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—subgraph generated models (SUGMs). Various subgraphs—e.g., links, triangles, cliques, stars—are generated and their union results in a network. We show that SUGMs are identified and establish the consistency and asymptotic distribution of parameter estimators in empirically relevant cases. We show that a simple four-parameter SUGM matches basic patterns in empirical networks more closely than four standard models (with many more dimensions): (i) stochastic block models; (ii) models with node-level unobserved heterogeneity; (iii) latent space models; (iv) exponential random graphs. We illustrate the framework's value via several applications using networks from rural India. We study whether network structure helps enforce risk-sharing and whether cross-caste interactions are more likely to be private. We also develop a new central limit theorem for correlated random variables, which is required to prove our results and is of independent interest.

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

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

Date: Revision: November 25, 2024.

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 minimum distance that were not part of the original paper. We thank Alberto Abadie, Isaiah Andrews, Emily Breza, Aureo de Paula, Paul Goldsmith-Pinkham, Bryan Graham, Han Hong, Guido Imbens, Michael Leung, Shane Lubold, Elena Manresa, Tyler McCormick, Angelo Mele, Joe Romano, Elie Tamer, and the referees, as well as seminar participants, for helpful comments and suggestions. We thank Shreya Chaturvedi, Vasu Chaudhary, Shobitha Cherian, Andres Drenik, Anoop Singh Rawat, and Meghna Yadav for valuable research assistance. Chandrasekhar is grateful for support from the NSF Graduate Research Fellowship Program, NSF grant SES-1156182, and the Alfred P. Sloan Foundation. Jackson gratefully acknowledges financial support from the NSF under grants SES-1629446 and SES-2018554 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; member of J-PAL.

[⋆]Department of Economics, Stanford University; external faculty member of the Santa Fe Institute.

1. Introduction

Networks of interactions impact many economic behaviors including insuring one's self (e.g., Cai, deJanvry, and Sadoulet (2015)), participating in microfinance (e.g., Banerjee et al. (2013)), educating one's self (e.g., Calvo-Armengol, Patacchini, and Zenou (2009); Carrell, Sacerdote, and West (2013)), and engaging in criminal behavior (e.g., Glaeser, Sacerdote, and Scheinkman (1996); Patacchini and Zenou (2008)). Networks of interactions are also essential to understanding financial contagions (e.g., Gai and Kapadia (2010); Elliott, Golub, and Jackson (2014); Acemoglu, Ozdaglar, and Tahbaz-Salehi (2015)), as well as world trade (e.g., Chaney (2016)), inter-state war (e.g., Jackson and Nei (2015); König, Rohner, Thoenig, and Zilibotti (2017)), and a host of other economic phenomena. As such, the structure that a network takes has profound consequences—changing the possibility of contagions, the decisions that people make, and the beliefs that people hold—and so it essential to understand and estimate network formation. Moreover, networks are of interest precisely because there are externalities—one agent's behavior impacts the welfare and behaviors of others.[1] This feature means that connections between agents are not independent, and so appropriate models of network formation must admit correlations in connections.

Despite the importance of network formation, general, flexible, and tractable econometric models for the estimation of network formation are lacking. This stems from two challenges: the aforementioned dependence in connections and the fact that many studies involve one (large) network. Thus, one is often confronted with estimating a model of formation by taking advantage of the large number of connections, but having them all be dependent observations. Despite the dependence, it is possible that the many relationships in a network still provide rich enough information to consistently estimate the parameters of a network model and test hypotheses from a single observed network, at least hypothetically. Here we develop a class of models that admit correlations in links and also provide practical techniques of estimating the models, showing that they are estimable, even with just a single network.

Before describing our model, it is useful to discuss some of the alternative approaches.

1.1. Alternative Models of Network Formation.

The most basic models are what are known as 'stochastic block models', in which links may depend on node characteristics but are (conditionally) independent of each other. That approach requires correlation between links to be well-approximated by observable characteristics, and may not be sufficient for most applications.[2] In particular, stochastic block models are not a good option for estimation in applications in which there is substantial clustering (triangles) or other cliques in the network. In fact, in Section 5 we show that our model (even with only four parameters) models the graph structure of real-world data better than a stochastic block model even when the block model admits a rich set of covariates and unobserved node level heterogeneity (fixed effects)

[1] For detailed discussions see Jackson, Rogers, and Zenou (2016) and Jackson (2019).

[2] A variation on this is community detection where nodes are estimated to belong to certain groups, though this calculation is NP-hard. See Bickel et al. (2011) for a "non-parametric view" of network formation, and Jackson and Storms (2017) for an approach to estimating the blocks even if they are latent.

(Chatterjee et al., 2010; Graham, 2017).[3] Although there are challenges in taking such models to data, they are useful if link correlation is not a concern.

Given the importance of clustering and other local network architectures in many applications, a literature spanning several disciplines (sociology, statistics, economics, and computer science) turned to using exponential random graph models—henceforth "ERGMs". ERGMs admit link interdependencies and have become the workhorse models for estimating network formation.[4] However, from the onset of the use of ERGMs, researchers realized that the parameter estimators could be unstable on all except excessively small networks. It has been shown that maximum likelihood and Bayesian estimators are not computationally tractable (the required Gibbs sampler will take exponential time to mix) nor consistent for important classes of such models—and in particular for the ERGMs that include many link dependencies of interest (and neither parameter estimators nor standard errors can be trusted). For details see Bhamidi, Bresler, and Sly (2008); Shalizi and Rinaldo (2012); Chandrasekhar and Jackson (2012).[5]

A set of models that allow for link dependencies and are estimable is the class of models based on explicit link formation algorithms (e.g., Barabasi and Albert (1999); Jackson and Watts (2001); Jackson and Rogers (2007); Currarini, Jackson, and Pin (2009, 2010); Christakis, Fowler, Imbens, and Kalyanaraman (2020); Bramoullé, Currarini, Jackson, Pin, and Rogers (2012)). These models can be estimated since the algorithms are particular enough so that one can directly derive how parameters in the model translate into aggregate network statistics, such as the degree distribution or homophily levels. The advantage of such models is that a specific algorithm allows for estimation. The disadvantage is that the specificity of the algorithms also necessarily results in highly-structured models. Thus, these approaches are useful in some contexts, but they are not designed, nor intended, for general statistical testing of a wide variety of network formation models and hypotheses. For instance, such models cannot generate considerable triadic closure (where links correlated across triples of nodes—so if two people have a friend in common, are they more likely to be friends with each other than if link formation were independent).[6]

Another approach has roots in the spatial statistics literature. Such models organize nodes such that pairs can be evaluated in terms of distance, with linking probabilities decaying in distance. The distance may be latent (unobserved) or in observed characteristic space

[3]In fact, correlations can be viewed as driven by unobserved heterogeneity (Chatterjee, Diaconis, and Sly, 2010), which has links be uncorrelated conditional on all (observed and unobserved) characteristics (as extended by Graham (2017)). See also Charbonneau (2017) for related work that is a directed networks version of Graham (2017). Such models have been studied in the mathematics and statistics literatures (e.g., Holland and Leinhardt (1981); Park and Newman (2004); Blitzstein and Diaconis (2011)).

[4]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 is to work with regression models at the link level, but to allow for dependent error terms, as in the "MRQAP" approach (e.g., Krackhardt (1988)).

[5]Recent work has made progress on both the speed of convergence of estimation algorithms as well as characterizing the asymptotic distribution of sufficient statistics in some classes of ERGMs that avoid extensive link dependencies (see e.g., Mele (2017, 2022); Mele and Zhu (2023)).

[6]The Jackson and Rogers (2007) model does have a parameter that affects triadic closure, but in that model closure cannot be separated from the shape of the degree distribution. So, it is best suited for growing random networks where new nodes are born over time.

(such as geography or demographics). Such models have foundations in the mathematics literature on random geometric graphs (Penrose, 2003)—where nodes are distributed in a latent space according to some Poisson point process and linking is much more likely among proximate nodes—and have been analyzed in the statistics literature in work on latent space models such as in Hoff et al. (2002). Links between distant-enough pairs of nodes are asymptotically independent and such models have been developed in more detail in the econometrics literature (e.g., Boucher and Mourifié (2017); Leung (2014)). This approach holds promise for some enormous networks with appropriate spatial structures—in which the graph can almost be decomposed into independent pieces.[7] However, there are many applications for which these latent space (and generally spatial) models—particularly the geometry of the space the nodes—overly dictates and limits the structure of link correlation. Using these models may in fact require estimating an unobserved manifold, which presents its own challenges.[8] Our model dispenses with these problems in a straightforward way, allowing correlations across nodes but not forcing correlations generated through distances in unobserved or characteristic space.

Finally, there is a large literature on the theory of network formation from a strategic perspective (for references, see Jackson (2005, 2008)). Since the first writing of this paper, researchers have started to derive versions of such models that can be taken to data. One approach builds upon the relationship between certain classes of strategic network formation models and potential games; some of which leverage subgraphs, but in a rather different way from us (Butts (2009); Mele (2017); Badev (2021)). Another derives restrictions on parameters of an observed network under the presumption that it is in equilibrium (pairwise stable) (De Paula, Richards-Shubik, and Tamer (2018); Sheng (2020)). The latter makes the observation that by using pairwise stability restrictions of Jackson and Wolinsky (1996) on subnetworks, one can partially identify preference parameters in the model, whereas doing so on the full graph can be computationally infeasible.[9,10] Although the progress to date requires strong restrictions on how links can enter agent's payoffs, they provide important first steps in deriving implications of the arsenal of strategic network formation models. Below, we also provide ways to incorporate strategic formation in SUGMs, thus in part bridging our approach here and the strategic formation approach.

1.2. **Our Subgraph Model Approach.** Our approach is distinct from all of the above, both in terms of the approach (working with subgraphs as the basic building blocks) and the technicalities of allowing nontrivial conditional correlations. Our contribution is to develop models of network formation that admit considerable interdependency without spatial restrictions, and still prove consistency and asymptotic normality of parameter estimators. As part of this, we develop a new central limit theorem for non-trivially correlated random

[7]McCormick and Zheng (2015) merge the insights from the unobserved heterogeneity and the latent space distance models, and Breza, Chandrasekhar, McCormick, and Pan (2020) evaluate its empirical performance.

[8]Estimating the latent geometry by using the network data to identify the underlying metric signature is the subject of one of the authors' related work in (Lubold et al., 2023).

[9]For a recent overview of the recent literature, see De Paula (2017).

[10]In both the potential games and the partial identification literatures, subgraphs play very different roles from their role here.

variables that moves away from relying on spatial-style mixing arguments that force decaying dependence in distance.

The paucity of flexible models that are computable and can be used across many applications for hypothesis testing and inference is what motivates our work here. Although our models are simple conceptually, we provide different applications that illustrate how such models admit strategic network formation, general covariates, and generate rich network features.

In Section 2 we introduce *subgraph generated models*—henceforth SUGMs. In these models, various subgraphs (e.g., links, triangles, cliques, and 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; villagers may form specific bilateral or multilateral agreements independently, each to sustain some collection of favors between those individuals involved in the agreement. 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. In this section, we also introduce three motivating applications to demonstrate how this model could be used: (i) descriptively modeling network structure, (ii) motives for risk-sharing, and (iii) incentives to link across social boundaries.

The statistical challenge is that often only the final network is observed: a survey may ask people to list their friends or acquaintances, or links may be observed on a social platform, or emails or phone calls are observed, etc., but the original formation process is often not observed. Subgraphs may overlap and may incidentally generate new subgraphs: e.g., three links may form and result in a triangle. Thus, the true rate of formation of the subgraphs cannot generally be inferred just by counting their presence in the resulting network.[11]

Despite this, in Section 3 we prove that every subgraph generated model is identified. That is, if we consider a SUGM—a collection of subgraphs that can potentially form together with a set of parameters governing the probabilities of each subgraph forming—then any two distinct set of parameters necessarily has two distinct set of distributions over the set of possible networks. Furthermore, we explore specific cases that are of empirical relevance—for instance, links and triangles models—and demonstrate that not only are the distributions generally distinct, but that simple statistics (such as the share of links or triangles that form) allow us to identify the parameters of interest.

Next we turn to estimation of the underlying parameters describing subgraph formation rates in Section 4. We show that we can consistently estimate the parameters and we derive the asymptotic distribution of the estimators so we can conduct inference. There are two situations that a researcher may face.

In the first case, the researcher has access to "many networks". This could be because they have collected network data from numerous schools, many villages, or so on. Here

[11]The closest work to ours is a Bollobás et al. (2011) piece on random graph theory, which looks at percolation processes, giant component structure, and degree distributions in a model where the observed graph is generated by a set of atoms (subgraphs in our language). That paper focuses on a specific rate of arrival of subgraphs (to maintain a sparsity where a core problem we study is ruled out) and is not interested in statistical estimation.

we demonstrate (using standard results) that the parameters governing the SUGM can be estimated consistently with maximum likelihood estimators that are asymptotically normally distributed. For some empirically relevant classes of models, we demonstrate that there are computationally simple, minimum distance estimators which satisfy consistency and asymptotic normality.

