_{\alpha}X_{\eta}X_{\gamma}] = o \left( a_N^{3/2} \right), \\
(5.2) \quad & \sum_{\alpha, \alpha', \eta \in \Delta(\alpha, N), \eta' \in \Delta(\alpha', N)} \text{cov} \left( (X_{\alpha} - \mu)(X_{\eta} - \mu), (X_{\alpha'} - \mu)(X_{\eta'} - \mu) \right) = o \left( (a_N)^2 \right), \\
(5.3) \quad & \sum_{\alpha, \eta \not\in \Delta(\alpha, N)} \text{cov} \left( X_{\alpha}, X_{\eta} \right) = o \left( a_N \right), \text{ and} \\
(5.4) \quad & \mathbb{E}[(X_{\alpha} - \mu)(X_{\eta} - \mu)|X_{\eta}] \geq 0 \text{ for every } \alpha, \eta \not\in \Delta(\alpha, N). 
\end{align}

Even though $\mathbb{E}[(X_{\alpha} - \mu)(X_{\eta} - \mu)|X_{\eta}] \geq 0$ in most applications (as subgraphs either incidentally generate each other or don’t overlap at all, but do not tend to interact negatively), we can do without the condition - it is used to provide a simpler statement of (5.3).\footnote{In the appendix, we prove a stronger version of the theorem with a combined version of (5.3) and (5.4) that only requires that $\sum_{\alpha, \eta \not\in \Delta(\alpha, N)} \mathbb{E}[(X_{\alpha} - \mu)(X_{\eta} - \mu) \cdot \text{sign}(\mathbb{E}[(X_{\alpha} - \mu)|X_{\eta}](X_{\eta} - \mu))] = o(a_N)$; without requiring the nonnegative conditional covariance, $\mathbb{E}[(X_{\alpha} - \mu)(X_{\eta} - \mu)|X_{\eta}] \geq 0$ for every $\alpha, \eta \not\in \Delta(\alpha, N)$. See (A.14).}

Condition (5.3) is an intuitive one that states that covariances between subgraphs outside of each other’s dependency sets have a lower order of covariance than within the dependency sets. Essentially, this just captures that dependency sets are properly defined and therefore the variance of the sum is captured by the sum of variances and covariances within dependency sets.

Conditions (5.1) and (5.2) are conditions that limit the extent to which there are dependencies between more than two subgraphs at a time, requiring that these be of lower order than interactions between two at a time. As we show below, these are satisfied by basic examples of SUGMs. Some such conditions are clearly needed since excessive dependence leads to a failure of a central limit theorem, and these extend the literature sufficiently to cover our SUGMs, which were not covered before, except for degenerate cases.

**Theorem 1.** If (5.1)-(5.4) are satisfied, then $S^N \sim N(0,1)$.

It is useful to consider the special case in which $\Delta(\alpha, N) = \{\alpha\}$, which still extends and nests many standard central limit theorems. This corollary is particularly useful when we get to the case of sparse networks, where incidental networks are unlikely and the correlation between different subgraphs becomes small.

**Corollary 1.** If $\mathbb{E}[(X_{\alpha} - \mu)(X_{\eta} - \mu)|X_{\eta}] \geq 0$ for every $\eta \neq \alpha$, and\footnote{Condition (i) can be weakened to $N^{-1/3}\mu^2 = o(\text{var}(X_{\alpha}))$, as shown in the appendix.}
(i) \(\text{var}(X) \geq \mu^2 N^{-1/3+\varepsilon}\) for some \(\varepsilon > 0\) and large enough \(N\),
(ii) \(\sum_{\alpha \neq \eta} \text{cov}((X_\alpha - \mu)^2, (X_\eta - \mu)^2) = o\left(N^2 \text{var}(X)\right),\) and
(iii) \(\sum_{\alpha \neq \eta} \text{cov}(X_\alpha, X_\eta) = o\left(N \text{var}(X)\right),\)

then \(S^N \sim N(0,1)\).

Moreover, if the \(X_\alpha\)'s are Bernoulli random variables and have \(E[X_\alpha] \to 0\), then (ii) is implied by (iii).

Note that (ii) is often satisfied whenever (iii) is, and so this is an easy corollary that is based on two intuitive conditions: the variance of the variable in question cannot vanish too quickly (as there needs to be enough variation/information about the variables to get convergence), and the covariance between variables cannot be too large. An application of this corollary is given in Section B of the Supplementary Appendix, and the proof of the ‘Moreover’ statement appears there.

We outline the steps of the proof of Theorem 1, which follow techniques pioneered by Stein (1972, 1986) (see also Bolthausen (1982); Baldi and Rinott (1989); Ross (2011)) adapted to our setting, and a detailed proof appears in the appendix.

Recall that the crux of the proof requires showing that

\[
|E\left[ S f \left( S \right) \right] - E\left[ f \left( S \right) \right]| \tag{5.5}
\]

tends to zero.

In working with (5.5) it is useful to break it into pieces, and so it is useful to define the (normalized) sum over the terms not in the dependency neighborhood:

\[
S_\alpha := \sum_{\eta \notin \Delta(\alpha,N)} (X_\eta - \mu) \, / \, a_N^{1/2}.
\]

In order to see how we show (5.5), let us start with the \(E\left[ S f \left( S \right) \right]\) term:

\[
E\left[ S f \left( S \right) \right] = E \left[ \frac{1}{a_N^{1/2}} \sum_\alpha (X_\alpha - \mu) \cdot f \left( S \right) \right] \quad \text{(by definition)}
\]

\[
= E \left[ \frac{1}{a_N^{1/2}} \sum_\alpha (X_\alpha - \mu) \left( f \left( S \right) - f \left( S_\alpha \right) \right) \right] + E \left[ \frac{1}{a_N^{1/2}} \sum_\alpha (X_\alpha - \mu) \cdot f \left( S_\alpha \right) \right].
\]

The second line contains two pieces. The second piece is handled by (5.3) and several lemmas that appear in Appendix A. So, then we have to bound the first part, which we do via a Taylor expansion. We add and subtract a term containing \(f(S_\alpha)\), which is useful for a Taylor expansion (below). This generates an extra term, which in the usual dependency
graph literature is assumed to be zero, which cannot be used for our purposes. We allow for modest but nontrivial amounts of correlation in these terms and still establish the result.

Now using the first expression from the above, we can rewrite our crucial expression as

\[ |E[SF(S)] - E[f'(S)]| \leq E \left[ \frac{1}{\sigma_N^{1/2}} \sum_{\alpha} (X_\alpha - \mu) \left( f(S) - f(S_\alpha) - (S - S_\alpha) f'(S) \right) \right] \]

bounded by 2nd order remainder

\[ + E \left[ \frac{1}{\sigma_N^{1/2}} \sum_{\alpha} (X_\alpha - \mu) (S - S_\alpha) f'(S) - f'(S) \right] + o(1). \]

Both (5.1) and (5.2) are then used to show that these terms go to zero as \( N \to \infty \). Specifically, the first term, which is bounded by a second order remainder term, involves expectations over products of triples and therefore is controlled by (5.1). The second term can be factored to show that it involves the terms of the form in (5.2). These sorts of terms arise in the dependency graph literature as well as the spatial statistics literature as well, but our bounds on them are looser and more permissive, which turns out to be critical for our applications.

5.2. Applying the Theorem to SUGMs.

We now show how to apply this central limit theorem to SUGMs.

5.2.1. Moving from Statistics to Parameters.

To begin with the application, note that the Central Limit Theorem that we have proven applies to the statistics: e.g., counts of links, triangle, \( K \)-stars, four-cliques, etc. We need to show that this also applies to the parameter estimates themselves. This conversion from statistics to parameters does not really depend on the details of our particular Central Limit Theorem or the networks setting, but follows a fairly standard argument (e.g., see Newey and McFadden (1994)), adapted to our setting. We include this for completeness.

Consider a SUGM with \( k \) types of subgraphs. Recall that

\[ S^n_\ell (g) = \frac{S^n_\ell (g)}{r^n_\ell}. \]

We set \( r^n_\ell = N_\ell(n) \cdot pe(\beta^m_\ell) \),\(^18\) where \( pe(\beta^m_\ell) \) is the probability of a given subgraph forming under the true vector of parameters \( \beta^m_\ell \) (which we assume is differentiable in \( \beta \)) and \( N_\ell(n) \) is the number of possible subgraphs of type \( \ell \).\(^19\)

\(^18\) Although the researcher will generally not know \( r^n_\ell \), the GMM estimation can be accomplished without this normalization by standard two-step or other iterative procedures, and this rate provides the exact rate of convergence to normality.

\(^19\) If there are no node characteristics and the subgraph was symmetric then this would just be \( \binom{n}{m_\ell} \), but that is a special case. A triangle has \( \binom{3}{3} \) but a two-star has \( 3 \times \binom{3}{3} \) because each node of the triple could be the center.
Let us consider SUGMs that are well-balanced, so: $r^n_\ell = \Theta(r^n_{\ell'})$ for every $\ell, \ell'$. This means that the relative frequencies of different sorts of subgraphs are roughly balanced, and a link has the same order of probability of being generated by any type of subgraph. We index these so that $r^n_h = \Theta(n^h)$ for some $h$, which is now independent of $\ell$. For example, one case of links and triangles where the probabilities are $\beta^n_{0,L} = b_{0,L}/n$ and triangles are $\beta^n_{0,T} = b_{0,T}/n^2$. In this case, $r^n_2 = \left(\frac{n}{2}\right)\beta^n_{0,L} = (n-1)b_{0,L}/2 = \Theta(n)$ and $r^n_1 = \frac{n-1}{2}(n-2)b_{0,T}/6 = \Theta(n)$ so they are of the same rate, and this is an example in which $h = 1$.

It is useful to define some terms. The parameter sequence is given by $\beta^n_{0,\ell} = \frac{b_{0,\ell}}{n^h}$, where $b_{0,\ell} \in B_\ell$, where $B_\ell$ a compact subset of $\mathbb{R}_{++}$. Then let $R_n := \text{diag} \left\{ n^{h_\ell} : \ell \in [k] \right\}$ and

$$M^n(\beta^n_0) := \nabla_{\beta=\beta^n_0} \left( \mathbb{E}^{\bar{S}^n} - \mathbb{E}^\beta \left[ \mathbb{S}^n \right] \right) R_n^{-1} = -\nabla_{\beta=\beta^n_0} \mathbb{E}^\beta \left[ \mathbb{S}^n \right] R_n^{-1},$$

with $M := \lim_{n \to \infty} M^n$.