The second case is where the researcher has one (or just a few) "large network". This could be because they have collected very rich network data with resource constraints in just a few communities, or because they are looking at a single market, or because they are looking at one social media platform, etc. In this case, the asymptotics are more technically challenging for two reasons. First, the network cannot be too sparse, as enough subgraphs must form to make estimation possible, nor too dense because it becomes impossible to distinguish which subgraph likely generated a candidate link. So formally, we have "rate requirements" on the parameters governing the probabilities of subgraphs forming, although these turn out to be quite accomodating. Second, 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. We overcome this problem by developing a new central limit theorem and use it to characterize when certain classes of SUGMs have estimators that are consistent and asymptotically normally distributed.[12]

With the statistical properties established, we turn to our empirical applications in Section 5. In each application we use the detailed network data we collected in 75 villages in Karnataka, India (Banerjee et al., 2019). We begin by comparing SUGMs to four archetypical models from the literature in terms of how well they model real-world data. Specifically, we fit each model to the data and then draw from the distribution at the estimated parameters for each model. We are interested in a variety of economically relevant network features (none of which are directly used to estimate any of the models). We find that across the board a four parameter SUGM outperforms a stochastic block model with flexible covariates; a model of unobserved heterogeneity at the node level as well as rich covariates; a latent space model with unobserved locations and heterogeneity as well as observed covariates; and an exponential random graph model with rich covariates. Only the SUGM comes close to capturing the average path length, homophily, maximal eigenvalue, size of the giant component, isolates, and clustering. Having established this, the second example turns to whether the structure of the networks is consistent with the idea that there are stronger incentives to have supported relationships for risk sharing links rather than informational links (Jackson et al., 2012) and we find evidence consistent with this. The third example explores whether linking across social boundaries—here links between upper caste and lower caste (Dalit communities)—is more likely to form in private (bilateral) rather than group (triadic) settings and we find exactly this. Together, these examples demonstrate the utility of our general framework.

[12]An interesting consideration for future work is to employ the techniques in Bhattacharyya et al. (2015), who develop a bootstrapping method to estimate the distribution of empirical counts of different subgraphs in enormous networks.

In Section 6 we return to state our Central Limit Theorem, which is of independent interest. We provide covariance conditions that are high-level but also straightforward to interpret, check, and micro-found. We use a powerful lemma from Stein (1986) in our proof. Many CLTs build upon Stein's method,[13] but we allow for much richer dependence—all random variables can have non-zero correlation—which admits the correlations in subgraph counts that can arise due to the incidental generation discussed above; and we also allow for triangular arrays. We discuss the relationship of our Central Limit Theorem and its proof to precursors in Section 6.

2. A Model of Network Formation via Subgraphs

2.1. **Networks.** $n \geq 3$ is the number of nodes on which a network is formed. Nodes may have characteristics, such as age, profession, gender, race, caste, etc., that we denote by the vector X_i for a generic $i \in \{1, \ldots, n\}$. The X_i have finite support.[14] As such nodes can be classified by a finite set of observable types.[15]

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. 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 refer to the undirected case. $\mathcal{G}^n$ denotes the set of all networks on n nodes (which given our definition of a network above, is the set of all labeled (undirected) graphs on the set $\{1, \ldots, n\}$).

2.2. **Subgraphs and SUGMs.** In a subgraph generation model, subgraphs are each 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

[13] For instance, Bolthausen (1982) uses a pre-cursor lemma from Stein (1972) to derive CLTs from some mixing conditions. In time-series and spatial econometrics, a non-exhaustive but illustrative list of papers using Bolthausen (1982) include Conley (1999), Jenish and Prucha (2009), Bester, Conley, and Hansen (2011), among others.

[14] This is a limitation since there are network models that do not require discrete covariates. While continuous variables can be discretized, this is a trade-off.

[15] We conjecture that our results extend to allow for continuous covariates as well, though that requires specifying parametric functions for the probability of subgraphs as a function of covariates and so remains beyond the scope of this paper. If one expands the set of covariates, the number of parameters to fit increases. In the limit, if one allows continuous covariates, one then has to fit functions for every type of subgraph (e.g., probability of a triangle as a function of the covariates of the three nodes). That is only simplified if one imposes restrictions on those functions (e.g., some form of linearity).

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.

A SUGM is formally defined as follows.

- There are finitely many types of nonempty subgraphs, indexed by $\ell \in \{1, \ldots, k\}$.
- Each of the k subgraph types corresponds to a set $(G_\ell)_{\ell \in \{1,\ldots,k\}}$, where each $G_\ell \subset \mathcal{G}^n$ is a set of possible subgraphs on $m_\ell \leq n$ nodes.
- For each ℓ and pair of subgraphs $g' \in G_\ell$ and $g'' \in G_\ell$ there exists a bijection π on $\{1, \ldots, n\}$ such that $ij \in g'$ if and only if $\pi(i)\pi(j) \in g''$.
- No subgraph is contained in two different sets: if $\ell \neq \ell'$ then $g \in G_\ell$ implies that $g \notin G_{\ell'}$.

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 examples below as well as the supplementary appendix. As an example, in the links and triangles case $\ell \in \{L, T\}$. Then, for $n = 4$ and $\ell = T$, G_T is the set of triangles, where $m_\ell = 3$ and $G_T = \{\{12, 23, 31\}, \{12, 24, 41\}, \{13, 34, 41\}, \{23, 34, 42\}\}$.

Note, however, that the definition does not require that G_ℓ contain all triangles. In examples in which node characteristics matter, different triangles could be categorized into different G_ℓs. In particular, definitions of the subgraph types can have restrictions based on node characteristics, for instance, requiring that the characteristics X_i and $X_{\pi(i)}$ be the same—e.g., G_ℓ for some ℓ could be the set of "triangles that involve one child and two adult nodes". As another example, the set G_ℓ for some ℓ could be all stars with one central node and four other nodes, and another $G_{\ell'}$ could be all the links that involve people of different castes, and so forth.

A few examples are pictured in Figure 1.

The probability that various subgraphs form is described by a vector of parameters, denoted $\beta = (\beta_1, \ldots \beta_\ell, \ldots, \beta_k) \in \mathcal{B}^k$, where $\mathcal{B}$ is (unless otherwise noted) a compact subset of $[0, 1)^k$.[16] For instance, $\beta = (\beta_L, \beta_T) \in \mathcal{B} \subset [0, 1)^2$ in a links and triangles example.[17]

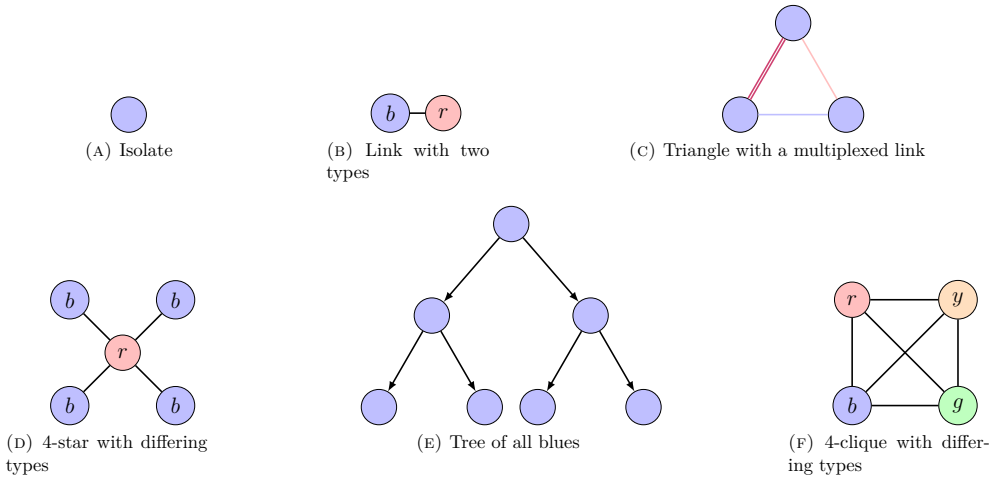
A network g on n nodes is randomly formed as follows:

(1) Each of the possible subgraphs $g_\ell \in G_\ell$ forms with probability β_ℓ independently of all other subgraphs (including others in G_ℓ).
(2) The resulting network, g, is the union of all the links that appear in any of the generated subgraphs.

2.3. **An Example with Node Characteristics.** Suppose that nodes come in two colors: blue and red (for instance different genders, age groups, religions, etc., and clearly this extends directly to more than two colors). In our example of links and triangles, there

[16]We treat vectors as row or columns as is convenient in what follows.

[17]In some examples below, we expand this demonstrating how β can have entries that are monotone functions of preference parameters (or equilibrium behavior), which allows us to study certain economic questions. Estimating β allows us to either recover the parameters or behavior of interest in some cases or conduct loose hypothesis testing using our estimates of β.

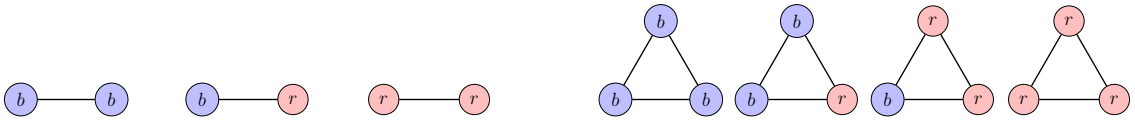


(A) Isolate (B) Link with two types (C) Triangle with a multiplexed link

(D) 4-star with differing types (E) Tree of all blues (F) 4-clique with differing types

FIGURE 1. Examples of subgraphs. Links could be directed or undirected or even multiplexed (take on multiple edge types) and nodes can have different characteristic combinations (denoted by node colors and labels).

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) which comprise the set of subgraphs indexed by ℓ.



(A) All Links (B) All Triangles

FIGURE 2. Panel (A) shows all possible links and Panel (B) shows all possible triangles when a node has characteristic $X_i \in \{red,\ blue\}$.

Thus, in this example the sets of subgraphs are

$$G_{(blue,blue)} = \{ij : X_i = blue, X_j = blue\}$$

and

$$G_{(blue,blue,red)} = \{ijk : X_i = blue, X_j = blue, X_k = red\},$$

and so forth, as depicted in Figure 2. The parameters

$$\{\beta_{(blue,blue)}, \beta_{(blue,red)}, \beta_{(red,red)}, \beta_{(blue,blue,blue)}, \beta_{(blue,blue,red)}, \beta_{(blue,red,red)}, \beta_{(red,red,red)}\},$$

are the probabilities that the corresponding subgraphs form.

One could restrict or enrich the model by having simpler or more complex sets of parameters – for instance requiring that $\beta_{(blue,blue)} = \beta_{(red,red)}$, or by having preference parameters that govern the probabilities of various subgraphs forming, as we discuss below.

2.4. **Links and Triangles as Our Leading Example.** The bulk of our illustrations and applications are based on link and triangle SUGMs, though other subgraphs can be included and are covered by our general results (e.g., Theorems 1, 2, and 3). Our illustrations focus

on links and triangles for two reasons: first, this case is simple to understand and illustrates the main points since it exhibits correlated links and incidental generation; second, the link and triangle model already matches the moments that are of interest in many research projects (larger cliques are rare). In fact, as we show below, simply looking at a links and triangle SUGM tagged with whether the nodes involved are homogenous or heterogeneous in demographics (e.g., just a 4 parameter model), replicates real-world network features better than far-richer models. Still, we leave further specification to the researcher as it will depend on their context and the phenomenon being modeled. If there are other the types of subgraphs that are hypothesized to arise in some particular context, then that model can be constructed and estimated in the ways we outline and are covered by our general results.[18]

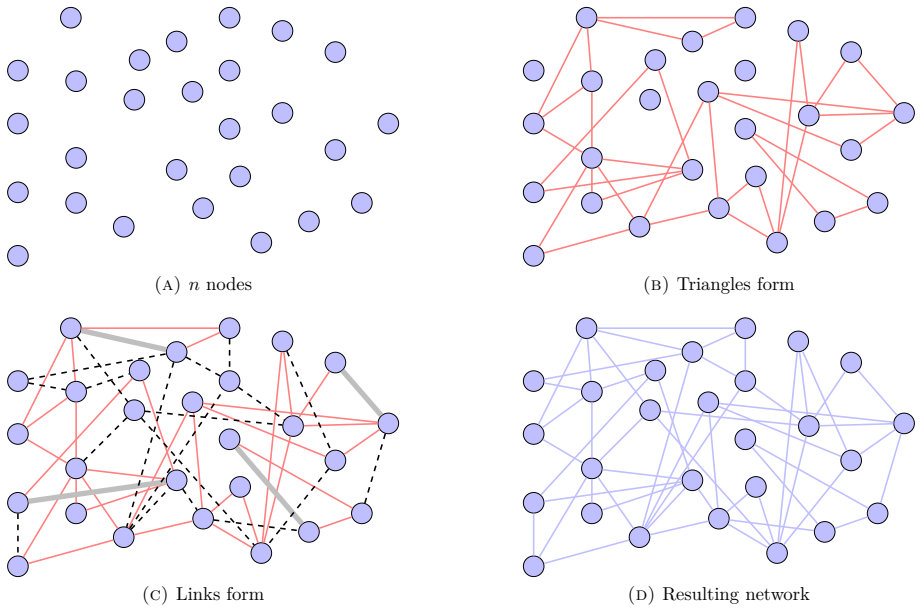
3. Identification

3.1. **The Challenge of Identification.** The researcher's goal is to use the observed data—from one or more networks—to recover the parameters of interest, for example, the (β_L, β_T) in a SUGM of links and triangles. If the researcher observed the links and triangles that were formed directly, then estimation would be straightforward. Indeed, in some instances a researcher has direct 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 and thus observes papers with two authors, three authors, and so forth. Instead, for instance, it may be that there are groups of three people who commonly share favors and risks together—who really form a triangle, but the researcher only has information from a survey asking with which alters a given person interacts (as in networks derived from the Add Health data set as in Currarini et al. (2009)), or who borrows from whom and who lends kerosene and rice to whom and other bilateral nominations (as in our Indian village data Banerjee et al. (2013)), or from observing that who are friends on a social platform (as in Facebook network data as in Bailey et al. (2016); Chetty et al. (2022)), or from observing that two people phone each other or remit payments to each other (as in Blumenstock et al. (2011)).