Additionally, it is useful to normalize the sequence of parameters appropriately by the number of effective observations. $R_n$ tracks the (inverse of) the rates at which the true parameter sequence tends to zero. $M^n$ is an appropriately normalized matrix of gradients of the moment function, evaluated at the true parameter value. Under our assumption of balance, $M^n$ converges uniformly to $M$ over the parameter space. We abuse notation and write $M^n(\beta^n) = M^n(R_n^{-1}b)$, and sometimes simply write it as a function of $b$, denoted $M^n(b)$.

**Lemma 2.** Consider a sequence of well-balanced and well-sequenced SUGMs with balancing rate $n^h$ and the GMM estimators $\hat{\beta}^n$ associated with weighting matrices $W^n \xrightarrow{P} W$ for some positive definite $W$. If

1. $\delta \left( \hat{\beta}^n, \beta^n_0 \right) \xrightarrow{P} 0$,
2. $\mathbb{E}_\beta \left[ \tilde{S}^n \right]$ is continuously differentiable over $\mathcal{B}$ and $\nabla_\beta \mathbb{E} \left[ \tilde{S}^n \right]$ has bounded derivative over $\mathcal{B}$,
3. $\sqrt{n^h} \left( \tilde{S}^n - \mathbb{E}_\beta^n \left[ \tilde{S}^n \right] \right) \rightsquigarrow \mathcal{N}(0, \Sigma)$,
4. $\lim_{n \to \infty} M^n(b) = M(b)$ exists for every $b \in \mathcal{B}$,
5. $\inf_n \lambda_{\min} \left( M^n(W^n M^n) \right) > 0$,

then

$$n^{h/2} R_n \left( \hat{\beta}^n - \beta^n_0 \right) \rightsquigarrow \mathcal{N} \left( 0, (M'WM)^{-1} M' \Sigma W' M (M'WM)^{-1} \right).$$

The lemma shows that if a sequence of SUGMs is consistent and the central limit theorem applies to the statistics, then the appropriately normalized estimator of the parameter is also centered around the true parameter with an asymptotic normal distribution. Thus, we just need to check identification (as already done above) and that our central limit theorem applies to the statistics.
5.2.2. Application of the Central Limit Theorem to Links and Triangles.

Applying our central limit theorem to our links and triangles example illustrates the results. We apply it using both the theorem and the corollary versions, in order to cover different cases and illustrate both results. The calculations are straightforward - basically, just expanding the various covariance expressions and checking that the rate conditions are satisfied. The corollary cases are much more direct, since the $\Delta$ sets are then singletons.

**Proposition 5.1.** Consider a links and triangles SUGM with associated parameters $\beta_{0,L}^n, \beta_{0,T}^n = \left(\frac{b_{0,L}}{n^{h_L}}, \frac{b_{0,T}}{n^{h_T}}\right)$ such that

$$h_L > 1/2 \quad \text{and} \quad h_T \in (h_L + 1, 5h_L - 1)$$

or

$$h_L > 1 \quad \text{and} \quad h_T \in (2, \min[1/2 + (3/2)h_L, h_L + 1, 3]).$$

Then the model satisfies the conditions of Theorem 1.

The proof appears in Section B of the Supplementary Appendix.

5.2.3. Simulations.

We return to the simulations from Section 4.4.1. We look at the asymptotic distribution of the GMM estimates $\hat{\beta}_L, \hat{\beta}_T$ in an example, to show that the asymptotic normality is a good approximation in finite samples.

Figure 5 plots the CDFs of the normalized parameter estimates across 500 simulations on $n = 350$ nodes for the example described before for the case where the average degree is $\approx 38$, as well as the standard normal CDF. The simulation shows that the normality approximation is a good one.

![Figure 5](image_url)

**Figure 5.** CDF of standardized parameter estimates $\hat{\beta}_L, \hat{\beta}_T$ across 500 simulations for $\beta_{0,L}, \beta_{0,T}$. 
6. Sparsity

While the results presented to this point provide general methods of estimating SUGMs, they involve calculations that can be circumvented in many cases of interest. In particular, many observed networks are sparse. For instance, even looking within limited settings, typical researchers have at most dozens of co-authors even though there are tens of thousands of researchers in any field. Generally, average degrees of nodes are of small order compared to the number of nodes, and this means that very simple direct counting methods can be accurate in estimating SUGMs in many applications - one can directly estimate parameters from observed counts. We give precise definitions of sparsity that ensure that the direct count estimates are accurate estimators for parameters in such sparse domains; and these restrictions apply in many settings of interest.

We call these estimators direct estimators, and denote them by \( \tilde{\beta} \) to distinguish them from the GMM estimators, \( \hat{\beta} \), though obviously these too correspond to a set of moments. The core idea is that when subgraph formation is sufficiently sparse, it is rare for a smaller subgraphs to incidentally generate larger ones. So, starting by counting the frequency of larger subgraphs (e.g., triangles in this case), then we can directly and accurately estimate the parameter that drives their formation. Next, after removing the triangles (since they always incidentally generate links), we then can count the relative frequency of links on the remaining pairs of nodes, which consistently estimates link formation.

Note that even though the parameters are estimated based on direct counts of subgraphs, there is still an important logic that needs to be imposed on how subgraphs are counted - for example, only estimating the frequency of links once we have removed the triangles. The ordering in which we do our counting is important since even in a sparse network larger subgraphs can still incidentally generate smaller subgraphs, but smaller ones will rarely incidentally generate larger ones.

Thus, we use the following convention in ordering the counting of statistics.


Consider a SUGM and order the classes of the subgraphs, \( G_1, \ldots, G_{\ell}, \ldots, G_k \), from ‘largest’ to ‘smallest’. In particular, we choose the ordering of \( 1, \ldots, k \) so that a subgraph in \( G_{\ell}^{n} \) cannot be a subnetwork of the subnetworks in \( G_{\ell'}^{n} \) for \( k \geq \ell' > \ell \geq 1 \):

\[
g_{\ell} \in G_{\ell}^{n} \text{ and } g_{\ell'} \in G_{\ell'}^{n} \implies g_{\ell} \not\subseteq g_{\ell'}.
\]

There exists at least one such ordering - for instance, any ordering in which subgraphs with more links are counted before subgraphs with fewer links. In an example with links, 2-stars and triangles: triangles precede 2-stars which precede links. Note that this is a partial order: for instance, a ‘three link line’ \( ij, jk, kl \) is neither a subgraph nor a supergraph of a
‘3-star’ \( ij, ik, il \), which is also a three link subgraph on four nodes. It is irrelevant in which order subgraphs with the same number of links are counted.

So, we count subgraphs in this order, and after having removed links associated with all of the subgraphs already counted, denoted \( \tilde{S}_n^\ell \):\(^{20}\)

\[
\tilde{S}_n^\ell (g) = \left| \{ g_\ell \in G_n^\ell : g_\ell \subset g \text{ and } g_\ell \cap g_\ell' \text{ for any } g_\ell' \in G_n^{\ell'} \text{ such that } g_\ell' \subset g \text{ for some } \ell' < \ell \} \right|
\]

**6.2. Direct Parameter Estimation.**

To define the direct parameter estimates, \( \bar{\beta} \)'s, from the counts, we then need to divide by the number of possible subgraphs that could exist on the remaining pairs of nodes after having removed the larger subgraphs. In particular, let \( \tilde{r}_n^\ell (g) \) denote the number of potential remaining subgraphs of type \( \ell \) exist after removing all those of types \( \ell'' < \ell \):\(^{21}\)

\[
\tilde{r}_n^\ell (g) = \left| \{ g_\ell \in G_n^\ell : g_\ell \cap g_\ell' \text{ for any } g_\ell' \in G_n^{\ell'} \text{ such that } g_\ell' \subset g \text{ for some } \ell' < \ell \} \right|
\]

In our links and triangles example, then \( \tilde{r}_n^T (g) = \binom{n}{3} \) and \( \tilde{r}_n^L (g) = \frac{n(n - 1)}{2} - L(\tilde{S}_T (g)) \) where \( L(\tilde{S}_T (g)) \) is the number of links that are part of triangles in \( g \). Typically, in sparse networks, the adjustments of the denominators to account for the deletion of links already in larger subgraphs will be inconsequential (see the proof of Proposition 2 below). For instance, there are relatively few triangles relative to what could be present and not many links will be lost to triangles.\(^{22}\)

The direct estimator \( \bar{\beta}_n^\ell \) is then

\[
(6.1) \quad \bar{\beta}_n^\ell = \frac{\tilde{S}_n^\ell (g)}{\tilde{r}_n^\ell (g)}
\]

For example, in a links and triangles model, direct estimators are

\[
(\bar{\beta}_T, \bar{\beta}_L) = \left( \frac{\# \text{ of triangles}}{\# \text{ of triples of nodes}}, \frac{\# \text{ of links not in triangles}}{\# \text{ of pairs of nodes that are not already together in some triangle}} \right)
\]

As we prove, under a sparsity condition these direct estimators are consistent estimates of the true parameters, and they are asymptotically Normally distributed.

\(^{20}\)Note in terms of the notation here, counting in order from ‘largest’ to ‘smallest’ subnetworks means that we count things from smallest to largest index \( \ell \): so the specification of how we ordered labels moves in the opposite direction of the size of the subgraphs.

\(^{21}\)This ignores the fact that not all of these could be formed without incidentally generating more larger networks. For instance, with links and triangles on 4 nodes, if we remove triangle 123, we are left with links 14, 24, 34, and at most one of those could form without incidentally generating another triangle. This bias will disappear in the sparse case with large number of links.

\(^{22}\)This does not mean that we can simply do away with our ordered-counting convention entirely - as the presence of directly formed links could still be of a similar order as the presence of links in triangles, it is just that both are relatively rare. So, the adjustments in the numerator associated with counting subgraphs are essential, while the ones in the denominator to track how many could have been present are not essential, but improve small-sample accuracy.
As an illustration, consider Figure 6 in which links and triangles are formed on 41 nodes. There are 9 truly generated triangles, but 10 observed overall. So, the frequency of triangles, \( \tilde{S}^n_T(g) \), is overestimated by using 10 instead of 9. The true frequency was 9/10660 but is estimated as 10/10660.

With respect to links, there were actually 25 truly directly generated, but one becomes part of an incidentally generated triangle and two others overlap on existing triangles, and so \( \tilde{S}^n_L(g) \) becomes 22 instead. Here we count them just out of the 820 – 30 = 790 remaining pairs of nodes that are not in triangles, so we estimate 22/790 while the true frequency was 25/820.