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. For example, if three links are observed between i, j, and k, 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 3 provides an illustration.

The overlap and incidental generation present 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

[18]One could also have a list of subgraphs as a possible basis for the SUGM with only a subset of them actually forming the true SUGM; allowing the data to tell the researcher which to include. Some of that can be done here, including the various subgraphs that might be involved and then seeing which have nontrivial parameter estimates. This marries SUGMs with model selection, a topic which could be explored further in future research.



(A) n nodes (B) Triangles form (C) Links form (D) Resulting network

FIGURE 3. The network that is formed and eventually observed is shown in panel D. The process comes from forming triangles with probability β_T as in (B) in red; and forming links, in grey, with probability β_L as in (C)—all independently. 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.

end up as parts of triangles. We show that despite this difficulty, the parameters can be recovered by careful study of the observed patterns. In particular, we show that a SUGM is *always* identified, and also provide techniques for recovering the parameters.

3.2. **A General Identification Result.** We first show that as the parameters of any SUGM change, so does the distribution over networks, and hence SUGMs are identified models.

Let P_β denote the probability distribution over a network g on n nodes under a vector of parameters β describing the probabilities of subgraph types $(G_\ell)_{\ell\in\{1,\dots,k\}}$.

THEOREM **1.** *Every SUGM is identified. That is, for any finite collection of distinct types of subgraphs $(G_\ell)_{\ell\in\{1,\dots,k\}}$ on n nodes, $\beta \neq \beta' \implies \mathrm{P}_\beta \neq \mathrm{P}_{\beta'}$.*

Recalling the general definition of the SUGM, this means that for every SUGM (even one comprised of subgraphs that could have nodes with varying (discrete) covariates and allowing for multiplexing, etc.) is identified.

To understand why this holds, for instance in the case of links and triangles, note that as one varies (β_L, β_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. Theorem 1 shows that this is true not only for links and triangles, but for any collection of distinct subgraphs.

We emphasize, of course, that identification does not imply that the parameters are easily estimated, especially on a very small number of nodes. We provide results on consistency below, which require observation of a sufficiently large network and/or sufficiently many networks.

3.2.1. *Identification from Link and Triangle Counts.* Although Theorem 1 shows that SUGMs are always identified—i.e., distinct parameters yield distinct distributions—it is often convenient to use minimum distance based estimators based on simple moments of the network. Thus, it is useful to show that identification can be achieved from simple statistics. We illustrate that this can be done with direct counts of the relative frequency of appearances of the subgraphs. In particular, in Proposition 1 we show that a links and triangles SUGM can be identified directly from the counts of links and triangles: $S(g) = (S_L(g), S_T(g))$. This does not mean that one can ignore incidental generation, but it does mean that the information one has to use can be simple counts.

Further below, in Theorem 3, we show conditions under which such direct counts not only identify the parameters for general subgraphs, but can also be used to derive consistent and normally distributed estimators of the parameters.

To understand the identification, consider Figure 4. Each configuration involves two triangles, but the graph in Panel B with only five links is *relatively* more easily incidentally formed than the one in Panel A. Thus, by looking at the combination of how many triangles and how likely links there are, we can sort out relative rates of the two parameters.

(A) Node adjacent triangles

(B) Edge adjacent triangles

FIGURE 4. Two different configurations of two triangles; one has a count of 6 total links and the other has a count of 5 links. (A) is more relatively more likely to come directly from the formation of two triangles, and (B) is relatively more likely to come from a combination of links and triangles. The likelihoods of links and triangles can thus be deduced via careful deductions from the combination of the counts of links and triangles.

PROPOSITION **1.** *A SUGM of links and triangles is identified from moments* $S(g) = (S_L(g), S_T(g))$ *for any* $\beta = (\beta_L, \beta_T) \in [0,1)^2$. *That is, if* $(\beta'_L, \beta'_T) \neq (\beta_L, \beta_T)$ *then* $\mathrm{E}_{\beta'}[S(g)] \neq \mathrm{E}_{\beta}[S(g)]$.

Let us outline the basic ideas behind the proof, with the full proof appearing in the appendix. Let $\widetilde{q}_L(\beta_L, \beta_T)$ denote the probability that any given link forms conditional upon

exactly one particular triangle that it could be a part of not forming, which depends on the βs. For instance, for nodes ij it is the probability that ij is formed either as a link or as part of a triangle that is *not* triangle hij for some other node h. Although this is not an immediately obvious parameter to define, it allows us to write the probability that a given link forms as $\beta_T + (1-\beta_T)\widetilde{q}_L(\beta_L, \beta_T)$. This expression turns to be useful as it helps us to compare the rate at which links form to the rate at which triangles form in a way that shows how they are identified. In particular:

$$\mathrm{E}_{\beta_L,\beta_T}[S_L(g), S_T(g)] = \left[\beta_T + (1-\beta_T)\widetilde{q}_L(\beta_L,\beta_T), \beta_T + (1-\beta_T)\left(\widetilde{q}_L(\beta_L,\beta_T)\right)^3\right]. \tag{3.1}$$

For instance, note that the term $\beta_T + (1-\beta_T)\left(\widetilde{q}_L(\beta_L,\beta_T)\right)^3$ is the probability that a triangle forms, either directly (β_T), or does not form directly ($1-\beta_T$) but then each of the links then forms on its own $\left(\widetilde{q}_L(\beta_L,\beta_T)\right)^3$.[19] This is helpful in showing how different parameters lead to different rates of formation of links and triangles since we can isolate the difference via the $\widetilde{q}_L(\beta_L,\beta_T)$ versus $\left(\widetilde{q}_L(\beta_L,\beta_T)\right)^3$ expressions.

Analogs of this proposition extend to cases with covariates and multiplexing, simply with more complicated extensions of (3.1) accounting for the specific types of triangles or links. Also, a general version of asymptotic identification is a by-product of Theorem 3, below.

4. Estimation and Asymptotics

We now provide conditions under which various estimators of the parameters are consistent and describe their asymptotic distributions. We consider two asymptotic frames, in which at least one of either the size of the network or the number of networks becomes large enough for consistent estimation. We discuss two different estimators for each frame for a total of four estimators.

4.1. Data and Asymptotic Frames.

Suppose that the researcher observes $R \geq 1$ independently, and identically drawn graphs $(g_1, \ldots, g_R)$, on at least n nodes each, drawn from a SUGM with a list of k subgraphs and parameters $\beta \in [0,1)^k$. Each of the k subgraphs involves no more than n nodes. For simplicity in notation, we work with each network having exactly n nodes, but one can directly extend the results by simply selecting n nodes for each network and applying all of our estimation to those subgraphs.

The first asymptotic frame, studied in Section 4.2, covers situations in which the number of different realizations of networks R tends to infinity. Here researchers have access to many networks and the empirical moments of interest converge to their expectations via observation of independent networks. This applies when a researcher is studying, for instance a number of schools, classrooms, villages, etc. In this case estimation and inference is straightforward. There are a growing number of independent draws from the distribution and we have already proven identification in Theorem 1. Our Theorem 2 shows that the maximum likelihood estimator from R networks—which we denote by $\hat{\beta}_R^{\mathrm{ML}}$—is consistent and asymptotically normally distributed as R grows.

[19] Conditional upon the triangle not forming directly, the links are then independent.

Given the difficulty in calculating the likelihoods for networks, also consider a second computationally-simpler minimum-distance estimator (presented for the case of links and triangles), denoted by $\hat{\beta}_R^{\text{MD}}$. We show in Proposition 2 that this minimum distance estimator is consistent and asymptotically normally distributed.

The second asymptotic frame is studied in Section 4.3 and it holds the number of networks observed R fixed, without loss of generality at $R = 1$, and then lets the number of nodes grow: $n \to \infty$. Examples include when the researcher has detailed information about a large community, friendships on social media platform, citation networks, etc. Clearly, this extends to cases with large n and more than one network, but we consider $R = 1$ for ease of notation. This is the more challenging perspective as the observations of various parts of a network are not independent. Also, the identification result from Theorem 1 does not guarantee that the empirical moments converge to their expectations in a single large network.

There are two cases of interest with a single large network. The first is what we call the sparse case (which we explicitly characterize), and this is a situation in which certain types of incidental generation of subgraphs become asymptotically negligible. For the sparse case, we prove that identification and asymptotic consistency and normality is possible from an easy variation on direct counts of observed subgraphs. Namely, one begins with the largest subgraph in the model, count how many of them are present, then remove links associated with them and step down to the next largest and so on. The estimator corresponding to this procedure is what we call a direct count estimator—denoted by $\breve{\beta}_n^{\text{DC}}$—as it is essentially directly calculating the linking rate for each subgraph type. We prove the consistency and asymptotic normality of the direct count estimator under suitable sparsity conditions in Theorem 3.

It is possible to verify whether a network is sparse enough to permit the direct estimator in the following way. One can take relevant parameter values for the SUGM (which one can find by a first crude estimation from the data) and then generate a network with those parameter values and then check to see if the direct estimators recover these parameters. If there is too much incidental generation, then the parameters will not be recovered and then our fourth estimator is needed, as is our new central limit theorem.

In particular, Theorem 3 requires a level of sparsity that makes certain kinds of incidental generations rare. For denser graphs (which can still be sparse, but permitting nontrivial incidental generation) we work with a minimum distance estimator that matches the moments of the shares of the subgraphs—which we denote by $\breve{\beta}_n^{\text{MD}}$. In Proposition 3, we show the consistency and asymptotic normality of this minimum distance estimator. We focus on the links and triangles model since the calculations are idiosyncratic based on the specific SUGM the researcher wants to employ, but the logic extends. The proof of asymptotic normality in this case of potentially dense SUGMs requires using our new central limit theorem for correlated random variables, Theorem 4, which is the focus of Section 6.

Appendix Dprovides simulations verifying consistency, asymptotic normality, and convergence. We also show how $\breve{\beta}_n^{\text{DC}}$ and $\breve{\beta}_n^{\text{MD}}$ both perform well when incidental generation is sufficiently small but that as the networks become denser $\breve{\beta}_n^{\text{DC}}$ is biased while $\breve{\beta}_n^{\text{MD}}$ is consistent.

We let β possibly depend on n and/or R as described below. We take the list of the types of subgraphs to be analyzed to be fixed.

4.2. **The Many Networks Case.** We keep the presentation of this first frame brief since it follows standard statistical arguments (e.g., Newey and McFadden (1994)).

One has a collection of R networks, each drawn independently according to the same SUGM with the same parameter vector β_0. We hold the set of nodes $\{1,\ldots,n\}$ (and their covariates) fixed. Theorem 2 states that a maximum likelihood estimator of the parameters is consistent and asymptotically normally distributed.

THEOREM **2.** *Consider a SUGM of k distinct types of subgraphs with $\beta_0 \in int(\mathcal{B})$, for $\mathcal{B}$ a compact subset of $[0,1)^k$. Let g_r for $r = 1,\ldots,R$ denote i.i.d. draws from this distribution. Let $\hat{\beta}_R^{\mathrm{ML}}$ denote the maximum likelihood estimator $\hat{\beta}_R^{\mathrm{ML}} = \mathrm{argmax}_{\beta\in\mathcal{B}} \frac{1}{R}\sum_r \log \mathrm{P}_\beta(g_r)$. Then $\hat{\beta}_R^{\mathrm{ML}} \xrightarrow{\mathrm{P}} \beta_0$. If in addition $J := \mathrm{E}[\nabla_\beta \log \mathrm{P}_{\beta_0}(g_r) \nabla_\beta \log \mathrm{P}_{\beta_0}(g_r)']$ is non-singular, then $\sqrt{R}\left(\hat{\beta}_R^{\mathrm{ML}} - \beta_0\right) \rightsquigarrow \mathcal{N}\left(0, J^{-1}\right)$.*

Although Theorem 2 demonstrates that a consistent and asymptotically normally distributed estimator exists, calculating the likelihood function of arbitrary networks as a function of the parameters can be computationally intensive for large networks. Thus, we also present a result on a minimum distance estimator which is computationally straightforward since it simply involves calculating frequencies of certain subgraphs. We present it based on links and triangles as the typical case that researchers will need, but the technique extends as a researcher requires. As before, let $S_L(g)$ and $S_T(g)$ denote the fraction of links and triangles in the network g, with $S = (S_L, S_T)'$.

Let

$$h(g_r, \beta) = S(g_r) - \mathrm{E}_\beta\left[S\left(g_r\right)\right],$$

be a moment function comparing observed subgraph statistics to expected ones for a given β.

Let $\hat{\beta}_R^{\mathrm{MD}}$ denote the minimum distance estimator,

$$\hat{\beta}_R^{\mathrm{MD}} := \underset{\beta\in\mathcal{B}}{\mathrm{argmin}} \left(\frac{1}{R}\sum_r h(g_r,\beta)\right)' \left(\frac{1}{R}\sum_r h(g_r,\beta)\right).$$

PROPOSITION **2.** *Consider a SUGM of links and triangles with parameters $\beta_0 \in int(\mathcal{B})$, a compact subset of $[0,1)^2$. Let g_r for $r = 1,\ldots,R$ denote i.i.d. draws from this distribution. Then,*

$$\hat{\beta}_R^{\mathrm{MD}} \xrightarrow{\mathrm{P}} \beta_0 \text{ and } \sqrt{R}\left(\hat{\beta}_R^{\mathrm{MD}} - \beta_0\right) \rightsquigarrow \mathcal{N}\left(0, (H'\Omega^{-1}H)^{-1}\right)$$

where $H := \mathrm{E}[\nabla_\beta h(g_r,\beta_0)]$ and $\Omega = \mathrm{E}\left[h(g_r,\beta_0)h(g_r,\beta_0)'\right]$.