![Figure 6](image)

**Figure 6.** A network is formed on 41 nodes and is shown in panel D. The process can be thought of as first forming triangles as in (B), and links as in (C). Note that two links form on triangles, and a third link incidentally generates an extra triangle. In this network we would count \( \tilde{S}^n_T(g) = 10 \), and \( \tilde{S}^n_L(g) = 22 \) from (D), while the true process generated 9 triangles and 23 links directly. The estimates become \( \hat{\beta}^n_T = \frac{10}{10660} \) and \( \hat{\beta}^n_L = \frac{22}{790} \), while the true frequencies were \( \frac{9}{10660} \) and \( \frac{25}{820} \).

6.3. Generating Classes.

To define sparsity, we have to track how many ways a potential subnetwork \( g' \in G^n_T \) could be incidentally generated, many of the ways being equivalent up to relabelings. For instance, many different combinations of triangles and edges could incidentally generate a
triangle $g' = \{12, 23, 31\}$. However, notice that there are only eight ways in which it can be done if we ignore the labels of the nodes outside of $g'$: 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.

We first provide a precise specification of what it means to be incidentally generated. We say that a subgraph $g' \in G_\ell$ for some $\ell$ can be incidentally generated by the subgraphs $\{g^j\}_{j \in J}$, indexed by $J$, if $g' \subset \bigcup_{j \in J} g^j$.

Consider any potential subgraph $g' \in G_n^\ell$ that can be incidentally generated by a set of subnetworks $\{g^j\}_{j \in J}$ with associated indices $\ell_j$ 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 $g'$ if there exists a bijection $\pi$ from $J$ to $J'$ such that $\ell_j = \ell_{\pi(j)}$ and $|g_j \cap g'| = |g_{\pi(j)} \cap g'|$. So the equivalent generating sets have the same configurations in terms of numbers and types of subgraphs, and in terms of how many nodes each of those subgraphs intersects the given network.

So, for instance a triangle 123, could be incidentally generated by links 12, 23, and triangle 134; and an equivalent generator is links 12, 23, and triangle 135, and another is links 23, 13; and triangle 128, and so forth.

Given this equivalence relation, ignoring the specific labels of subgraphs we can define generating classes for any type of subgraph $G_\ell$. We just keep track of the number and type of subgraphs needed, as well as how many nodes each has subgraph intersecting with the given incidentally generated subgraph.

So, each generating class $\mathcal{C}$ of some $G_n^\ell$ is a list $\mathcal{C} = (\ell_1, c_1, \ldots, \ell_C, c_C)$ consisting of a list of types of subgraphs used for the incidental generation and how many nodes each has intersecting with the given incidentally generated subgraph. Thus, $\mathcal{C} = (\ell_1, c_1, \ldots, \ell_C, c_C)$ is such that there $\exists g' \in G_n^\ell$ generated by some $\{g^j\}_{j \in J}$ for which $|J| = C$ and for each $j$: $g^j \in G_n^\ell_j$ and $c_j = |g^j \cap g'|$.

We order generating classes so that the indices are ordered: $\ell_j \leq \ell_{j+1}$, and lexicographically $c_j \leq c_{j+1}$ whenever $\ell_j = \ell_{j+1}$. This ensures that we avoid counting the same class twice.\(^\text{23}\)

We only need to work with a small set of generating classes, so we restrict attention to the following:

- generating classes that are minimal: in the above $J$ there cannot be $j'$ such that $g' \subset \bigcup_{j \in J, j \neq j'} g^j$, and
- generating classes that only involve smaller subgraphs: $\ell_j \geq \ell$ for all $j \in J$.

The second condition states that we can ignore many generating classes because of our counting convention: when counting any given subgraph type, we only have to worry about incidental generation by the remaining (weakly smaller) subgraphs.

\(^\text{23}\) However, a generating class of two links and a triangle is a different generating class than one link and two triangles - this numbering just avoids the double counting of two links and a triangle separately from a triangle and two links.
So, for a links and triangles example, where \( G^n = (G^n_T, G^n_L) \) are triangles and links, 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 one triangle, or three links.\(^{24}\) Under the last condition above, there are no generating classes for links to worry about, since they cannot be incidentally generated by themselves and we only count them after removing all triangles.

6.4. Relative Sparsity.

Consider a set of (ordered) subgraphs \( G^n = (G^n_1, \ldots, G^n_k) \) and any \( \ell \in \{1, \ldots, k\} \) and any generating class of some \( \ell, C = (\ell_1, c_1; \ldots; \ell_C, c_C) \). Let\(^{25}\)

\[
M_C = (\sum_{j=1}^{\ell} c_j) - m_\ell.
\]

For example, in forming a triangle from any combination of triangles and links, each \( c_j = 2 \) and so \( M_C = 6 - 3 = 3 \).

We say that a sequence SUGMs with associated (ordered) subgraphs \( G^n = (G^n_1, \ldots, G^n_k) \) and parameters \( \beta_0^n \) is relatively sparse if and for each \( \ell \) and associated generating class \( C \) with associated \( (\ell_j, c_j)_{j \in 1, \ldots, \ell} \):

\[
\frac{\prod_{j=1}^{\ell} E_{\beta_0^n}(S^n_{\ell_j}(g))}{n^{MC} E_{\beta_0^n}(S^n_{\ell}(g))} \to 0,
\]

and

\[
\frac{\prod_{j=1}^{\ell} E_{\beta_0^n}(S^n_{\ell_j}(g))}{n^{MC} E_{\beta_0^n}(S^n_{\ell_{j'}}(g))} \to 0,
\]

for each \( j' \in 1, \ldots, C \). Note that this condition applies to the actual frequencies of subgraphs \( S^n_{\ell} \) - a condition on primitives - rather than the directly counted subgraphs \( \tilde{S}_n^n \), although the condition turns out to be equivalent when it holds. This is a condition that provides us with the necessary bounds the relative frequency with which subgraphs are incidentally generated (the numerator) compared to directly generated (the denominator). It applies in two ways: one is that new graphs are not being incidentally generated at too fast a rate, and secondly, that given subgraphs are not disappearing into incidentally generated larger

---

\(^{24}\)Here, then we would represent a generating class of two triangles and a link as \((T, 2; T, 2; L, 2)\), where this indicates that two triangles were involved and each intersected the subgraph in question in two nodes and then \( L, 2 \) indicates that a link was involved intersecting the subgraph in two nodes.

\(^{25}\)Note that \( M_C \geq 1 \) since \( C \geq 2 \) and some set of \( c_j \) nodes intersects with at least one other set of \( c_{j'} \) nodes for some \( j' \neq j \) (noting that the incidentally generated subgraph is not a collection of disconnected links). Recall that under the ordering, lower-indexed subgraphs cannot be generated as a subset of some single higher-indexed one.
subgraphs at too fast a rate. Although notationally complex, they are easily checked with links and triangles, for instance if \( \beta_L = b_L/n \) and \( \beta_T = b_T/n^2 \), or for many other values.\(^{26}\)

6.5. Estimation of Sparse Models.

In order to have \( \delta \left( \bar{\beta}^n, \bar{\beta}_0^n \right) \xrightarrow{P} 0 \), beyond the network being relatively sparse, it must also be that the potential number of observations of a particular kind of subgraph grows as \( n \) grows. For instance, if nodes have different characteristics (say some demographics), and we are counting triangles and links by node types, then it also has to be that the number of nodes that have each demographic grows as \( n \) grows. If there were never more than 5 nodes with some demographic, then we cannot get an accurate estimate of link formation among those nodes.

We say a SUGM is growing if the probability that \( S^n_\ell (g) \to \infty \) for each \( \ell \) goes to 1.

**Proposition 2** (Consistency and Asymptotic Normality of Direct Estimators of Sparse SUGMs). Consider a sequence of growing and relatively sparse SUGMs with associated subgraph statistics \( \bar{S}^n = (\bar{S}_1^n, \ldots, \bar{S}_k^n) \) and parameters \( \beta^n_0 = (\beta^n_{0,1}, \ldots, \beta^n_{0,k}) \). Consider the direct estimator \( \bar{\beta} \) described above. Then

1. \( \delta \left( \bar{\beta}^n, \bar{\beta}_0^n \right) \xrightarrow{P} 0 \) and
2. \( \nabla^{-1/2}(\beta^n - \beta^n_0) \sim \mathcal{N}(0, I) \) where \( \nabla_{\ell, \ell} = \frac{\beta^n_{0,\ell}(1 - \beta^n_{0,\ell})}{\kappa(\bar{m}_\ell)} \) and the off-diagonals are all 0.

Proposition 2 states that growing and relatively sparse SUGMs are consistently estimable via a very simple estimation technique that is easily computable.

The proof of the proposition involves showing that, under the growing and sparsity conditions, the fraction of incidentally generated subnetworks vanishes for each \( \ell \), and the observed counts of subnetworks converge to the truly generated ones. And then, by a standard limiting argument applied to the truly generated subgraphs (which are independent), the appropriately normalized vector of subgraph counts are asymptotically normally distributed (with an approximately independent distribution).

7. Applications

We consider two applications to data. In the first, we show that SUGMs are much better than link-level models with covariates, at replicating several features (beyond subgraph counts) of network data. Second, we show how SUGMs can be used to test a hypothesis about preferences for link formation.

---

\(^{26}\) Again, we emphasize that many empirical applications have degrees that are fairly constant with network size, and \( \beta_L = b_L/n \) and \( \beta_T = b_T/n^2 \) covers a case in which nodes have expected degree roughly \( b_L + b_T/2 \) irrespective of how large the network grows.
7.1. Network properties generated by SUGMs.

In our first application, we compare a simple and standard (‘stochastic block’) model that estimates linking probabilities based on node characteristics - caste and geography - to a SUGM based on links and triangles. The idea is to compare how well each of these models replicates various features of actual 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, Chandrasekhar, Duflo, and Jackson (2013) data consisting of Indian villages with an average of 220 nodes. Here we focus on “advice” networks: an edge represents whether a household speaks to another household when having to make an important decision and we use an undirected, unweighted graph. 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.

Beyond average degree and clustering (which turn out to be well-captured by links and triangles), we are interested whether a very basic SUGM does a good job of replicating observed networks in terms of characteristics other than those that directly involve link and triangle counts. 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, Borgs, Chayes, and Riordan (2010); 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 – namely, for any subset of nodes the number of links leaving the subset relative to the number of links within the subset. 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, Vayanos, and Zwiebel (2003); 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 also consider average path length (in the largest component).

Our procedure is as follows. For every village, we estimate two different network formation models. The first network formation model is a link-based model where the probabilities can depend on geographic and caste covariates. In particular, pairs of household are categorized

\[ T_{ij} = \frac{g_{ij}}{\sum_k g_{ik}}, \]

where either \( g_{ii} = 1 \), or \( g_{ik} > 0 \) for some \( k \neq i \), as this captures the set of people to whom \( i \) listens.

\[ \text{In these data we actually have 75 different villages, which each provides its own network (in fact we have networks for several different types of relationships for each village). This is unusual, where a researcher might often have access to only one or a few networks. Having more also enables us to verify that the model is working as claimed, but is not needed to apply the model.} \]

\[ \text{The stochasticized adjacency matrix } T \text{ is defined as } T_{ij} = \frac{g_{ij}}{\sum_k g_{ik}}, \text{ where either } g_{ii} = 1, \text{ or } g_{ik} > 0 \text{ for some } k \neq i, \text{ as this captures the set of people to whom } i \text{ listens.} \]
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. There, we categorize triangles as being “close” if all nodes are of the same caste and each pairs is below the median distance, and “far” otherwise. 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 for each of the 36 villages and for each of the two models. We then compare the aforementioned network characteristics from the simulations with the actual data.\(^{29}\)

Table 1 presents the results.\(^{30}\) 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](image)

<table>
<thead>
<tr>
<th>Models are fit to different combinations of these statistics.</th>
<th>Data</th>
<th>Link-based model with covariates</th>
<th>SUGM with links and triangles</th>
<th>SUGM with isolates, links and triangles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Unsupported Links</td>
<td>160.8</td>
<td>236.2</td>
<td>161.2</td>
<td>161.8</td>
</tr>
<tr>
<td>Number of Triangles</td>
<td>39.2</td>
<td>3.1</td>
<td>39.7</td>
<td>39.5</td>
</tr>
<tr>
<td>Average Degree</td>
<td>2.3243</td>
<td>2.3260</td>
<td>2.5916</td>
<td>2.5219</td>
</tr>
<tr>
<td>Number of Isolates</td>
<td>54.9722</td>
<td>25.7222</td>
<td>31.4444</td>
<td>65.9167</td>
</tr>
<tr>
<td>Average Clustering</td>
<td>0.0895</td>
<td>0.0105</td>
<td>0.1268</td>
<td>0.0829</td>
</tr>
<tr>
<td>Fraction in Giant Component</td>
<td>0.7061</td>
<td>0.8315</td>
<td>0.7982</td>
<td>0.6718</td>
</tr>
<tr>
<td>First Eigenvalue</td>
<td>5.5446</td>
<td>3.8578</td>
<td>4.6762</td>
<td>5.3025</td>
</tr>
<tr>
<td>Spectral Gap</td>
<td>0.9550</td>
<td>0.3354</td>
<td>0.6884</td>
<td>1.0617</td>
</tr>
<tr>
<td>Second Eigenvalue of Stochastized Matrix</td>
<td>0.9573</td>
<td>0.9632</td>
<td>0.9559</td>
<td>0.9069</td>
</tr>
<tr>
<td>Average Path Length</td>
<td>4.6921</td>
<td>5.6565</td>
<td>5.1215</td>
<td>4.1180</td>
</tr>
</tbody>
</table>

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

\(^{29}\) We have complete GPS and caste data for only 36 villages.

\(^{30}\) An extended version of Table 1 with standard errors appears in Appendix F.
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. Not surprisingly, it fits isolates better than either of the previous models. The more interesting aspects are in the other features to which none of the models are fit. 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 the second eigenvalue (the homophily measure). However, that is likely because the 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 counter-act that, but is beyond our illustrative purposes here.

We also examine distributional outcomes. In Figure 7, 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 7a 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 7b 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.

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 basic 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. Finally, knowing that our model is better able to capture the realistic correlation of links within observed networks should make us more confident in trusting the results of the empirical application in Section 7.2. When we look at links across social boundaries, we can be comfortable that to first order, thinking about a SUGM with links and triangles across and within caste groups is a reasonable cut of the data.

7.2. Links across social boundaries.

As another illustration, let us answer the question about propensities to link across castes that we mentioned in the introduction.

Individuals are associated with groups and identities that can lead to strong social norms - prescriptions and proscriptions - about interactions across groups. For instance, in much of India there are strong forces that influence if and when individuals form relationships across castes. 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? To answer this we need models that account for link dependencies, as cliques of three or more may dictate greater adherence to a group norm prohibiting certain inter-caste relationships, while the norm may be circumvented in isolated bilateral relationships.

To analyze this, we examine network data from Indian villages (from our study Banerjee, Chandrasekhar, Duflo, and Jackson (2014) that we discuss in more detail below). We link two households if members of either engaged in favor exchange with each other: that is, they borrowed or lent goods such as kerosene, rice or oil in times of need. 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 (Munshi and Rosenzweig, 2006). Scheduled castes and scheduled tribes are those defined by the Indian government as being disadvantaged. This is a fundamental distinction over which the strongest cultural forces are likely to focus. Additional norms are at work with finer caste (jati) distinctions, but those norms are more varied depending on the particular castes in question while this provides a clear barrier.

As a simple model to address this issue, consider a process in which 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 prefer to form the link, and a triangle is formed if and only if all three individuals prefer to form it. This minimally complicates an independent-link model enough to require modeling link interdependencies.

In particular, there are probabilities, denoted $\pi_L(diff), \pi_L(same)$, that a given link has an opportunity to form (i.e., the pair meets 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 described by a random utility framework (McFadden, 1973). Individual $i$’s utility of having a relationship with $j$ can by influenced by whether they share caste and is given by

$$ u_i(ij) = \alpha_{0,i} + \beta_{0,i} SameCaste_{ij} + \delta_{0,i} X_{ij} - \epsilon_{L,ij}, $$

where $SameCaste_{ij}$ is a dummy for whether both individuals are members of the same caste, $X_{ij}$ is a vector of covariates depending on $X_i$ and $X_j$. For expositional simplicity here, we set $\delta_L = 0$. The outside option is normalized to zero, so $p_L(same)$ is the probability that an individual desires to form a link with an individual of the same caste group, and $p_L(diff)$ is the probability that an individual desires to form a link with an individual of a different caste group.
The crucial point is that $i$ can have returns that depend on being in a multilateral relationship with $j$ and $k$ – that is conceptually distinct from having these two bilateral relationships – and this can be given by

$$u_i(ijk) = \alpha_{0,L} + \beta_{0,T} \text{SameCaste}_{ijk} + \delta_{0,T} X_{ijk} - \epsilon_{T,i,ijk},$$

where $\text{SameCaste}_{ijk}$ is a dummy for whether all three individuals are members of the same caste, $X_{ijk}$ is a vector of covariates depending on $X_i$, $X_j$, and $X_k$. Again for expositional simplicity, we set $\delta_{0,T} = 0$. Correspondingly, $p_T(\text{same})$ is the probability that an individual desires to form a triangle when all individuals are of the same caste group, and $p_T(\text{diff})$ is the probability that an individual desires to form a triangle when it consists of people from both caste groups.\(^{31}\)

The hypothesis that we explore is that $p_T(\text{diff})/p_T(\text{same}) < p_L(\text{diff})/p_L(\text{same})$ 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 $p_T(\text{diff})/p_T(\text{same}) = p_L(\text{diff})/p_L(\text{same})$.

Note that the probability that a “same” link forms is

$$P_L(\text{same}) = p_L(\text{same})^2 \pi_L(\text{same})$$

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

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

Analogously for triangles we have

$$P_T(\text{same}) = p_T(\text{same})^3 \pi_T(\text{same}) \quad \text{and} \quad P_T(\text{diff}) = p_T(\text{diff})^3 \pi_T(\text{diff}),$$

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 fair 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.

One 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 we still have identification as the meeting bias makes 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.

\(^{31}\)This is a simplified model for illustration, but one can clearly consider preferences conditional on any string of covariates. This extends a model such as that of Curra{r}ini, Jackson, and Pin (2009, 2010) to allow for additional link dependencies. We could also be interested in higher order relationships.
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 8. (Variations on this sort of biased meeting process appear in Currarini, Jackson, and Pin (2009, 2010); Bramoullé, Currarini, Jackson, Pin, and Rogers (2012).)

**Figure 8.** Geographically driven meeting process where agents spend 3/4 of their time in their own community.

**Lemma 3.** A sufficient condition for $\frac{p_T^{(diff)}}{p_T^{(same)}} < \frac{p_L^{(diff)}}{p_L^{(same)}}$ is that $\frac{p_T^{(diff)}}{p_T^{(same)}} < \left( \frac{p_L^{(diff)}}{p_L^{(same)}} \right)^{3/2}$.

The proof appears in the appendix, but follows from straightforward calculations.

Given Lemma 3, we can test our hypothesis directly from a SUGM that compares relative link and triangle counts (we can also include isolated nodes, but those do not impact this hypothesis). In particular, we only need examine whether $\frac{p_T^{(diff)}}{p_T^{(same)}} < \left( \frac{p_L^{(diff)}}{p_L^{(same)}} \right)^{3/2}$.

Figure 9 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 at the 99 percent level.\(^{32}\)

In Figure 9, Panel (A), 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 balanced (e.g., see McPherson, Smith-Lovin, and Cook (2001); Currarini, Jackson, and Pin (2009, 2010)). Here we see that the

\(^{32}\)This is from doing a conservative nonparametric test: under the null that the number of villages for which the ratio is less should be 1/2 with a binomial distribution on the number above or below.
Figure 9. Comparison of the relative propensity to form cross-caste versus same-caste relationships for triangles (y-axis) compared to links (x-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 in Panel (A) distinguishes those villages that have above/below the median size minority group. Panel (B) presents standard errors that would be used if the researcher had access to data from that singular network.

social pressures against mixed-caste triangles are stronger when the two caste designations are more evenly balanced.

Panel (B) presents the standard errors associated with the parameter estimate ratios. Most researchers have access to a single or handful of networks at best. So the standard errors computed in (B) are based on single network asymptotics (for each network as if it were the whole data set) as described in this paper. A simple interpretation of Panel (B) therefore is a meta-study plot of the various inferences that 75 different papers would have made, each having the dataset from one village in our sample.

8. Additional Models and Extensions


Another perspective on forming networks, is not that people are faced with various subgraphs (possibly randomly) and then choose which ones to form, but instead they put in
effort which affects how many subgraphs of various types that they have an opportunity to be a part of (i.e., agents actively “network”). The theory literature has studied models with search intensities as drivers for network formation (Curra\-ri\-ni, Jack\-son, and Pin, 2009, 2010; Borgs, Chayes, Ding, and Lucier, 2010; Golub and Livne, 2010), and such models can be incorporated here.

The basic idea is that an agent with attributes $X_i$ can put in search effort to form a clique with $m$ nodes consisting of characteristic vector $X$. From this match, an agent obtains a utility

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

The simplest possible version of this is $u(X_i, m, X) = u_m$, meaning that characteristics of parties do not matter, only the subgraph itself does.