4.3. **The Large Network Case.** Next we turn to the case where researchers have access to at least one large network (modeled as $n \to \infty$). For the exposition, we let $R = 1$, but clearly this extends directly to having observations of more than one network.

This case is considerably more challenging as it involves correlated observations generated within a network. Network data tend to be sparse, but still have local patterns such as clustering, so that people have relatively few connections compared to the potential number

of links, but where a person's neighbors tend to be linked to each other with much higher than an independent probability (e.g., see the background in Newman (2003); Jackson (2008)). Such clustering is the challenging aspect of the asymptotics since subgraphs are not only directly generated but also incidentally generated. Thus, we need new techniques for our asymptotic results.

4.3.1. *Sequences of Large Random Networks.* To describe how parameter estimators behave as a function of the number of nodes n, it is useful to allow the parameters to also be indexed by n. This approach is standard in the random graphs literature (e.g., see the classic book of Bollobas (2001)) as it is needed to accommodate most applications. Specifically, 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.[20] As a concrete example, consider friendships among high school students in the U.S. based on the Add Health data set (e.g., see (Currarini, Jackson, and Pin, 2009, 2010)). There are some high schools with only 30 students and others with around 3000 students. The average degree ranges between 6 and 8 over the high schools, but this means that the link probability must shrink dramatically with n: average degree d corresponds to a link probability of roughly $d/30$ in the small schools, but only $d/3000$ in the large schools. Thus, irrespective of the size of their school, students have numbers of friends of the same order of magnitude; and so the true underlying parameters describing friendship formation must decrease with n to match the data.

Thus, we consider a sequence of SUGMs with subgraphs $(G_1^n, ..., G_k^n)$ that form on n nodes that are generated with probabilities $\beta^n = (\beta_1^n, ..., \beta_k^n)$. The superscript on the β^n indicates the dependence on n to allow for true subgraph formation rates to vary along the sequence.

4.3.2. *Direct-Count Estimators for Negligible Incidental Generation.* It is convenient to express the β_ℓ^ns in the form

$$\beta_\ell^n = \frac{b_\ell}{n^{h_\ell}}$$

for some $b_\ell > 0$ and $h_\ell > 0$ fixed in n. This allows us to encode the rates at which the parameters vary with n, and is a general way of encoding the rates that could come from meeting, time budgets, costs, or any other constraints that gives rise to sparse networks.

We consider the case in which $m_\ell > h_\ell$ (where recall that m_ℓ is the number of nodes in the subgraph of type ℓ and is fixed along the sequence), as otherwise the expected number of subgraphs in the whole network could be bounded as n grows, precluding estimation.

The researcher can make assumptions on h_ℓ, either its value or the possible range of values that are admissible for their model. In fact, the magnitude may be straightforward to observe with simple subgraph counts. For example, if across networks of varying size, one sees some growing function of links, triangles, and so on, one can infer what values of h_ℓ are needed

[20]See Chandrasekhar (2016) 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.

to be consistent with data. In fact, in most models of network formation, such assumptions are implicitly made, knowingly or not.

We show that even without knowing the b_ℓ or h_ℓ, the parameters β_ℓ^n can be well-estimated, provided the network model is not so sparse that subgraphs are never observed, nor so dense so that they scale linearly in n.

To develop the estimator, first we need some definitions and notation.

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

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

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.

We count subgraphs in this order after having removed links associated with all of the subgraphs already counted. The resulting counts are denoted $\widetilde{S}_\ell^n$:[21]

$$\widetilde{S}_\ell^n(g) = |\{g_\ell \in G_\ell^n : g_\ell \subset g \text{ and } g_\ell \not\cap g_{\ell''} \text{ for any } g_{\ell''} \in G_{\ell''}^n \text{ such that } g_{\ell''} \subset g \text{ for some } \ell'' < \ell\}|.$$

The logic of this is that incidental generation is more often in one direction than another: a triangle incidentally generates three links, while it can be much rarer that three links happen to independently form to make a triangle. This manner of counting motivates a simple estimator that we call the *direct-count* estimator. We then divide by the number of possible subgraphs of that variety.

For the direct count estimator $\breve{\beta}_n^{\mathrm{DC}}$ we presume that, for each ℓ, G_ℓ^n includes all subgraphs that are relabellings of each other. Thus, we work without demographics on the subgraphs, but these counts can easily be adjusted accordingly by normalizing by $|G_\ell^n|$. Let κ_ℓ^n denote the (finite number) of relabelings to count different subgraphs in G_ℓ^n on a given set of m_ℓ nodes.[22] Then $\kappa_\ell^n \binom{n}{m_\ell}$ is the number of possible subgraphs of type ℓ.

The direct-count estimator $\breve{\beta}_n^{\mathrm{DC}}$ is

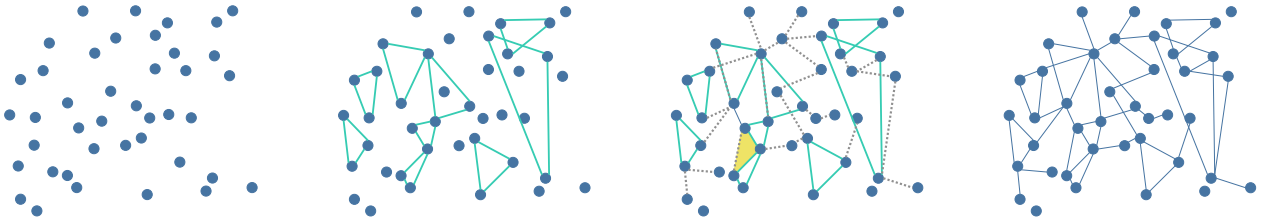
$$\breve{\beta}_{n,\ell}^{\mathrm{DC}} = \frac{\widetilde{S}_\ell^n(g)}{\kappa_\ell \binom{n}{m_\ell}}. \tag{4.1}$$

As we prove in Section 4.3.2, under suitable conditions, incidental generation is low and the direct estimators are consistent estimators of the true parameters and are asymptotically normally distributed.

[21]Note that counting in order from 'largest' to 'smallest' subnetworks means that we count things from smallest to largest index ℓ since the specification of how we ordered labels moves in the opposite direction of the size of the subgraphs.

[22]For example, note that $\kappa_\ell^n = 1$ 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.

As an illustration, consider Figure 5 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, $\widetilde{S}_T^n(g)$, 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 $\widetilde{S}_L^n(g)$ becomes 22 instead. So we estimate 22/820 while the true frequency was 25/820. These errors are already small on a network on just 41 nodes, and as we prove below, the errors disappear completely as n grows.



(A) n nodes (B) Triangles form (C) Links form (D) Resulting network

FIGURE 5. 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 $\widetilde{S}_T^n(g) = 10$, and $\widetilde{S}_L^n(g) = 22$ from (D), while the true process generated 9 triangles and 23 links directly. The estimated parameters are $\widehat{\beta}_{n,T}^{DC} = \frac{10}{10660}$, and $\widehat{\beta}_{n,L}^{DC} = \frac{22}{820}$, while the true frequencies were $\frac{9}{10660}$ and $\frac{25}{820}$.

To understand when the direct-count estimator is appropriate, we need to characterize the rate of incidental subgraph formation. To do this we track how many ways a subnetwork $g' \in G_\ell^n$ could be incidentally generated.

We first provide a precise specification of what it means to be incidentally generated. We say that a subgraph $g' \in G_\ell^n$ for some ℓ can be *incidentally generated* by the subgraphs $\{g^j\}_{j\in J}$, indexed by J, if $g' \subset \cup_{j\in J} g^j$. For instance a triangle $g' = 123$ can be incidentally generated by links $g^1 = 12$, $g^2 = 23$, and triangle $g^3 = 134$; or by link 12 and triangles 234 and 135, etc. Some of these incidental generations are equivalent to each other (e.g., involve two links and one triangle) and so it is useful to define equivalence classes of generators.

Consider any potential subgraph $g' \in G_\ell^n$ that can be incidentally generated by a set of subnetworks $\{g^j\}_{j\in J}$ with associated indices ℓ_j and also by another set $\{g^{j'}\}_{j'\in J'}$. We say that $\{g^j\}_{j\in J}$ and $\{g^{j'}\}_{j'\in J'}$ are *equivalent generators* of g' if there exists a bijection π 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. For instance a triangle 123 is not only incidentally generated by links 12, 23, and triangle 134; but also by an equivalent generator of links 12, 23, and triangle 135, or links 23, 13; and triangle 128, and so forth.

Given this equivalence relation, we simplify by ignoring the specific labels of subgraphs and defining *generating classes* for any type of subgraph G_ℓ. We just track the number and type of subgraphs needed, as well as how many nodes each subgraph has intersecting with the given incidentally generated subgraph.

In particular, each generating class $\mathcal{C}$ of some G^n_ℓ 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 $c_j = |g^j \cap g'|$. For example, if we consider a triangle, then it can be incidentally generated by two other triangles and a link; and we represent that 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.

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.[23]

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

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

The first 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. We do it this way, since we first count the largest subgraphs, and having accounted for them, we worry about the remaining subgraphs, and so forth. The second condition restricts attention to the smallest generators. For instance a triangle could be generated by two links and two triangles. However, in that case either one of the links or one of the triangles can be dropped. The minimal classes for the triangle only involve three subgraphs: three links, two links and one triangle, one link and two triangles, or three triangles. Under the first condition, 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.

The following conditions ensure that the direct estimation parameters are arbitrarily accurate for large enough networks.

First, for each ℓ let

$$h_\ell > m_\ell - 2. \tag{4.2}$$

This condition ensures that the overall degree of any node grows more slowly than the size of the graph. This comes from the fact that any given node can be a part of $\binom{n}{m_\ell - 1}$ subgraphs of type ℓ, each of which forms with probability $\frac{b_\ell}{n^{h_\ell}}$. Expecting over these gives an upper

[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.

bound on the number of links (up to a proportional constant) that a given node is part of from graphs of type ℓ, and the condition is that this be smaller than n. The average degree can still grow with n, but sublinearly. In particular, this condition ensures that the chance that any given link is part of multiple subgraphs is vanishing.

Next, for each ℓ consider any (minimal)[24] generating class with index J of subgraphs no larger than ℓ. Let

$$h_\ell < \sum_{j \in J} h_{\ell_j} + c_{\ell_j} - m_{\ell_j} \tag{4.3}$$

and

$$h_{\ell_{j'}} + m_\ell - m_{\ell_{j'}} < \sum_{j \in J} h_{\ell_j} + c_{\ell_j} - m_{\ell_j} \tag{4.4}$$

for each $j' \in J$.

(4.3) is the requirement that a given subgraph is more likely to form directly than indirectly. h_ℓ governs the direct formation, and $\sum_{j \in J} h_{\ell_j} + c_{\ell_j} - m_{\ell_j}$ governs the rate of incidental generation, and so the exponent on the direction formation must be less than the sum of the exponents of the graphs needed for incidental generation, subtracting off how many variations on each of these there are (captured by the $-(m_{\ell_j} - c_{\ell_j})$ coming from how many nodes are free to be chosen for each incidentally generating subgraph). (4.4) is the requirement that a given subgraph of some type $\ell_{j'}$ that is part of a generating class of some ℓ appear at a fast enough rate to ensure that it is not always becoming part of incidentally generated ℓs, but can be distinguished. This is a similar calculation of rates.

Under these conditions, we prove identification in addition to consistency and asymptotic normality on a single large network.

Define the variance-covariance matrix

$$V_n = \operatorname{diag}\left\{ n^{2h_\ell} \frac{\beta^n_{0,\ell}(1 - \beta^n_{0,\ell})}{\kappa_\ell \binom{n}{m_\ell}} \right\}.$$

We say that a sequence of SUGMs with k types of subgraphs is complete and growing if for each $\ell \in \{1, \ldots, k\}$, G^n_ℓ includes all subgraphs that are relabellings of each other and $G^n_\ell \subset G^{n+1}_\ell$. So, this implies that ℓ has the same meaning (e.g., triangles or $k-stars$) across n.

THEOREM **3.** *Consider a growing and complete sequence of SUGMs of k distinct types of subgraphs. If they have associated true parameters $0 < b_{0,\ell}$ such that $\beta^n_{0,\ell} = \frac{b_{0,\ell}}{n^{h_\ell}}$ and (4.2)-(4.4) hold for each ℓ and associated (minimal) generating classes, then $|\breve{b}^{\mathrm{DC}}_n - b_0| \xrightarrow{\mathrm{P}} 0$ and $V_n^{-1/2}\left(\breve{b}^{\mathrm{DC}}_n - b_0\right) \rightsquigarrow \mathcal{N}(0, I)$.*

Although the conditions may appear hard to understand, they are actually fairly straightforward, and it is easy to see sufficient conditions that ensure them.

[24] If the condition is satisfied by minimal classes, it is automatically satisfied by larger classes.

For example, suppose that each $h_\ell = m_\ell - x$ for some same $x \in (0,2)$, so that each node has the same order probability of being a part of different sorts of subgraphs. This is the natural case, as otherwise some subgraphs become infinitely more likely than others.

In that case, all three conditions are automatically satisfied whenever the subgraphs are all cyclic (cliques, or other subgraphs in which all nodes are parts of cycles). If some of the subgraphs are not cyclic (e.g., lines or stars), then all three conditions hold if $x \in (0,1)$.