Let

$$e(X_i, m, X) \in [0, \tau_m]$$

denote the search effort for cliques of size $m$ with agents characteristics $X$ by an agent with characteristics $X_i$.

Whether or not an actual clique $Cl_m$ actually forms is stochastic, depending on the vector of efforts of those involved

$$p(e_1, \ldots, e_m) = p(e_1(X_1, m, X), \ldots, e_m(X_m, m, X)),$$

and presumed to be non-decreasing in each argument.

Continuing our simple example, we may have

$$p(e_1, \ldots, e_m) = \sum_{j \in [1, \ldots, m]} e_j,$$

where by assumption $\tau_m$ is such that $m \tau < 1$. This means that no agents can put in enough effort to guarantee a successful meeting.

Effort is costly, so let the cost function be $c ((e_i(X_i, m, X))_{m,X}, X_i)$ which can depend on her characteristics and her search effort for all cliques of all attribute types. The form of course is up to the modeler.

Here a simple example is

$$c ((e_i(X_i, m, X))_{m,X}, X_i) = \sum_m \phi_m e_i^2 / 2.$$

In sum, the expected utility is

$$(8.1) \sum_{m,Cl_m,i \in Cl_m} u(x_i, m, x) \cdot p((e_j(x_j, m, x))_{j \in Cl_m}) - c ((e_i(x_i, m, x))_{m,x}, x_i).$$
Returning to our simple example, we have

$$\sum_{m, C_l, i \in C_l} u_m \cdot \left( \sum_{j \in [m]} e_{j,m} \right) - \sum_{m} \phi_m \frac{e_{1,m}^2}{2}.$$ 

This can have multiple equilibria. If $u$’s are nonnegative, this defines a supermodular game. Agent $i$’s change in payoff from increasing any dimension of $(e_i(X_i, m, X))_{m,X}$ is nondecreasing in the vector of strategies $(e_j(X_j, m, X))_{j \neq i, m, X}$. Since in such games pure strategy equilibria exist and form a complete lattice (e.g., see Topkis (2001)), it is up to the econometrician to specify additional conditions on $p(\cdot)$, $u(\cdot)$, and $c(\cdot)$ to ensure uniqueness of equilibrium. One may also appeal to equilibrium selection.33

In our example, we would find that

$$e_{i,m}^* = \frac{u_m}{\phi_m}.$$ 

Normalizing a cost to benefits parameter $\beta_m := \frac{u_m}{\phi_m}$, and noting that the frequency of observing a clique therefore is

$$p_m = p(e_1^*, \ldots, e_m^*) = me_{1,m}^* = m\beta_m,$$

we see how to map from the frequency of a subgraph to the structural parameter.

Again, we note that this is certainly not the only model. This is an example of a simple version of the model to give the reader an idea as to how it works. One could certainly condition on covariates or make different functional form assumptions.

It should be clear that this model corresponds to a network formation process where the frequencies of cliques of size $m$ consisting of agents with characteristics described by the profile $X$ are the subgraphs that are counted. Note that though the discussion is described for cliques, it can easily be adjusted for any subgraphs (for instance an agent may value being the center of a star with $m$ agents).

8.2. Directed Network Formation.

The SUGMs that we have defined include directed networks. Although most of our discussion has been for the undirected case, the formal definitions apply to the directed case, as subgraphs are easily defined with directed links. The model also easily accommodates multigraphs.

9. Continuous Covariates and using Logistic Regression for Estimation

---

33Because there are positive spillovers, generally the maximal equilibrium will Pareto dominate the others. The standard refinement would be to look at the Pareto efficient equilibrium which is then unique and pure (see Vives (2007) for some background).
As we have mentioned, our analysis extends easily to cases in which continuous covariates are included. For example, if we allow the probability of some subgraph type $\ell$ to depend on a continuous covariate, the model extends fairly directly, and this is easy in the sparse case. We discuss this at length in Supplementary Appendix E.

10. Summary

We have provided a new class of network models, SUGMs, that are well-suited to practical statistical estimation of social and economic networks, especially as the edges in such networks tend to be correlated. We have shown that a broad class of these models are well-identified and can be easily estimated by GMM, or even more basic and direct techniques in the case of sparse networks. As a by-product we have also proven a new central limit theorem for correlated random variables. The power of these models was illustrated in applications.

References


Banerjee, A. V., A. Chandrasekhar, E. Duflo, and M. O. Jackson (2014): “Gossip: Identifying Central Individuals in a Social Network,” Available at SSRN 2425379. 7.2


Bollobás, B. (2001): Random Graphs, Cambridge University Press. 4.1


Appendix A. Proofs

Proof of Proposition 1.

Write

\[ \beta^n = \left( \frac{b^n_L}{n^{h_L}}, \frac{b^n_T}{n^{h_T}} \right) \quad \beta^n_0 = \left( \frac{b^n_{L0}}{n^{h_L}}, \frac{b^n_{T0}}{n^{h_T}} \right), \]

where \( b^n_L, b^n_T, b^n_{L0}, b^n_{T0} \) lie in \([D, D]\).

Let \( r^n_L = 1/n^{h_L} \) and \( r^n_T = 1/n^{h_T} \).

First, note that \( 1 - (1 - \beta^n_T)^x \) is the probability that some link is formed as part of at least one triangle out of \( x \) possible triangles that could have it as an edge (independently of whether it also forms directly).

Next, note that the probability that a link forms conditional on some particular triangle that it could be a part of not forming is

\[ q^n_L = \beta^n_L + (1 - \beta^n_L) \left( 1 - (1 - \beta^n_T)^{n-3} \right). \] (A.1)

So, we can write the probability of some triangle forming as

\[ q^n_T := E_{\beta^n_L, \beta^n_T} [ST(g)] = \beta^n_T + (1 - \beta^n_T) (\bar{q}^n_L)^3, \] (A.2)

where the first expression \( \beta^n_T \) is the probability that the triangle is directly generated, and then the second expression \( (1 - \beta^n_T) (\bar{q}^n_L)^3 \) is the probability that it was not generated directly, but instead all three of the edges formed on their own (which happen independently, conditional on the triangle not forming, which has probability \( (\bar{q}^n_L)^3 \)).

It is useful to note that since \( \beta^n_L = o(1), (1 - \beta^n_L) \to 1 \) and since \( h_T > 1, (1 - \beta^n_T)^{n-3} \to 1 - \frac{b^n_{T0}}{n^{h_T}} \). Thus,

\[ \bar{q}^n_L = \Theta \left( \frac{1}{n^{h_L}} + \frac{1}{n^{h_T-1}} \right) = \Theta \left( \frac{1}{n^{h_L}} \right) \]

where the second equality follows since \( h_T \geq h_L + 1 \).

Next, note that the probability that a link forms is

\[ q^n_T := E_{\beta^n_L, \beta^n_T} [SL(g)] = \beta^n_L + (1 - \beta^n_L) \left( 1 - (1 - \beta^n_T)^{n-2} \right), \] (A.3)

where the first expression \( \beta^n_L \) is the probability that the link is directly generated, and then the second expression \( (1 - \beta^n_L) (1 - (1 - \beta^n_T)^{n-2}) \) is the probability that it was not generated directly, but instead appeared as an edge in some triangle (and there are \( n - 2 \) such possible triangles).

---

34We allow the constants to depend on \( n \) to capture that some applications have both rates and constants that adjust with scale, and we may want to fit across data of networks of varying sizes. But this is largely semantic, as estimating any particular network has only one \( b \), and one can ignore the superscripts on the \( b \)s if one likes.

35That is, consider a given pair of nodes \( i, j \) and a third node \( k \). Consider the probability that link \( ij \) is formed conditional on triangle \( ijk \) not forming directly as a triangle.
It is also useful to write this in a very different way:

(A.4) \[ q^n_L := E_{\beta^n_L, \beta^n_T}[S_L(g)] = \beta^n_T + (1 - \beta^n_T)\bar{q}^n_L, \]

noting that a link could form as part of a triangle that it is part of, or else form conditional upon that triangle not forming.

The first part of the Proposition can be derived from the expressions (A.2) and (A.4), noting that they only differ in the exponent on \( \bar{q}^n_L \), and is left to the reader (who can also prove it as below, without rate restrictions). We prove the second part of the Proposition regarding identifiable uniqueness.

The following derivative expressions are useful:

(A.5) \[
\begin{align*}
\frac{\partial \bar{q}^n_L}{\partial \beta^n_L} &= (1 - \beta^n_T)^{n-3} & \frac{\partial \bar{q}^n_L}{\partial \beta^n_T} &= (n - 3)(1 - \beta^n_L)(1 - \beta^n_T)^{-2}, \\
\frac{\partial \bar{q}^n_T}{\partial \beta^n_L} &= (1 - \beta^n_T)^{n-2}. & \frac{\partial \bar{q}^n_T}{\partial \beta^n_T} &= 3(1 - \beta^n_T)(\bar{q}^n_T)^2 \frac{\partial q^n_L}{\partial \beta^n_L} = 3(\bar{q}^n_L)^2(1 - \beta^n_T)^{n-2}. \\
\frac{\partial q^n_L}{\partial \beta^n_T} &= 1 - \bar{q}^n_L + (1 - \beta^n_T)\frac{\partial q^n_L}{\partial \beta^n_T} = 1 - \bar{q}^n_L + (n - 3)(1 - \beta^n_L)(1 - \beta^n_T)^{n-1}. & \frac{\partial q^n_T}{\partial \beta^n_T} &= 1 - (\bar{q}^n_T)^3 + 3(1 - \beta^n_L)(\bar{q}^n_T)^2 \frac{\partial q^n_L}{\partial \beta^n_T} = 1 - (\bar{q}^n_T)^3 + 3(\bar{q}^n_L)^2(n - 3)(1 - \beta^n_L)(1 - \beta^n_T)^{n-1}.
\end{align*}
\]

Given that \( \beta^n_L = o(1) \) (since \( h_L > 0 \)), \( \beta^n_T = o(1/n) \) (since \( h_T > 1 \)), and \( \bar{q}^n_L = \Theta\left(\frac{1}{n^{2h_L}}\right) \) the above expressions imply that:

(A.6) \[ \frac{\partial q^n_L}{\partial \beta^n_L} = 1 - o(1), \]

(A.7) \[ \frac{\partial q^n_L}{\partial \beta^n_T} = \Theta\left(\frac{1}{n^{2h_L}}\right), \]

(A.8) \[ \frac{\partial q^n_T}{\partial \beta^n_L} = n - 2 - o(1), \]

(A.9) \[ \frac{\partial q^n_T}{\partial \beta^n_T} = \Theta\left(\max[1, n^{1 - 2h_L}]\right). \]

Note that (A.6)-(A.9) hold for any parameters \( h_L > 0 \) and \( 3h_L > h_T \geq h_L + 1 \) - and thus uniformly for any \( \beta^n \) in a compact set of such \( h_L, h_T \), and thus as long as we restrict attention
to $\beta^n$ in that compact set, we have the same order derivatives and so then we approximate:

\[
\frac{E_{\beta^n} [S_L(g)] - E_{\beta_0^n} [S_L(g)]}{r_L^n} \approx n^{h_L} \left[ \frac{b^n_L - b^n_{L0}}{n^{h_L}} + (n - 2) \frac{b^n_T - b^n_{T0}}{n^{h_T}} \right]
\]

\[\approx b^n_L - b^n_{L0} + (b^n_T - b^n_{T0})\Theta(n^{h_L+1-h_T}),\]

and

\[
\frac{E_{\beta^n} [S_T(g)] - E_{\beta_0^n} [S_T(g)]}{r_T^n} \approx n^{h_T} \left[ \frac{b^n_L - b^n_{L0}}{n^{h_L}} \Theta(1/n^{2h_L}) + \frac{b^n_T - b^n_{T0}}{n^{h_T}} \Theta\left(\max[1, n^{1-2h_L}]\right) \right]
\]

\[\approx (b^n_L - b^n_{L0})\Theta(n^{h_T-3h_L}) + (b^n_T - b^n_{T0})\Theta\left(\max[1, n^{1-2h_L}]\right).\]

To establish identifiable uniqueness (given the additive separability of $Q^n(\beta)$ across $U, T$) it is sufficient to argue that for any $\varepsilon > 0$ there exists $\phi > 0$ such that for large enough $n$, if $\delta((\beta^n_L, \beta^n_T), (\beta^n_{0,L}, \beta^n_{0,T})) > \varepsilon$, then at least one of the following inequalities holds:

\[
\left| \frac{E_{\beta^n} [S_L(g)] - E_{\beta_0^n} [S_L(g)]}{r_L^n} \right| > \phi
\]

or

\[
\left| \frac{E_{\beta^n} [S_T(g)] - E_{\beta_0^n} [S_T(g)]}{r_T^n} \right| > \phi.
\]

Note that $\delta((\beta^n_L, \beta^n_T), (\beta^n_{0,L}, \beta^n_{0,T})) > \varepsilon$ translates into $|b^n_L - b^n_{L0}| > c\varepsilon$ and/or $|b^n_T - b^n_{T0}| > c\varepsilon$ for some $c > 0$. If the second inequality holds, then by (A.11) it follows that (A.13) holds. If (A.13) does not hold for any $\phi$, then by (A.11) it must be that $|b^n_L - b^n_{L0}| > c\varepsilon$ while $|b^n_T - b^n_{T0}| < \delta^n$ for a sequence $\delta^n \to 0$. In that case, noting that since $h_T \geq h_L - 1$ (and so the second term of (A.10) is of order at most 1 times $\delta^n$ while the first term is at least $c\varepsilon$ in magnitude), then by (A.10) it follows that (A.12) holds. ■

The following lemmas are useful in the proof of Theorem 1.

**Lemma A.1.** A solution to $\max_h E[Zh(Y)]$ s.t. $|h| \leq 1$ (where $h$ is measurable) is $h(Y) = \text{sign}(E[Z|Y])$, where we break ties, setting $\text{sign}(E[Z|Y]) = 1$ when $E[Z|Y] = 0$.

**Proof.** This can be seen from direct calculation:

\[
E[Zh(Y)] = \int_Y E[Z|Y]h(Y)dP(Y)
\]

Maximizing $E[Z|Y]h(Y)$ pointwise when $|h| \leq 1$ is achieved by setting $h(Y) = \text{sign}(E[Z|Y])$, and it is clearly ok to break ties by setting $\text{sign}(E[Z|Y]) = 1$ when $E[Z|Y] = 0$, as that makes no difference in the integral. ■
Lemma A.2. \(E[XYh(Y)]\) when \(h(\cdot)\) is measurable and bounded by \(\sqrt{\frac{2}{\pi}}\) satisfies
\[
E[XYh(Y)] \leq \sqrt{\frac{2}{\pi}} E[X \cdot \text{sign}(E[X|Y]Y)] .
\]

Proof. This follows from Lemma A.1, setting \(Z = XY\). ■

Proof of Theorem 1. By Lemma 5.1, it is sufficient to show that the appropriate sequence of random variables \(S_N^N\) satisfies
\[
\sup_{\{f: \|f\|, \|f''\| \leq 2, \|f'\| \leq \sqrt{2/\pi}\}} \left| E[f'(S_N^N) - S_N^N f(S_N^N)] \right| \to 0.
\]

Recall
\[
a_N = \sum_{\alpha, \eta \in \Delta(\alpha, N)} \text{cov} (X_\alpha, X_\eta),
\]
and
\[
S_N^N = S_N / a_1^{1/2}.
\]

Also let the size of the dependency set be given by \(M(N) = |\Delta(\alpha, N)|\).

Then define the following average covariances:
\[
c_1^N = \frac{\sum_{\eta \in \Delta(\alpha, N)} \text{cov}(X_\alpha, X_\eta)}{M(N)}, \quad c_2^N = \frac{\sum_{\eta \notin \Delta(\alpha, N)} \text{cov}(X_\alpha, X_\eta)}{N - M(N)}.
\]

Note that \(a_N = NM(N)c_1^N\).

For ease of notation, we omit the superscript \(N\)s below.

Let
\[
S_\alpha := \sum_{\eta \notin \Delta(\alpha, N)} (X_\eta - \mu) \quad \text{and} \quad \overline{S}_\alpha := S_\alpha / a_1^{1/2}.
\]

Observe that
\[
E \left[ \overline{S} f \left( \overline{S} \right) \right] = E \left[ \frac{1}{a_1^{1/2}} \sum_\alpha (X_\alpha - \mu) \cdot f \left( \overline{S_\alpha} \right) \right]
\]
\[
= E \left[ \frac{1}{a_1^{1/2}} \sum_\alpha (X_\alpha - \mu) \left( f \left( \overline{S} \right) - f \left( \overline{S_\alpha} \right) \right) \right] + E \left[ \frac{1}{a_1^{1/2}} \sum_\alpha (X_\alpha - \mu) \cdot f \left( \overline{S_\alpha} \right) \right].
\]

The first step is to show that
\[
\left| E \left[ \frac{1}{a_1^{1/2}} \sum_\alpha (X_\alpha - \mu) \cdot f \left( \overline{S_\alpha} \right) \right] \right| = o(1),
\]
by (5.3).

To see this, observe that if \(E[X_\alpha - \mu|X_\eta - \mu]\) is nonnegative for all \(\alpha, \eta\), then it follows that \(\text{sign}(E[(X_\alpha - \mu)(X_\eta - \mu)](X_\eta - \mu))\) is always either 0 or 1. Therefore we can reduce (5.3) to \(\sum_{\alpha, \eta \notin \Delta} E[(X_\alpha - \mu)(X_\eta - \mu)] = o(a_N)\).
In order to apply this, we can expand the term to
\[
\left| E \left[ \frac{1}{a_N^{1/2}} \sum_{\alpha \in A} (X_\alpha - \mu) \cdot f \left( \widehat{S}_\alpha \right) \right] \right| = E \left[ \frac{1}{a_N^{1/2}} \sum_{\alpha \in A} (X_\alpha - \mu) \cdot f \left( \frac{1}{a_N^{1/2}} \sum_{\eta \notin \Delta(\alpha,N)} (X_\eta - \mu) \right) \right] \\
\leq E \left[ \frac{1}{a_N^{1/2}} \sum_{\alpha \in A} (X_\alpha - \mu) \cdot f (\mu) \right] = 0 \text{ since } E[X_\alpha - \mu] = 0.
\]
\[
+ E \left[ \frac{1}{a_N^{1/2}} \sum_{\alpha \in A} (X_\alpha - \mu) \cdot \left( \frac{1}{a_N^{1/2}} \sum_{\eta \notin \Delta(\alpha,N)} (X_\eta - \mu) \right) \cdot f' \left( \widehat{S}_\alpha \right) \right] 
\]
where \( \widehat{S}_\alpha \) is an intermediate value between \( S_\alpha \) and \( \mu \).

To bound the second term, we apply Lemma A.2 to conclude that
\[
\left| E \left[ \sum_{\alpha \in A; \eta \notin \Delta(\alpha,N)} (X_\alpha - \mu) (X_\eta - \mu) \cdot f' \left( \widehat{S}_\alpha \right) \right] \right| \\
\leq \sqrt{\frac{2}{\pi}} E \left[ \sum_{\alpha \in A; \eta \notin \Delta(\alpha,N)} (X_\alpha - \mu) (X_\eta - \mu) \cdot \text{sign} \left( E \left[ (X_\alpha - \mu) (X_\eta - \mu) \mid (X_\eta - \mu) \right] \right) \right] \text{.}
\]

Thus, it is sufficient that
(A.14)
\[
E \left[ \sum_{\alpha \in A; \eta \notin \Delta(\alpha,N)} (X_\alpha - \mu) (X_\eta - \mu) \cdot \text{sign} \left( E \left[ (X_\alpha - \mu) (X_\eta - \mu) \mid (X_\eta - \mu) \right] \right) \right] = o(a_N)
\]
to ensure that
\[
\left| E \left[ \sum_{\alpha \in A; \eta \notin \Delta(\alpha,N)} (X_\alpha - \mu) (X_\eta - \mu) \cdot f' \left( \widehat{S}_\alpha \right) \right] \right| = o(1).\]

Note that (A.14) is weaker than (5.3), since if \( E \left[ (X_\alpha - \mu) (X_\eta - \mu) \mid (X_\eta - \mu) \right] \geq 0 \), then (using our tie-breaking convention from the lemma)
\[
\text{sign} \left( E \left[ (X_\alpha - \mu) (X_\eta - \mu) \mid (X_\eta - \mu) \right] \right) = 1,
\]
and so (A.14) becomes
\[
E \left[ \sum_{\alpha \in A; \eta \notin \Delta(\alpha,N)} (X_\alpha - \mu) (X_\eta - \mu) \right] = o(a_N),
\]
which is ensured by (5.3).
The second step is to apply the same reasoning as in Ross (2011), with an o(1) adjustment, to write

\[
|E \left[ f'(\mathcal{S}) - \mathcal{S} f(\mathcal{S}) \right] | \leq |E \left[ \frac{1}{a^{1/2}} \sum_\alpha (X_\alpha - \mu) (f(\mathcal{S}) - f(\mathcal{S}_\alpha) - (\mathcal{S} - \mathcal{S}_\alpha) f'(\mathcal{S})) \right] |
\]

\[+ |E \left[ f'(\mathcal{S}) \left( 1 - \frac{1}{a^{1/2}} \sum_\alpha (X_\alpha - \mu) (\mathcal{S} - \mathcal{S}_\alpha) \right) \right] | + o(1).\]