COROLLARY **1.** *Consider a growing and complete sequence of SUGMs of k. If they have associated true parameters $0 < b_{0,\ell}, n^{h_\ell}$ such that $\beta^n_{0,\ell} = \frac{b_{0,\ell}}{n^{h_\ell}}$, and such that $m_\ell - h_\ell = x$ for each ℓ and some $x \in (0,2)$ and either all subgraphs are cyclic or else $x < 1$, then $|\breve{b}^{\mathrm{DC}}_n - b_0| \xrightarrow{\mathrm{P}} 0$ and $V_n^{-1/2}\left(\breve{b}^{\mathrm{DC}}_n - b_0\right) \rightsquigarrow \mathcal{N}(0, I)$.*

In both results, although we state them in terms of bs, it is also true that the ratio of $\breve{\beta}^{\mathrm{DC}}_{n,\ell}$ to $\beta_{0.\ell}$, tends to one. Furthermore, as we show in the proof, if we normalize the difference between the estimated probability and the truth by the standard deviation, this is asymptotically normally distributed. This is an equivalent representation of the above result, but is helpful to note as it does not require knowledge of h_ℓs but rather just that they satisfy the relevant bounds, which is true of many human networks.

4.3.3. *Minimum Distance Estimator for Non-Negligible Incidental Generation.* Theorem 3 holds for parameter values for which incidental generation becomes small as a function of the overall counts of the subgraphs, and works for arbitrary subgraph varieties. However, we may want an estimator that works when incidental generation is not ignorable, even in the limit.

For SUGMs with specific subgraph types, we can explicitly calculate all the incidental rates and account for them, and develop an estimator that is more accurate in small samples where there can be nontrivial incidental generation and also works asymptotically even when there is incidental generation. In particular, in this section we consider a links and triangles SUGM based and provide an estimator that fully accounts for incidental generation (with extensive details in Appendix C.We prove identification as well as consistency and asymptotic normality of a minimum distance estimator.

In order to show the properties of the minimum distance estimator, we show that the following moments converge:

$$\frac{S^n_L(g) - \mathrm{E}_{\beta^n_0}[S_L(g)]}{\sigma^n_L} \rightsquigarrow \mathcal{N}(0,1), \text{ and } \frac{S^n_T(g) - \mathrm{E}_{\beta^n_0}[S_T(g)]}{\sigma^n_T} \rightsquigarrow \mathcal{N}(0,1),$$

and jointly as well, where $(\sigma^n_L)^2 := \mathrm{var}\left(S^n_L(g)\right)$ and $(\sigma^n_T)^2 := \mathrm{var}\left(S^n_T(g)\right)$. Since

$$S^n_L(g) = \frac{\sum_{i<j} g_{ij}}{\binom{n}{2}} \text{ and } S^n_T(g) = \frac{\sum_{i<j<k} g_{ij} g_{ik} g_{jk}}{\binom{n}{3}}$$

and g^n_{ij} and g^n_{ik} are correlated for any k, S^n_L involves correlated random variables, and since any two triples in S^n_T that involve a common link are correlated, we need to prove a central limit theorem that shows that such correlation does not cause problems.

Let $S^n(g) = (S_L^n(g), S_T^n(g))'$ be the stacked vector of both shares. It is useful to define the variance-covariance matrix of the moments

$$V_n = \begin{pmatrix} \text{var}(n^{h_L} S_L) & \text{cov}(n^{h_L} S_L, n^{h_T} S_T) \\ \text{cov}(n^{h_L} S_L, n^{h_T} S_T) & \text{var}(n^{h_T} S_T) \end{pmatrix}.$$

Finally, let $R_n = \text{diag}\left\{n^{h_L}, n^{h_T}\right\}$. With this defined we can state our result.

Define the minimum distance estimator for a single large network by

$$\breve{\beta}_n^{\text{MD}} := \underset{\beta}{\text{argmin}} (S^n(g) - \text{E}_{\beta_0^n}[S^n(g)])' R_n^2 (S^n(g) - \text{E}_{\beta}[S^n(g)]).$$

PROPOSITION **3.** *Consider a links and triangles SUGM with associated parameters* $\beta_{0,L}^n, \beta_{0,T}^n = \left(\frac{b_{0,L}}{n^{h_L}}, \frac{b_{0,T}}{n^{h_T}}\right)$ *with* $0 \leq \underline{D} < b_{0,L}, b_{0,T} < \overline{D}$ *such that* $h_L \in (2/3, 2)$ *and* $h_T \in [h_L{+}1, 3h_L]$, *with* $h_T < 3$. *Then the minimum distance estimator is consistent,* $|\breve{b}_n^{\text{MD}} - b_0| \xrightarrow{\text{P}} 0$, *and* [25] *and asymptotically normal,* $V_n^{-1/2}\left(\breve{b}_n^{\text{MD}} - b_0\right) \rightsquigarrow \mathcal{N}(0, I)$.

The proof makes use of Theorem 4, below. The proof is in Appendix C.

Proposition 3 covers a wide range of link and triangle densities, ranging from average degree on the order $n^{1/3-\delta}$ to $n^{-1+\delta}$ for any $\delta > 0$. This covers the order constant and logarithmic growth rates of average degree studied in the literature (Newman et al., 2001; Bollobas, 2001; Jackson, 2008; Graham, 2017), for instance.

In particular, Proposition 3 covers situations in which the rate of incidental generation (e.g., the proportion of triangles that are generated incidentally) does not vanish asymptotically. Not only does the estimator have better small sample properties (see the simulations below), but it also works asymptotically in cases that Theorem 3 does not.

The restrictions are easily interpretable. $h_T \geq h_L + 1$ ensures that triangles are not so numerous that almost all of the links in the network lie in triangles: that n^{3-h_T} does not dwarf n^{2-h_T}. $h_T \leq 3h_L$ ensures the opposite: that triangles are not always incidentally formed by links and never formed directly: $n^{3(1-h_T)}$ is not dwarfed by n^{3-3h_L}. $h_L > 2/3$ ensures that links and triangles are disentangled by imposing a density cap. Finally, $h_L < 2, h_T < 3$ ensure that there is information in the network—enough links and triangles are present to estimate their formation.

Again we note that although the results are stated in terms of b, these are equivalent statements to saying that ratio of the estimated ($\breve{\beta}_n^{\text{MD}}$) and true ($\beta_0^n$) frequencies tend to one. And, that, when self-normalized by the standard deviations, the empirical frequencies estimated are asymptotically normally distributed. Thus, the result requires no knowledge of h_ℓs other than that they satisfy the relevant bounds.

4.3.4. *Discussion of incidental generation and estimators.* It is instructive to summarize the difference in assumptions and performance of $\breve{\beta}_n^{\text{DC}}$ and $\breve{\beta}_n^{\text{MD}}$. Again, the first requires more sparsity—less incidental generation specifically—than the latter. Relative to Theorem 3, we can see that Proposition 3 covers cases where incidental generation is not ignorable. Namely, one can check that our result on the $\breve{\beta}_n^{\text{DC}}$ requires $h_T > 2$ (which means that the probability of a triangle is going to faster at a rate faster than $1/n^2$). But $\breve{\beta}_n^{\text{MD}}$ only requires a rate

[25] The expression for V_n is different when $h_T = h_L + 1$, and is given in the proof of the proposition.

faster than $h_T > 5/3$ or $1/n^{1+2/3}$. This means that triangles (and therefore links, checking the conditions) can appear at a faster rate and still be estimated under the minimum distance estimator but not through direct-counts. We also see evidence of this in our simulations, in Appendix D, where for very sparse networks both estimators give the same result but the direct-count becomes biased as we increase density whereas the minimum distance estimator remains unbiased.

5. Applications

SUGMs are useful for a number of purposes. First, purely as a statistical modeling tool, SUGMs—even ones with just links and triangles—generate higher-order features of empirically observed social networks that link-based models (even those accounting for characteristics, unobserved characteristics, geography, and latent locations) cannot. This is important for prediction. For example, if one wants to see which networks might form under a hypothetical policy, a model is only useful if it can generate networks that are likely to occur at a variety parameter values. As we demonstrate, our model outperforms stochastic block models, models with node-level fixed effects, latent space models, and ERGMs in generating realistic distributions of networks even with considerably fewer parameters (e.g., 4 parameter SUGMs versus over 200 (or even 400) parameters in some alternatives).

Second, a SUGM can be used to test which incentives underlie link formation. There are many theories (e.g., Coleman (1988); Jackson, Rodriguez-Barraquer, and Tan (2012)) predicting that triangles and other cliques play special roles in maintaining cooperation in favor exchange. In order to test such theories, we need a statistical model that allows us to test whether cliques appear significantly more often than being randomly generated by links, and whether they appear in configurations that would be predicted by the game theory.

Third, SUGMs can be used for structural estimation. There are parsimonious microfoundations—models of mutual consent or search—that give rise to SUGMs. Structural parameters are useful for welfare analyses, and also aid in examining counterfactuals or policy evaluation. Such parameters are recoverable from SUGM parameters.

We provide three examples. Our first example shows that SUGMs model a myriad of network features much better than other standard models. The other two examples build models of network formation to address specific economic questions. In both cases, the equilibrium network is a random draw from a SUGM with interpretable parameters.

5.1. **Data.** We use the Banerjee, Chandrasekhar, Duflo, and Jackson (2013, 2019) data (https://doi.org/10.7910/DVN/U3BIHX) consisting of a variety of social and economic networks from 75 Indian villages as well as detailed demographic background.[26] Having 75 villages allows us to show not only how the model scales with the number of nodes, but also to cover both of our asymptotic frames.

The networks have households as nodes. There are an average of $n = 220$ households per village. We surveyed adults, asking them about a variety of their daily interactions, as well as their demographics (caste, education, profession, religion, family size, wealth

[26]See Banerjee, Chandrasekhar, Duflo, and Jackson (2013) for more information about the data.

variables, voting and ration cards, self-help group participation, savings behavior, etc.). We have network data from 89.14 percent of the 16,476 households based on interviews with 65 percent of all adults between the ages of 18 and 55. We have data concerning twelve types of interactions: (1) whose houses he or she visits, (2) who visits his or her house, (3) his or her relatives in the village, (4) non-relatives who socialize with him or her, (5) who gives him or her medical help, (6) from whom he or she borrows money, (7) to whom he or she lends money, (8) from whom he or she borrows material goods (e.g., kerosene, rice), (9) to whom he or she lends material goods, (10) from whom he or she gets important advice, (11) to whom he or she gives advice, (12) with whom he or she goes to pray (e.g., at a temple, church or mosque).

The answers are aggregated to the household level, but one can also work with the individual-level networks to get similar results to those presented below. How a link is defined varies based on the application. We use undirected,[27] unweighted networks that may allow for multiplexing. This also means that we observe 98.8% of the potential links between pairs.[28]

For much of what follows, we work with the borrowing and lending of material goods (questions 8 and 9, with any positive answer indicating a link being present) that we call "favor" links, and the exchange of advice (questions 10 and 11, with any positive answer indicating a link being present) that we call "info" links.

5.2. **Example 1: Matching Features of Empirical Network Data.** A challenge for network formation models has been to capture more than one or two observed features of networks at a time. For instance, many observed social networks are sparse but clustered, which motivates developing models that reflect this (Watts and Strogatz, 1998). They also have a variety of differing degree distributions ((Barabasi and Albert, 1999; Jackson and Rogers, 2007) and exhibit high levels of homophily (McPherson, Smith-Lovin, and Cook, 2001; Currarini, Jackson, and Pin, 2009, 2010), which can lead to poverty traps and inequalities (Calvo-Armengol and Jackson, 2007; Jackson, 2023). There are also features such as the expansion properties of a network that are described by maximal eigenvalue of the adjacency matrix and governs diffusion processes operating on the network (Bollobas (2001)). The depth of the max flow min cut speaks to several things such as consensus time in a social learning process Golub and Jackson (2012) as well as the sustainable degree of cooperation (Karlan, Mobius, Rosenblat, and Szeidl, 2009).

We show that a SUGM fits economically-relevant network features in the data far better than four prominent alternatives. Importantly, these features were not used to fit the model. They are the size of the giant component, average path length, and various spectral properties of the adjacency matrix (e.g., the largest eigenvalue and an eigenvalue measure of homophily).

[27]Some links are not reciprocated, but that is true at similar rates for the questions regarding relatives as compared to the other questions, and so much of the failure of reciprocation may simply be measurement error rather than true one-way relationships. For our purposes here, which are purely to illustrate the ability of the models to match data, this distinction is inconsequential.

[28]This is a new wave of data relative to our original microfinance study that includes more surveys. Note that $1-(1-0.8914)^2 = 0.988$.

A simple SUGM outperforms the alternative models despite the fact that the alternative models have many more dimensions such as numerous covariates, n fixed effects, or even n latent space variables, that should give them an advantage in fitting.

Specifically, the alternative models are (a) a standard stochastic block model that includes flexible controls for continuous covariates that influence edge probabilities; (b) an extension of that model that includes n parameters to capture node fixed effects (e.g., Graham (2017)); (c) a latent space model (Hoff, Raftery, and Handcock, 2002) in which nodes have unobserved arbitrary locations in $\mathbb{R}^3$ to be estimated and the probability of linking declines in their latent positions; and (d) an exponential random graph model with links, triangles, and a rich set of covariates.

Before we proceed, we review the features of the graph structure that we examine and why they are interesting. 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)). 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.[29] 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)), but is also closely related to homophily (Golub and Jackson (2012)) and is labeled as such in the table below. Additionally, we look at the fraction of nodes that belong to the giant component of the network, as well as the number of isolates, as empirical networks are often not completely connected. Finally, we also consider average path length (in the largest component).