By a Taylor series approximation and given the bound on the derivatives of \( f \), it follows that

\[
|E \left[ f'(\mathcal{S}) - \mathcal{S} f(\mathcal{S}) \right] | \leq \frac{|f''|}{2a^{3/2}} \sum_\alpha E \left[ |X_\alpha - \mu| \left( \sum_{\eta \in \Delta(\alpha,N)} (X_\eta - \mu) \right)^2 \right]
\]

\[+ \frac{|f''|}{2a^{3/2}} \sum_\alpha E \left[ \mu \left( \sum_{\eta \in \Delta(\alpha,N)} (X_\eta - \mu) \right)^2 \right]
\]

\[+ \frac{|f''|}{2a^{3/2}} \sum_{\alpha,\eta \in \Delta(\alpha,N), \gamma \in \Delta(\alpha,N)} E \left[ \mu (X_\eta - \mu) (X_\gamma - \mu) \right]
\]

\[= \frac{|f''|}{2a^{3/2}} \sum_{\alpha,\eta \in \Delta(\alpha,N), \gamma \in \Delta(\alpha,N)} \left( E[X_\alpha X_\eta X_\gamma] - \mu E[X_\alpha X_\eta] - \mu E[X_\alpha X_\gamma] + \mu^2 E[X_\alpha] \right)
\]

\[+ \frac{|f''|}{2a^{3/2}} \sum_{\alpha,\eta \in \Delta(\alpha,N), \gamma \in \Delta(\alpha,N)} \mu \text{cov} [X_\eta, X_\gamma]
\]

\[= \frac{|f''|}{2a^{3/2}} \sum_{\alpha,\eta \in \Delta(\alpha,N), \gamma \in \Delta(\alpha,N)} \left( E[X_\alpha X_\eta X_\gamma] - \mu \left( \text{cov} [X_\alpha, X_\eta] + \mu^2 \right) - \mu \left( \text{cov} [X_\alpha, X_\gamma] + \mu^2 \right) + \mu^3 \right)
\]

\[+ \frac{|f''|}{2a^{3/2}} \sum_{\alpha,\eta \in \Delta(\alpha,N), \gamma \in \Delta(\alpha,N)} \mu \text{cov} [X_\eta, X_\gamma]
\]

\[= \frac{|f''|}{2a^{3/2}} \sum_{\alpha,\eta \in \Delta(\alpha,N), \gamma \in \Delta(\alpha,N)} \left( E[X_\alpha X_\eta X_\gamma] - \mu \text{cov} [X_\alpha, X_\eta] - \mu \text{cov} [X_\alpha, X_\gamma] - \mu^3 \right)
\]
Proof of Lemma 1. The lemma is essentially the consequence of standard results (e.g., Newey and McFadden (1994)). For convenience, define the expected moment function

$$H(\beta) := E_{\beta;\Theta} \left[ S^n - E_{\beta} \left[ S^n \right] \right]$$
and the empirical moment function

\[ \begin{aligned} \hat{H}(\beta) &:= \mathbb{S}^n - \mathbb{E}_B \left[ \mathbb{S}^n \right]. \end{aligned} \]

It is useful to make a change of variables via some diagonal normalizing matrix \( R_n \) to a parameter vector \( b_0 := R_n \beta_0^n \), so that the parameter vector does not change with \( n \). This is convenient for the argument, but not the natural parameter from the graph-theoretic perspective.

Observe that \( \delta (\hat{\beta}, \beta_0^n) \xrightarrow{P} 0 \) if and only if \( \hat{b} \xrightarrow{P} b_0 \), and consistency in the \( \delta \)-metric holds by assumption (1) of the Lemma.

Also define the expected and empirical moment functions in terms of this rescaled parameter,

\[ m(b) := \hat{H}(\beta = R_n^{-1}b) \]

and

\[ \hat{m}(b) := \hat{H}(\beta = R_n^{-1}b). \]

The first order condition satisfies the following, which by cross multiplying the normalizing matrix, can be rewritten in terms of the normalized parameter \( \hat{b} \). We can then take a mean value expansion around the true normalized parameter \( b_0 \)

\[ \begin{aligned} 0 &= \nabla \hat{H} \left( \hat{\beta}^n \right)' W_n \hat{H} \left( \hat{\beta}^n \right) \\
0 &= \nabla \hat{m} \left( \hat{b} \right)' W_n \left[ \hat{m} \left( b_0 \right) + \nabla \hat{m} \left( \hat{b} \right) \left( \hat{b} - b_0 \right) \right], \end{aligned} \]

by applying the mean-value theorem.\(^{36}\) Note that the mean value \( b \) is evaluated component by component in the matrix \( \nabla \hat{m} \left( \hat{b} \right) \). This abuse of notation is standard (e.g., Newey and McFadden (1994)).

We can solve for the \( \hat{b} - b_0 \) and multiply both sides by \( \sqrt{n^h} \)

\[ \sqrt{n^h} \left( \hat{b} - b_0 \right) = - \left[ \nabla_b \hat{m} \left( \hat{b} \right)' W^n \nabla_b \hat{m} \left( \hat{b} \right) \right]^{-1} \times \nabla_b \hat{m} \left( \hat{b} \right)' W^n \times \sqrt{n^h} \hat{m} \left( b_0 \right). \]

Observing that \( R_n \left( \hat{\beta} - \beta_0^n \right) = (\hat{b} - b_0^n) \), it follows that

\[ (A.15) \quad n^{h/2} R_n \left( \hat{\beta} - \beta_0^n \right) = - \left[ \nabla_b \hat{m} \left( \hat{b} \right)' W^n \nabla_b \hat{m} \left( \hat{b} \right) \right]^{-1} \times \nabla_b \hat{m} \left( \hat{b} \right)' W^n \times \sqrt{n^h} \hat{m} \left( b_0 \right). \]

Note that the last term is converging in distribution to a mean zero random variable. This follows by definition of the moment function and applying assumption (3),

\[ \sqrt{n^h} \hat{m} \left( b_0 \right) = \sqrt{n^h} \hat{H} \left( \beta_0^n \right) = \sqrt{n^h} \left( \mathbb{S}^n - \mathbb{E}_{\beta_0^n} \left[ \mathbb{S}^n \right] \right) \xrightarrow{D} \mathcal{N} \left( 0, \Sigma \right). \]

\(^{36}\)This is valid because \( b_0 \) is assumed to lie in the interior of \( B \), a compact set, which then implies the sequence of \( B^n \) under consideration.
If we can show that the preceding terms \( \left[ \nabla \widehat{m} \left( \widehat{b} \right) W^n \nabla \widehat{m} \left( \widehat{b} \right) \right]^{-1} \times \nabla_b \widehat{m} \left( \widehat{b} \right) \) converge in probability to non-degenerate limit matrices \( (M'WM)^{-1}M' \), then we can apply Slutzky’s theorem to obtain the result that \( n^{h/2}R_n \left( \widehat{\beta} - \beta_0^n \right) \) converges in distribution to a mean zero normal random vector as desired.

First, observe that
\[
\nabla \widehat{m}(b_0) = \nabla \widehat{H}(\beta_0^n) R_n^{-1} = M^n(b_0).
\]

Second, we show that
\[
\sup_{b \in B} |M_n(b) - M(b)| \to 0.
\]

There is pointwise convergence \( M_n(b) \to M(b) \) for every \( b \in B \) by assumption (4). Uniform convergence then follows from assumption (2): the fact that \( E(b/n)E[n] \) are continuously differentiable functions of \( b, b \in B \) where \( B \) is compact, and \( \nabla_b E[n] \) has a bounded derivative.

Third, by assumption (1) and (2), \( \| \tilde{b} - b_0 \| \overset{P}{\to} 0 \) and \( \nabla_b E[n] \) has bounded derivative, so we can write
\[
\nabla \tilde{b} \widehat{m} \left( \tilde{b} \right) = \nabla \tilde{b} \widehat{m} \left( b_0 \right) + o_p(1) = M^n(b_0) + o_p(1) = M + o_p(1),
\]

where the last step follows from pointwise convergence of \( M_n(b_0) \) to \( M = M(b_0) \).

Finally, by assumption (5), \( M_n' W^n M_n \) is invertible, as is the limit.

Taken together, these imply that
\[
\left[ \nabla \widehat{m} \left( \widehat{b} \right) W^n \nabla \widehat{m} \left( \widehat{b} \right) \right]^{-1} \times \nabla_b \widehat{m} \left( \widehat{b} \right) \overset{P}{\to} (M'WM)^{-1}M'
\]

which then completes our proof.

**Proof of Corollary 1.** We apply Theorem 1 to the case in which \( \Delta(\alpha, N) = \{\alpha\} \).

In this case, note that (5.1) becomes
\[
\sum_{\alpha} E \left[ X_\alpha^3 \right] = o \left( \left( \sum_{\alpha} \text{var}(X_\alpha) \right)^{3/2} \right).
\]

which is satisfied if
\[
N \mu^3 = o \left( N^{3/2} \text{var}(X_\alpha)^{3/2} \right).
\]
or
\[
N^{-1/3} \mu^2 = o \left( \text{var}(X_\alpha) \right).
\]

which is implied by (i).

Next, (5.2) becomes
\[
\sum_{\alpha, \alpha'} \text{cov} \left( (X_\alpha - \mu)^2, (X_{\alpha'} - \mu)^2 \right) = o \left( N^2 \text{Var}(X_\alpha)^2 \right).
\]
Note that the terms on the left-hand-side where $\alpha = \alpha'$ are equal to $N \text{var} (X_{\alpha})^2$, and so are inconsequential to satisfying the equation, and thus it becomes
\[ \sum_{\alpha \neq \alpha'} \text{cov} \left( (X_{\alpha} - \mu)^2, (X_{\alpha'} - \mu)^2 \right) = o \left( N^2 \text{var} (X_{\alpha})^2 \right), \]
which is condition (ii).

Finally, (5.3) becomes
\[ \sum_{\alpha \neq \alpha'} \text{cov} (X_{\alpha}, X'_{\alpha}) = o \left( N \text{var} (X_{\alpha}) \right), \]
which is condition (iii).