We present the results for favor and info networks. These networks are reasonably connected (with more than ninety percent of the nodes being in the giant component) and yet also typically sparse.

Our procedure is as follows. For every village, we estimate six network formation models.

One network formation model is a link-based model (stochastic block model) in which the probabilities can depend on geographic distance, caste, the number of rooms households have, number of beds, quality of electricity provision, quality of latrines, household ownership status, and squared differences in non-binary variables. The probabilities are estimated using logistic regression and the model has 12 parameters.

The next is the model of Graham (2017). This is the same formulation of the preceding model, but adds unobserved heterogeneity in the form of node-fixed effects,

$$\mathrm{P}(g_{ij} = 1|X_{ij}) = \Lambda\left(\alpha_i + \alpha_j + \gamma' X_{ij}\right),$$

where $\Lambda(\cdot)$ is the logit link function and X_{ij} is the aforementioned vector of demographic characteristics and polynomials therein. This model has n+12 parameters per network.[30]

[29]The stochasticized adjacency matrix T is defined as $T_{ij} = \frac{g_{ij}}{\sum_k g_{ik}}$, 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.

[30]Consistency of all α_i in addition to β has been proven for a dense sequence of graphs (e.g., Chatterjee et al. (2010); Graham (2017)).

The third model is a latent space model,

$$\mathrm{P}(g_{ij} = 1|w_{ij}) = \Lambda\left(\alpha_i + \alpha_j - \eta \cdot \mathrm{dist}(w_i, w_j) + \gamma' X_{ij}\right),$$

where now w_i are unobserved positions in $\mathbb{R}^3$.[31] This has $2n + 12$ parameters.

The fourth model is a links and triangles ERGM with covariates. Specifically,

$$\mathrm{P}(g) \propto \exp(\theta_L \cdot S_L(g) + \theta_T \cdot S_T(g) + \gamma' X).$$

Turning to SUGMs, in contrast, we consider only low-dimensional models. One is a the basic SUGM with links and triangles. Pairs of household are categorized as either being "close" or "far," where "close" refers to pairs of nodes that are of the same caste and "far" to those that differ in caste. Similarly, we categorize triangles as being "close" if all nodes are of the same caste and "far" otherwise. Thus, we allow for four parameters, close and far link parameters and close and far triangle parameters. The other model is a slightly richer SUGM in which we allow some nodes to be isolates, which adds one more parameter.[32] Neither includes any other demographic covariates nor unobserved heterogeneity. We estimate both via a variation on the minimum distance estimator of Proposition 3, $\breve{\beta}_n^{\mathrm{MD}}$, since there appears to be enough incidental generation that needs to be accounted for.[33]

To make the strongest point, we compare these stark SUGMs that use only same/different caste variables to account for homophily, to very rich covariate dependent (block) models that can incorporate a large set of covariates – including much richer demographics that are usually available to a researcher as well as node-level fixed effects in the unobserved heterogeneity model and node-level latent locations in the latent space model. We show that even though we have considerably more information on the nodes, such as geographic distance and demographic characteristics, and allow for such unobserved heterogeneities—and we do not make use of this information for the SUGMs—they recreate networks much more accurately than a link-based model that does takes advantage of a rich set of node characteristics. Adding over 12 parameters to the block model to flexibly control for demographic attributes, *or even n+12 parameters with unobserved heterogeneity or* $2n + 12$ *with latent locations*, does not come close to doing as well as the simple SUGMs. Moreover, since the specification developed here makes use of considerably richer data than those used in the two candidate

[31]We use $\mathbb{R}^3$ as Euclidean is commonly used in the literature, though it is not the only choice. The subject of choice of geometry is addressed in Lubold, Chandrasekhar, and McCormick (2023), which shows how to check isometric embedding conditions. In this data we find that 25% of networks are not consistent with *any* latent space from the family of simply connected, complete Riemannian manifolds of constant curvature, lending evidence to the idea that latent space models may not be a universally appropriate device to model correlation.

[32]With isolates, in a first stage some nodes are randomly chosen to be isolates with a given probability. For the subsequent formation of other subgraphs, those isolates can be considered as removed from the set of nodes and no subgraph that involves them forms in the subsequent subgraph formation.

[33]Specifically, we use a hybrid estimator of first directly (unbiasedly) estimating the link parameters from the frequency of links among pairs of nodes that have no common neighbors. We then fix this parameter in the minimum distance estimator to estimate the triangle parameters. This slightly simplifies the computations. The code appears in supplementary materials.

SUGM models, it suggests that by decomposing a network into a tapestry of random structures (triangles, links, and even isolates), considerable value is added in modeling higher order features of networks in a parsimonious way.

We estimate parameters village-by-village for each model and then generate random network from each model based on the estimated parameters. We do 100 such simulations for each village and model. We then compare the true network characteristics with those from the simulations for each of the various models.

TABLE 1. Network Properties

	Truth	Links/ Triangles SUGM	Links/ Triangles/ Isolates SUGM	Covariates (Block Model)	Covariates + Unobserved Heterogeneity (Latent Block Model)	Latent Space Model (with Covariates)	ERGM (Links/Triangles with Covariates)
Panel A: Information							
Degree	8.0960	8.2166	8.2064	8.8111	9.5860	13.2540	13.8812
	(0.2607)	(0.2754)	(0.2746)	(0.3126)	(0.3615)	(0.1257)	(0.1337)
Clustering	0.2198	0.1599	0.1478	0.0506	0.0742	0.0834	0.1287
	(0.0057)	(0.0034)	(0.0032)	(0.0030)	(0.0045)	(0.0007)	(0.0010)
Isolates	10.9718	3.3361	13.4597	0.5313	0.8369	10.7658	12.8454
	(0.8410)	(0.3853)	(0.9924)	(0.0977)	(0.1336)	(0.1448)	(0.1381)
% in Giant	0.9503	0.9844	0.9397	0.9977	0.9964	0.9434	0.9205
	(0.0030)	(0.0016)	(0.0033)	(0.0004)	(0.0005)	(0.0008)	(0.0013)
Maximal Eigenvalue	11.9138	10.6260	11.0178	10.3737	12.5418	16.2470	18.3364
	(0.3741)	(0.3293)	(0.3466)	(0.3222)	(0.4325)	(0.1293)	(0.1323)
Homophily	0.8865	0.8156	0.8029	0.6869	0.6795	0.8743	0.7921
	(0.0065)	(0.0090)	(0.0093)	(0.0104)	(0.0097)	(0.0009)	(0.0024)
Average Path Length	3.0273	2.9406	2.8580	2.7602	2.6399	3.1017	3.1163
	(0.0485)	(0.0422)	(0.0393)	(0.0403)	(0.0374)	(0.0106)	(0.0172)
Panel B: Favor							
Degree	7.0579	7.2192	7.2185	7.7614	8.5145	13.1301	16.6176
	(0.2611)	(0.3048)	(0.3052)	(0.3232)	(0.3835)	(0.1571)	(0.1347)
Clustering	0.2895	0.1894	0.1764	0.0467	0.0641	0.0724	0.1506
	(0.0054)	(0.0034)	(0.0032)	(0.0032)	(0.0040)	(0.0008)	(0.0008)
Isolates	10.0704	7.2859	16.0939	1.0423	3.4777	19.3338	15.8796
	(0.7670)	(0.6645)	(1.1103)	(0.1429)	(1.7734)	(0.2763)	(0.2046)
% in Giant	0.9510	0.9632	0.9245	0.9955	0.9820	0.8695	0.9108
	(0.0032)	(0.0031)	(0.0041)	(0.0005)	(0.0107)	(0.0021)	(0.0012)
Maximal Eigenvalue	10.0654	9.8417	10.1727	9.4778	11.3075	16.0454	21.5754
	(0.3337)	(0.3702)	(0.3826)	(0.3382)	(0.4236)	(0.1637)	(0.1201)
Homophily	0.9412	0.8716	0.8636	0.7325	0.7189	0.9074	0.7895
	(0.0044)	(0.0082)	(0.0085)	(0.0107)	(0.0111)	(0.0010)	(0.0025)
Average Path Length	3.5158	3.1591	3.0739	2.9140	2.7799	3.8148	2.8126
	(0.0659)	(0.0479)	(0.0441)	(0.0442)	(0.0443)	(0.0212)	(0.0149)

Table 1 presents the results, averaged across villages for each of the models. We use 71 villages out of the 75 since 4 villages have only one caste group. The ERGM is only estimated for 68 villages as it did not converge for 3 villages. Both of the SUGMs match the features of the networks substantially better than the conditional edge independent models (with and without node fixed effects). Including isolates in the SUGM further improves the fits not only for isolates, but also for fraction in the giant component and the maximum eigenvalue. This suggests that there are more isolated households in a village for a reason beyond randomness in network formation.

An obvious thing to note is that the link-based and also latent space models do extremely poorly when it comes to matching clustering while the SUGM does much better, and here adding unobserved dimensions to generate unconditional link correlations (e.g., clustering) does worse than a SUGM that allows correlated link formation directly. The ERGM performs better on clustering but a the cost of generating excessive density, diffusiveness, the spectral cut (homophily), connectedness, and average path length.

Including triangles in the SUGM is enough to deliver better matches on all dimensions, and the difference on homophily is perhaps most interesting, since one would imagine that the block models or even latent space models could get that right given that they include many covariates. This tells us that triangles and correlation between links play a subtle but important role in homophily—something that is better picked up by a SUGM than an independent link model even when that model includes rich demographics and unobserved heterogeneity.

It is important that SUGMs do a much better job at recreating a multitude of features of observed network structures that standard link-based models, especially with rich demographic information, models with unobserved heterogeneity, latent space models, and ERGMs. It suggests that there is a substantial value added of modeling the formation of triangles and isolates. 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 some other empirical applications. For example, when we look at links across social boundaries, we can be comfortable that to a first order, thinking about a SUGM with links and triangles across and within caste groups can do a good job of matching patterns in the data, and thus tracing them back to model parameters.

5.3. **Example 2: Do incentives for risk sharing drive network formation?**

5.3.1. *A model of mutual consent.* Consider a simple model in which individuals get utility from being in bilateral relationships, denoted by L, as well as trilateral relationships, denoted by T. The value of a partner j to i in a bilateral relationship is a function of their demographics (given by vector X_i) is given by u_i^L:

$$u^L\left(X_i; X_j\right) = X_i'\gamma_{L1} + X_j'\gamma_{L2} + \gamma_{L3} d_L(X_i, X_j) - \epsilon_{ij} =: \phi_L(X_i; X_j) - \epsilon_{ij}.$$

where $d_L(X_i, X_j)$ is a distance or other function comparing the demographics—for instance to allow for homophily. Similarly, the value of a triangle of relationships jk to i is given by u_i^T:

$$u^T\left(X_i; X_j, X_k\right) = X_i'\gamma_{T1} + f(X_j, X_k)'\gamma_{T2} + \gamma_{T3} d_T(X_i; X_j, X_k) - \varepsilon_{ijk} =: \phi_T(X_i; X_j, X_k) - \varepsilon_i jk,$$

where $f(\cdot,\cdot)$ is a function that is symmetric in arguments, and $d_T(X_i;\cdot,\cdot)$ is a function that is symmetric in the last two arguments. The value of the relationships depend on the characteristics of the people involved, as well as some idiosyncratic values to the relationships, $-\epsilon_{ij}$ and $-\varepsilon_{ijk}$, which may capture personalities, compatibilities, etc., distributed according to some distributions F_L and F_T respectively.

Forming relationships requires mutual consent (e.g., the pairwise stability of Jackson and Wolinsky (1996)), so the net utility must be positive to all agents. The probability that a

subgraph ij forms is

$$\beta_L\left(X_{ij},\gamma_L\right)=F_L\left(\phi_L(X_i;X_j)\right)\times F_L\left(\phi_L(X_j;X_i)\right)$$

and similarly the probability that subgraph ijk forms is

$$\beta_T\left(X_{ijk},\gamma_T\right)=F_T\left(\phi_T(X_i;X_j,X_k)\right)\times F_T\left(\phi_T(X_j;X_i,X_k)\right)\times F_T\left(\phi_T(X_k;X_i,X_j)\right).$$

The products capture that a link requires two consents and a triangle requires three.

By estimating the probabilities of subgraphs forming ($\beta_T(\cdot)$ and $\beta_L(\cdot)$), under suitable assumptions described below, one can recover the marginal effects of changes in covariates on preferences for being in various configurations (γ_T and γ_L). Since we have finite support for covariates, we label the subgraph formation probabilities β_{T,X_T} and β_{L,X_L} for pair and node covariate combination X_T and X_L respectively.

5.3.2. *Incentives for Risk-Sharing.* Jackson, Rodriguez-Barraquer, and Tan (2012) show that whether or not a link is supported can play an key role in maintaining informal favor exchange when it would not be self-sustaining without social pressure. It characterizes renegotiation-proof and robust pairwise stable networks and shows that (in the homogenous parameter case) all networks that incentivize exchange are quilts (a union of cliques with no cycle involving more than the minimal clique-size number of nodes), and in the inhomogenous parameter case every link must be supported (if i,j are linked then there exists k such that $g_{ik}=g_{jk}=1$).

Consider a variation on this model wherein now there are multiple link types: favors and information. We can use this to study the question raised by Jackson, Rodriguez-Barraquer, and Tan (2012). To make this simple ignore covariates, so all nodes are identical. Preferences are described by a random utility framework (McFadden, 1973), with the value of a link between i and j to i given by

$$u_i^{L,favor}(i)=\gamma_{L,favor}-\epsilon_{ij,favor},\ u_i^{L,info}(i)=\gamma_{L,info}-\epsilon_{ij,info},$$

and the value of a triangle given by

$$u_i^{T,favor}(jk)=\gamma_{T,favor}-\epsilon_{ijk,favor},\ u_i^{T,info}(jk)=\gamma_{T,info}-\epsilon_{ijk,info}.$$

In this case, due to mutual consent, $\beta_{L,favor}=F(\gamma_{L,favor})^2$ and $\beta_{T,favor}=F(\gamma_{T,favor})^3$. It is analogous for information. By the arguments of Jackson, Rodriguez-Barraquer, and Tan (2012), we can test the hypothesis that fraction of links that are supported is higher in favor exchange than in information links, which would be consistent with exchanging of favors needing to be incentivized while sharing of information not needing network incentives. In the language of this model the hypothesis is expressed in terms of parameters as follows:[34]

LEMMA **1.** *Under the above assumptions,*

$$\frac{\gamma_{T,favor}}{\gamma_{L,favor}}>\frac{\gamma_{T,info}}{\gamma_{L,info}}\ \textit{corresponds to}\ \frac{\beta_{T,favor}/\beta_{L,favor}^{3/2}}{\beta_{T,info}/\beta_{L,info}^{3/2}}>1.$$

[34]It is without loss of generality to take $F(\gamma)=\gamma$ which is just a bijection and is convenient to work with.

The (joint) hypothesis that we are testing is that exchanging material goods is more costly and/or happens less frequently for agents, and so requires more incentives and supporting enforcement than exchanging information which is less costly and/or more frequent.

Given that triangles can be incidentally generated, one cannot test this simply by examining the ratio of supported links to unsupported ones. If $\gamma_{L,info}$ was very high, then it could be that there are many incidentally generated information triangles, and few unsupported links, and so we need to estimate the underlying parameters using our techniques to account for incidental generation. To keep the illustration in this first example clear, we abstract from covariates. We illustrate the incorporation of covariates in the examples below.

TABLE 2. Parameter estimates by network type

	$\hat{\beta}^{\text{MD}}_{R,L}$	$\hat{\beta}^{\text{MD}}_{R,T}$
Information	0.0119	0.0001
	(0.0191)	(0.0001)
Favor	0.0109	0.0002
	(0.0307)	(0.0002)

Notes: Standard errors computed using the results of Proposition 2.

First, we estimate the four parameters in question under the many independent network (n fixed, $R \to \infty$) framework (Proposition 2). Table 2 presents the parameter estimates and standard errors. Although the point estimates are in line with the theory, the standard errors estimates are large and we cannot reject the null hypothesis that there is no difference in the support of favor relationships compared to information relationships.[35]

We cannot conclude that the data are consistent with the theory that incentives for favor exchange matters in network formation in these data, but in part because the parameters actually vary nontrivially across villages. This leads to high standard errors and also suggests that the more appropriate approach is to examine villages separately.

Thus, we push this further by estimating the parameters separately for each village v, with the large single network ($n \to \infty$, $R = 1$) paradigm for each village. This allows for heterogeneity in the parameters across villages by assuming they are drawn from entirely different distributions. Again, a we use the same variation on the minimum distance estimator of Proposition 3 used in Example 1.

We see the results in Figure 6, with standard errors omitted for visual clarity. We see that for almost all of villages, the favor over info ratios are higher for triangles compared to links (more than would occur at random at the 1 percent level if the two had the same distribution).

[35]Specifically, the p-value is computed for a test of the null hypothesis $\frac{\beta_{T,favor}}{\beta_{T,info}} = \frac{\beta^{3/2}_{L,favor}}{\beta^{3/2}_{L,info}}$, where the parameters are held to be common across all villages in the sample. If instead of using the conservative standard errors from Proposition 2, we bootstrap them then the hypothesis is rejected at the 1 percent level.

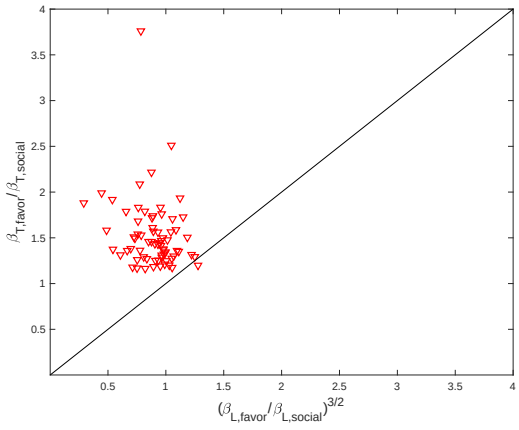


FIGURE 6. Plots of estimates of $\frac{\beta_{T,favor}}{\beta_{T,info}}$ against $\frac{\beta_{L,favor}^{3/2}}{\beta_{L,info}^{3/2}}$ by village.

5.4. **Example 3: Links across Social Boundaries.** Our next example shows how a SUGM can be used to investigate whether there are norms governing link-formation across different social groups. Identities can lead to strong social norms—prescriptions and proscriptions—concerning interactions across groups. For instance, in much of India there are strong forces that influence if and when individuals can form relationships across castes, particularly among "upper caste" Hindus and the "lower caste" communities, comprised of Dalits (Scheduled Castes, SC) and scheduled tribes (ST). The SC and ST communities 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 or subcaste distinctions, but those norms are more varied depending on the particular castes in question while this provides a clear barrier (Munshi and Rosenzweig, 2006).

Among many, one natural question concerns the norms around forming public versus private cross-caste group relationships. Namely, are members of upper and lower caste more likely to form cross-group 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 (and thus have a witness to the relationship)?

To answer this we need a model that accounts for link dependencies; cliques of three or more may exhibit greater adherence to a norm prohibiting certain inter-caste relationships, while the norm may be circumvented in isolated bilateral relationships. In particular, we can get at this hypothesis by testing whether the relative frequency of triangles compared to links is higher when the relationships are within caste than across caste.

This example is instructive because it is more subtle than the previous example and it demonstrates that a SUGM can be used for a hypothesis test even when preference parameters are not identifiable without additional restrictive assumptions. 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 include link interdependencies.

Individuals' probabilities to have opportunities to form links or triads can depend on the composition of castes of those involved. So let $\pi_L(diff), \pi_L(same)$ denote the probabilities 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. Analogously define $\pi_T(diff), \pi_T(same)$. Notice these are unlikely to be observed by the researcher.

As noted above, individual i's utility of having a relationship with j can by influenced by whether they share caste (x_{ij} a dummy variable for same caste) and is given by

$$u_i^L(j) = \alpha_{0,L} + \gamma_{0,L} x_{ij} - \epsilon_{L,ij}$$

and similarly for a triad,

$$u_i^T(jk) = \alpha_{0,L} + \gamma_{0,T} x_{ijk} - \epsilon_{T,i,jk},$$

where x_{ijk} is a dummy for whether all three individuals are members of the same caste.[36] The probability of an individual consenting to a subgraph of type $z \in \{L, T\}$ among the m_z nodes is

$$p_{z,same} = F(\alpha_{0,z} + \gamma_{0,z}) \text{ and } p_{z,diff} = F(\alpha_{0,z}).$$

The hypothesis that we explore is that

$$\frac{p_{T,diff}}{p_{T,same}} < \frac{p_{L,diff}}{p_{L,same}}$$

so that people are more reluctant to involve themselves in cross-caste relationships when those are "public" in the sense that other individuals observe those relationships.

The researcher does not observe either the meeting probabilities nor the probabilities within the mutual consent process. Rather, the researcher observes the compositions β_ℓ for $\ell \in \{L, T\} \times \{same, diff\}$ which are precisely SUGM parameters:

(1) $\beta_{L,same} = p_{L,same}^2 \pi_L(same)$ and $\beta_{L,diff} = p_{L,diff}^2 \pi_L(diff)$, and
(2) $\beta_{T,same} = p_{T,same}^3 \pi_T(same)$ and $\beta_{T,diff} = p_{T,diff}^3 \pi_T(diff)$.

There are two challenges. Recall the difference in the exponents reflects that it is more difficult to get a triangle to form than a link. Hence, to perform a proper 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. Further, identifying a preference bias is confounded by the meeting bias. Thus, we first model the meeting process $\pi_z(x)$ more explicitly and show that we still have identification as the meeting bias makes triangles relatively more likely to be cross-caste than links.

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

[36]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 Currarini, Jackson, and Pin (2009, 2010) to allow for additional link dependencies. We could also be interested in higher order relationships.

have f of its own types present and $1-f$ of the other type present.[37] This generates a conservative test in the sense that if we find cross-caste links relatively more likely, that is evidence for a preference bias.

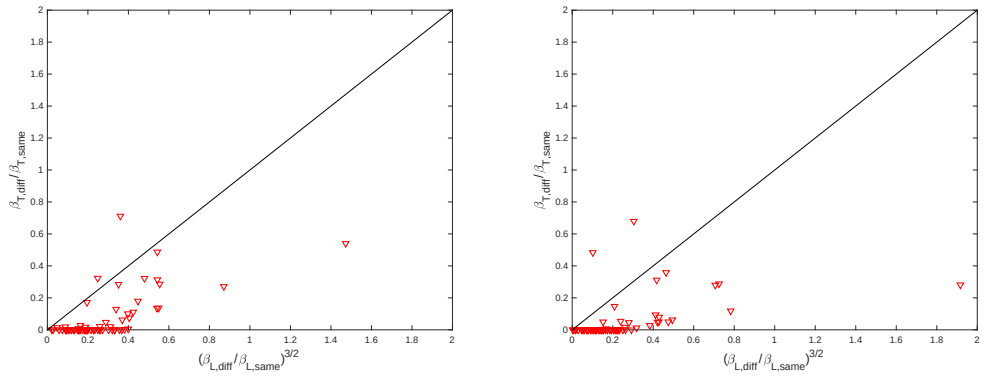
LEMMA **2.** *A sufficient condition for* $\frac{p_{T,diff}}{p_{T,same}} < \frac{p_{L,diff}}{p_{L,same}}$ *is that* $\frac{\beta_{T,diff}}{\beta_{T,same}} < \left(\frac{\beta_{L,diff}}{\beta_{L,same}}\right)^{3/2}$.

Turning to the data, we link two households if members of either engaged in favor exchange with each other: i.e., they borrowed or lent goods such as kerosene or rice in times of need.

TABLE 3. Parameter estimates by network type

	$\hat{\beta}^{\text{MD}}_{R,L,same}$	$\hat{\beta}^{\text{MD}}_{R,T,same}$	$\hat{\beta}^{\text{MD}}_{R,L,diff}$	$\hat{\beta}^{\text{MD}}_{R,T,diff}$
Information	0.016891	0.000309	0.006535	0.000002
	(0.020001)	(0.000156)	(0.008739)	(0.000048)
Favor	0.012652	0.000302	0.004281	0.000003
	(0.024433)	(0.000194)	(0.006796)	(0.000038)

Table 3 presents the parameter estimates using the estimator from Proposition 2, for the estimation in which we assume that all 75 networks are independent draws from the same distribution. Similar to the example above, the large standard errors on the triangle parameters lead us to fail to reject the null hypothesis that $\frac{p_T(diff)}{p_T(same)} = \left(\frac{p_L(diff)}{p_L(same)}\right)^{3/2}$. Again, this suggests that since the parameters vary by village, that we should work with estimation by village.



(A) Information (B) Favors

FIGURE 7. Plots of estimates of $\frac{\beta_{T,diff}}{\beta_{T,same}}$ against $\frac{\beta^{3/2}_{L,diff}}{\beta^{3/2}_{L,same}}$ by village for information and favors networks.

[37] Variations on this sort of biased meeting process appear in Currarini, Jackson, and Pin (2009, 2010); Bramoullé, Currarini, Jackson, Pin, and Rogers (2012).

Finally, Figure 7 shows the results when we allow the parameter estimates to vary by village, again using the variation on the estimator of Proposition 3 that we used in Example 1. For the bulk of villages, cross-caste relationships relative to within-caste relationships are more frequent as isolated links compared to being embedded in triangles, for both information and favor networks. This is a new observation in the literature and begs the question as to the foundations as to why members of two groups which meet less frequently and may have less affinity for each other may nonetheless be able to sustain interactions privately. They appear to give up support for the sake of not having the interaction be public.

6. A Central Limit Theorem for Correlated Random Variables

We now state a new central limit theorem that applies to a variety of settings in which all variables may be correlated (well-beyond network settings) but the total amount of covariance is bounded. We require it to prove Proposition 3, but the result is considerably more general. Our result is derived from Stein's method, which provides the foundations for many CLTs with dependent variables, but as will become clear, our treatment delivers more general conditions that also allow us to cover SUGMs (Stein, 1986).

To understand the necessity of such a theorem at a high intuitive level, begin with the simplest example where $n = 3$ and the only possible subgraphs are a triangle and three links. Here if there is a triangle, then there are necessarily three links, and if there is no triangle there are necessarily fewer than three links. There is a strong dependence in the subgraph counts, and this carries over in the estimates of parameters. The overall dependence of different subgraph counts is not so extreme as n grows, but there is always a correlation and adjacent subgraphs remain nontrivially correlated all along the sequence. Thus, we need a theorem that covers moments that are based on variables, some of which are highly correlated all along the sequence.

There are essentially two approaches used to prove CLTs relevant that allow for correlated random variables and the existing results in neither setting apply to SUGMs.