**Proof of Proposition 2.**

When obvious, we omit superscript $n$’s to simplify notation, but they are implicit.

Let $D_{\ell}(g)$ denote the set of all links which are deleted due to counting $\ell' < \ell$:
\[ D_{\ell}(g) = \{ij : ij \in g', g' \subset g, g' \in G_{\ell'}, \ell' < \ell\}. \]
For instance, $D_{\ell}(g)$ is the set of links that are members of triangles that appear in $g$ in the links and triangles SUGM, and therefore are not considered when counting links.

Note that under our relative sparsity condition, overall link presence vanishes$^{37}$, and so it then easily follows that $E[|D_{\ell}(g)|] = o_p(n^2)$ for any $\ell$. It follows that,

\[ (A.16) \quad \tilde{\beta}_\ell = \frac{\tilde{S}_\ell(g)}{\kappa_{\ell} \left( \begin{array}{c} n \\ m_\ell \end{array} \right) - D_{\ell}(g)} = \frac{\tilde{S}_\ell(g)}{\kappa_{\ell} \left( \begin{array}{c} n \\ m_\ell \end{array} \right)} (1 + o_p(1)) \]

\[ = \left( \frac{S_{\ell}^{\text{true}}}{\kappa_{\ell} \left( \begin{array}{c} n \\ m_\ell \end{array} \right)} + \frac{S_{\ell}^{\text{true}} - S_{\ell}^{\text{true}}}{\kappa_{\ell} \left( \begin{array}{c} n \\ m_\ell \end{array} \right)} + \frac{\tilde{S}_\ell(g) - S_{\ell}^{\text{true}}}{\kappa_{\ell} \left( \begin{array}{c} n \\ m_\ell \end{array} \right)} \right) (1 + o_p(1)), \]

where $S_{\ell}^{\text{true}}$ is the number of truly generated such subgraphs (unobserved) on the whole network, and $S_{\ell}^{\text{true}}$ is the number of truly generated such subgraphs (unobserved) on the networks that the after removing the links in $D_{\ell}(g)$, and $\left( \begin{array}{c} n \\ m_\ell \end{array} \right)$ counts the number of ways to pick $m_\ell$ nodes out of $n$, and $\kappa_{\ell}$ is the (finite number) of relabelings to count different subgraphs of type $\ell$ on a given set of $m_\ell$ nodes.$^{38}$

$^{37}$This can be checked from the condition directly, or simply by noting that if it did not, then the probability of any finite subgraph forming simply from link generation (and hence incidentally from whatever subgraphs form a particular link) would not vanish which would contradict parts of what is proven below.

$^{38}$For example, note that $\kappa_{\ell} = 1$ for a triangle but for a $K$-star it is $K$ since each star is different when a different member of the $K$ nodes is the center.
So, we show below that $|\bar{S}_t^{true} - S_t^{true}| = o_p(S_t^{true})$ and $|\bar{S}_t(g) - S_t^{true}| = o_p(S_t^{true})$; which then also implies that $\bar{S}_t(g) - S_t^{true} = o_p(S_t^{true})$. Together with (A.16), these tell us that

$$
\bar{\beta}_t = \left( \frac{S_t^{true}}{\kappa_t \binom{n}{m_t}} \right) (1 + o_p(1)).
$$

Given that the network is growing that $S_t(g)$ has a binomial distribution with parameter $\beta_{0,t}^n$, parts (1) and (2) follow directly.

To see this, let us define an estimator $\tilde{\beta}^{true}$ based on the $S^{true}$:

$$
\tilde{\beta}^{true}_t = \left( \frac{S_t^{true}}{\kappa_t \binom{n}{m_t}} \right),
$$

noting that this is a theoretical construct since $S^{true}$ is unobserved as noted above.

We know that

$$
\Sigma^{-1/2}(\tilde{\beta}^{true} - \beta_0^n) \sim N(0, I)
$$

where $\Sigma_{\ell,\ell} = \frac{\beta^n_0(1-\beta^n_0)}{\kappa_\ell \binom{n}{m_\ell}}$ and the off-diagonals are all 0. Then, since we have shown that $\bar{\beta} = \tilde{\beta}^{true}H$, where $H$ is a diagonal matrix with $H_{\ell\ell} = 1 + \varepsilon_\ell$, with $\varepsilon_\ell = o_p(1)$, (2) then follows.

So, to complete the proof we show that $|\bar{S}_t^{true} - S_t^{true}| = o_p(S_t^{true})$ and $|\bar{S}_t(g) - S_t^{true}| = o_p(S_t^{true})$.

To establish these claims, we establish two facts. One is that the probability that some observed subgraph of type $\ell$ was incidentally generated (by subgraphs that are no larger than it in the ordering) is $o_p(1)$. This establishes that $|\bar{S}_t(g) - S_t^{true}| = o_p(S_t^{true})$. The other is that a truly formed subgraph of type $\ell$ becomes part of an incidentally generated subgraph of type $\ell' < \ell$ is $o_p(1)$. This establishes that $|\bar{S}_t^{true} - S_t^{true}| = o_p(S_t^{true})$.

Let $z^n_\ell$ denote the probability that any given $g' \in G^n_\ell$ is incidentally generated. We now show that $z^n_\ell / \beta^n_{0,\ell} = o(1)$, which establishes the first claim.

Consider $g_\ell \in G^n_\ell$ and a (minimal, ordered) generating subclass $\mathcal{C} = (\ell_j, c_j)_{j \in J}$, and for which $\ell_j \geq \ell$ fr all $j$.

We show that the probability $z^n_\ell$ that it is generated by this subclass goes to zero relative to $\beta^n_{0,\ell}$ and since there are at most $M_\ell \leq k^{m_\ell}$ such generating classes, this implies that $z^n_\ell / \beta^n_{0,\ell} \to 0$.

Consider a subnetwork in $G^n_\ell$. The probability of getting at least one such network that has the $c_j$ nodes out of the $m_\ell$ in $g_\ell$ is no more than

$$
\kappa_{\ell_j} \binom{n}{m_{\ell_j} - c_j} \beta^n_{0,\ell_j} \leq \kappa_{\ell_j} n^{m_{\ell_j} - c_j} \beta^n_{0,\ell_j}.
$$

Then, we can bound the desired ratio by
\[
\frac{z_n^\ell}{\beta_n^0,\ell} \leq \frac{\prod_{j \in J} n^{m_{\ell_j} - c_j} \kappa_{\ell_j} \beta_n^0,\ell_j}{\beta_n^0,\ell} \leq \frac{\prod_{j \in J} n^{m_{\ell_j} - c_j} \kappa_{\ell_j} \beta_n^0,\ell_j}{n \sum_j c_j \beta_n^0,\ell_j}
\]
\[
\leq \frac{\prod_{j \in J} n^{m_{\ell_j} - c_j} \kappa_{\ell_j} \beta_n^0,\ell_j}{n M_C n^{m_{\ell_j}} \beta_n^0,\ell_j} \to 0,
\]
where the penultimate step follows from the fact that \(M_C = \sum_{j \in J} c_j - m_\ell\) and \(M_C \geq 1\) (since \(|J| \geq 2\) and some \(c_j\) intersects with at least one other set of \(c_j\) for some \(j' \neq j\), as the subgraph is not just isolated pairs of links) and the final step follows from the sparseness condition
\[
\frac{\prod_{j \in 1, \ldots, C} E[S_{\ell_j}^n]}{n^{M_C} E[S_{\ell_1}^n]} \to 0
\]
since the numerator of the final expression is of the order \(\Pi_{j \in 1, \ldots, C} E(S_{\ell_j}^n)\) while the denominator is of the order \(n^{M_C} E(S_{\ell_1}^n)\).

The second claim follows from a similar calculation. It is sufficient to show that the probability that some subgraph of type \(\ell_j\) becomes part of a subgraph of type \(\ell < \ell_j\) (where \(j' \in J\) is part of a generating class of some \(\ell < \ell_j\)), compared to the likelihood of the formation of a subgraph of type \(\ell_j\), is of vanishing order. Again, as there are a finite number of larger subgraphs, and a finite number of generating classes, it is sufficient to show this for a generic \(\ell < \ell_j\) and generic generating class. In the following, the numerator is on the order of the expected number of incidentally formed subgraphs of type \(\ell\) from this type of generating class, while the denominator is the expected number of the subgraphs of type \(\ell\).

\[
\frac{\kappa_{\ell'} \left( \frac{n}{m_j} \right) \prod_{j \in J} n^{m_{\ell_j} - c_j} \kappa_{\ell_j} \beta_n^0,\ell_j}{\kappa_{\ell'} \left( \frac{n}{m_j} \right) \beta_n^0,\ell_j} = \Theta \left( \frac{\prod_{j \in J} n^{m_{\ell_j} - c_j} \beta_n^0,\ell_j}{n \sum_{j=1, \ldots, C} c_j \beta_n^0,\ell_j} \right) \to 0.
\]

This convergence to 0 follows from the second part of the sparsity condition, which implies that
\[
\frac{\Pi_{j \in 1, \ldots, C} E_{\beta_n^0} (S_{\ell_j}^n (g))}{n^{M_C} E_{\beta_n^0} (S_{\ell_1}^n (g))} \to 0,
\]
for each \(j' \in 1, \ldots, C\).

**Proof of Lemma 3.** 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.\(^{39}\) Thus, the relative meeting
frequency of different type links compared same type links is
\[
\frac{\pi_L(dif f)}{\pi_L(same)} = \frac{1 - (f^2 + (1 - f)^2)}{f^2 + (1 - f)^2}.
\]
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 \( 1 - (f^2 + (1 - f)^2) \) of them being of mixed types, and so
\[
\frac{\pi_T(dif f)}{\pi_T(same)} = \frac{1 - (f^3 + (1 - f)^3)}{f^3 + (1 - f)^3}.
\]
It follows directly that for \( f \in (0, 1) \):
\[
(A.17) \quad \frac{\pi_T(same)}{\pi_T(dif f)} < \frac{\pi_L(same)}{\pi_L(dif f)}.
\]
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
\[
\frac{p_T(dif f)}{p_T(same)} < \frac{p_L(dif f)}{p_L(same)} \text{ if and only if } \left( \frac{P_T(dif f)}{P_T(same)} \right)^{1/3} < \left( \frac{P_L(dif f)}{P_L(same)} \right)^{1/2}.
\]
In summary, given (A.17), a sufficient condition for \( \frac{p_T(dif f)}{p_T(same)} < \frac{p_L(dif f)}{p_L(same)} \) is that
\[
(P_T(dif f)/P_T(same)) < (P_L(dif f)/P_L(same))^{3/2}
\]
which completes the argument.