The first approach to CLTs with dependence we can call "geometric": random variables carry indices locating them in some embedding space. This covers time series, spatial data, mixing random fields. 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. With enough moment conditions, using distance-based limits on correlation allows researchers to leverage Stein's lemma under mixing conditions to derive a CLT (e.g., (Bolthausen, 1982; Jenish and Prucha, 2009) just to name a few). In fact, a literature evolved focusing on mixing random fields leveraging such conditions (Froot, 1989; Conley, 1999; Driscoll and Kraay, 1998).[38] Some researchers working on network formation (e.g., Boucher and Mourifié (2017); Leung (2014)) exploit such spatial techniques by embedding nodes in some space so that only "nearby" nodes can link and "distant" nodes cannot link (e.g., following the logic in Penrose (2003); Hoff et al. (2002)) in order to satisfy mixing conditions and apply a

[38]See also Kuersteiner (2019), which studies conditional mixingale types of assumptions so that nodes that are "far" in characteristic space have decaying dependence.

central limit theorem like Bolthausen (1982). As $n \to \infty$ most nodes get further and further apart and therefore essentially never link.

SUGMs do not lend themselves to ordering the indices of random variables into time, space, lattices, or more general random fields. The reason that this does not work as it imposes specific structure on the adjacency matrix. For example, 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 to that node. While this is 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 university who are, ex ante, infinitely unlikely to link to each other, etc.). We can prove a CLT without this structure and that nests such previous theorems, and so it is worthwhile to do so.

A second approach to CLTs with dependence is to use dependency graphs (Baldi and Rinott (1989); Goldstein and Rinott (1996); Chen and Shao (2004); Ross (2011)).[39] There is another graph (not a SUGM but a graph among indices of random variables), where edges between random variable indices indicates conditional dependence (and independence if there is no edge).[40] While that literature leverages Stein (1986) to prove CLTs despite not forcing a geometric structure, they require most entries of the dependency graph to be zero. That is, they impose a sparse (dependency) graph with independence across many pairs of random variables (and usually do not consider triangular arrays). SUGMs violate both impositions since at each n, all links can have non-zero correlations with each other.

Our insight is that if the overall covariances satisfy some bounds, then one can still prove a CLT no matter how that dependency is arranged and even without numerous conditional independencies.

Note that both the geometric or dependency graph approaches limit the total correlation of the $\binom{n}{2}$ random variables with specific structure. We develop a CLT in which all random variables can have non-zero correlation by controlling the total size but not forcing zeros. We show that if one can "collect" what we call *affinity sets* for each index—that is a set of other random variables that can have larger correlation with the reference random variable—then as long as three covariance conditions hold, we can limit the overall correlation and apply Stein's method. Our conditions are that on average (1) within an affinity set, most of the correlation comes from the dependence between the reference variable and its members rather from than between two members; (2) the amount of correlation between members of two different affinity sets is small; (3) the correlation between the reference variable and all those outside its affinity sets is appropriately small. We argue that these are relatively intuitive (weighted covariance conditions), interpretable, easy to check, and more easily

[39] Aronow and Samii (2017) use Chen and Shao (2004) assuming conditional independence in treatment effects with spillovers, where certain interferences are ruled out. But this means one cannot study spillovers due to information, for example, because in principle information can flow to any node in the network, so treatment of one node leaves non-zero exposure to all others. Another application of Chen and Shao (2004) is in Leung and Moon (2022). They characterize dependence through radii of stability which correspond to changes in network features when dropping observed links.

[40] Some work, e.g., Chen and Shao (2004), focuses on exact finite sample Berry-Essen inequalities.

micro-founded than potentially a more complex but specific assumption on things like mixing random fields. It is in this sense that we believe this is of independent interest. In a follow-up paper (Chandrasekhar, Jackson, McCormick, and Thiyageswaran, 2023) we provide a number of other (non-network) applications and additional discussion of the literature.

We require some notation.

Consider a triangular array of (real-valued) random variables X_α^N with a set of labels $\alpha \in \Lambda^N$ such that $|\Lambda^N| = N$. For instance, in our SUGM setting the X_α^N may be an indicator of the appearance of some particular subgraph, such as a link or triangle, and α would track the pairs of nodes involved in a potential link (ij) or triples of nodes in a triangle (ijk). N captures the $\binom{n}{2}$ possible links or $\binom{n}{3}$ possible triangles. So when considering link counts α would track pairs of nodes involved in links and when considering triangles α would track triples.

Let us normalize the variables by their means:

$$Z_\alpha^N = X_\alpha^N - \mathrm{E}\left[X_\alpha^N\right].$$

We presume that the Z_αs are such that the $\mathrm{E}[|Z_\alpha^N|^3]/\mathrm{E}[(Z_\alpha^N)^2]^{3/2}$ is bounded above for all α, N.[41] We provide conditions under which a normalized statistic, as $N \to \infty$, converges to the standard normal distribution,

$$\frac{\sum_{\alpha \in \Lambda^N} Z_\alpha^N}{\sqrt{a_N}} \rightsquigarrow \mathcal{N}(0,1),$$

where the normalizer, a_N, is a measure of the variance of the sum, defined below.

6.1. **Affinity Sets.** For each α, N, we partition the index set Λ^N into two pieces: an *affinity set* and its complement. In particular, we define an affinity set, for each α, N, as

$$\mathcal{A}(\alpha, N) \subset \Lambda^N \text{ such that } \alpha \in \mathcal{A}(\alpha, N).$$

The conditions for $\eta \in \mathcal{A}(\alpha, N)$ are defined below. It is crafted in a specific manner to satisfy a few sufficient conditions for a CLT.

$\mathcal{A}(\alpha, N)$ includes indices η where the corresponding X_η's have relatively "high" correlation with X_α, and its complement includes the indices η where the corresponding X_η's that have relatively "low" correlation with X_α. There is substantial freedom in defining these sets, but an easy rule to applying them to (non-sparse) SUGMs is to set the $\mathcal{A}(\alpha, N)$ sets to include the other tuples of nodes with which the reference tuple of nodes shares some potential edges and therefore could face incidental generation.

We show that under conditions on the relative correlations inside and outside of the affinity sets a central limit theorem applies.

[41] This condition holds, for instance, if the Zs are from Bernoulli random variables, or if $\mathrm{E}[|Z_\alpha^N|^3]$ is bounded above and $\mathrm{E}[(Z_\alpha^N)^2]$ is bounded below across α, N, but also in cases where such individual bounds do not hold. Violations of this condition are, for example, random variables with finite second moments but infinite third moments.

6.2. **The Central Limit Theorem.** Let

$$a_N := \sum_{\alpha;\eta\in\mathcal{A}(\alpha,N)} \operatorname{cov}\left(Z_\alpha, Z_\eta\right),$$

be the total sum of variance-covariances across all the pairs of variables in each other's affinity sets , and recall this was used in normalizing the total sum above. In what follows, we maintain the assumption that $a_N \to \infty$, as otherwise there is insufficient variation to obtain a central limit theorem.

Finally, let $\mathbf{Z}_{-\mathcal{A}(\alpha,N)} := \sum_{\eta\notin\mathcal{A}(\alpha,N)} Z_\eta$ be the sum over all random variables not in the reference index α's affinity set.

The following are the key conditions for the theorem:

$$\sum_{\alpha;\eta,\gamma\in\mathcal{A}(\alpha,N)} \mathrm{E}\left[\,|Z_\alpha| Z_\eta Z_\gamma\right] = o\left(a_N^{3/2}\right), \tag{6.1}$$

$$\sum_{\alpha,\alpha',\eta\in\mathcal{A}(\alpha,N),\eta'\in\mathcal{A}(\alpha',N)} \operatorname{cov}\left(Z_\alpha Z_\eta, Z_{\alpha'} Z_{\eta'}\right) = o\left(a_N^2\right), \tag{6.2}$$

$$\sum_\alpha \mathrm{E}\left[\left|\mathrm{E}\left[Z_\alpha \mathbf{Z}_{-\mathcal{A}(\alpha,N)} \,\middle|\, \mathbf{Z}_{-\mathcal{A}(\alpha,N)}\right]\right|\right] = \sum_\alpha \mathrm{E}\left[\left|\mathbf{Z}_{-\mathcal{A}(\alpha,N)} \mathrm{E}\left[Z_\alpha \,\middle|\, \mathbf{Z}_{-\mathcal{A}(\alpha,N)}\right]\right|\right] = o\left(a_N\right). \tag{6.3}$$

Condition (6.1) captures the idea that most of the covariance between random variables in an affinity set α comes from covariances between the reference random variable X_α and one member of the neighborhood X_η, rather than from covariance between two other members X_η and X_γ. Some of them can have high covariance, but in total they cannot. The term on the left-hand-side consists of an integral over Z_α of covariances between Z_γ and Z_η, weighted by $|Z_\alpha|$, so the covariances cannot be too large exactly with large values of Z_α. So in constructing our normalizer a_N we need only consider the covariance terms between reference variables and members of their affinity sets.

Condition (6.2) is similar but it looks at the covariance between two members (η, η') of different affinity sets of two distinct reference nodes (α, α'). It says, again, that the *total* amount of covariance across members of different affinity sets, when considering any two pairs of reference nodes, is small relative to the total sum of variances.

Condition (6.3) states that covariances between reference nodes and all members outside of its affinity set are relatively small. This is intuitive and motivates the strategy in defining affinity sets in the first place. Note that if, for instance, $\mathrm{E}[Z_\alpha \mathbf{Z}_{-\mathcal{A}(\alpha,N)} | \mathbf{Z}_{-\mathcal{A}(\alpha,N)}] \geq 0$, then Condition (6.3) is simply that $\sum_{\alpha;\eta\notin\mathcal{A}(\alpha,N)} \operatorname{cov}\left(Z_\alpha Z_\eta\right) = o\left(a_N\right)$.

Although these conditions may seem complex, in Chandrasekhar, Jackson, McCormick, and Thiyageswaran (2023) we show how they can be directly verified in a number of applications.

THEOREM **4.** *If (6.1)-(6.3) are satisfied, then* $\sum_{\alpha\in\Lambda^N} Z_\alpha^N / \sqrt{a_N} \rightsquigarrow \mathcal{N}(0,1)$.

It is useful to consider the special case in which $\mathcal{A}(\alpha, N) = \{\alpha\}$, which extends but nests many standard central limit theorems. Here we use the notation $\mathbf{Z}_{-\alpha}$ to denote $\mathbf{Z}_{-\mathcal{A}(\alpha,N)}$ This is useful when we get to the case of sparse networks, where incidental networks are unlikely and the correlation between different subgraphs becomes small.

COROLLARY **2.** *If* $\mathrm{E}\left[Z_\alpha \mathbf{Z}_{-\alpha} | \mathbf{Z}_{-\alpha}\right] \geq 0$ *for every* α*, and*[42]

(i) $\sum_{\alpha,\eta} \mathrm{cov}(Z_\alpha^2, Z_\eta^2) = o\left(a_N^2\right)$*, and*
(ii) $\sum_{\alpha \neq \eta} \mathrm{cov}(Z_\alpha, Z_\eta) = o\left(a_N\right)$*,*

then $\sum_{\alpha \in \Lambda^N} Z_\alpha^N / \sqrt{a_N} \rightsquigarrow \mathcal{N}(0,1)$*.*

Moreover, if the X_α*s are Bernoulli random variables with* $\mathrm{E}[X_\alpha] \to 0$ *(uniformly), then (ii) implies (i).*

Note that (i) is often satisfied whenever (ii) is, so this is an easy corollary based on one intuitive condition: the overall sum of covariances between different variables cannot be too large relative to the sum of their variances.

7. Concluding Remarks

We have developed a new class of models—SUGMs—in which networks are formed via a basis set of subgraphs. The parameters are always identified and we study conditions when the parameters have estimators that are consistent and asymptotically normally distributed. We provide four such estimators to cover various data settings. En route, we develop a new central limit theorem for dependent random variables which extends the dependency graph literature and also does not require a geometric (lattice-like) ordering of covariances of the kind used in the time series and spatial literatures. We believe this is of independent interest.

Our model is useful for empirical work. We show that it models economically relevant features of real-world network data better than the standard alternatives: stochastic block models, unobserved heterogeneity models, latent space models, and ERGMs. Further, we have illustrated that it is easy to microfound a SUGM to test important hypotheses such as whether a network provides incentives to sustain informal contracts or whether people are willing to interact across caste publicly.

Future research can explore, among other things, richer inclusion of covariates in subgraphs, a data-driven approach to select subgraphs for inclusion in the model, statistical properties of other specific empirically-relevant SUGMs not studied here, and systematic bootstrap techniques for inference for use in complex implementations of these models.

Data Availability Statement: The code and data needed to reproduce all figures and tables in the paper are available at http://doi.org/10.5281/zenodo.14218442

References

ACEMOGLU, D., A. OZDAGLAR, AND A. TAHBAZ-SALEHI (2015): "Systemic Risk and Stability in Financial Networks," *American Economic Review*, 105(2), 564–608. 1

ANDREWS, D. W. (1992): "Generic uniform convergence," Tech. Rep. 2. C.1, C.2

ARONOW, P. M. AND C. SAMII (2017): "Estimating average causal effects under general interference, with application to a social network experiment," *Annals of Applied Statistics*, 11, 1912–1947. 39

[42] If $\mathrm{E}\left[Z_\alpha \mathbf{Z}_{-\alpha} | \mathbf{Z}_{-\alpha}\right] \geq 0$ does not hold, then (ii) can just be substituted by (6.3).

Badev, A. (2021): "Nash equilibria on (un) stable networks," *Econometrica*, 89, 1179–1206. 1.1

Bailey, M., R. Cao, T. Kuchler, and J. Stroebel (2016): "Social Networks and Housing Markets," *unpublished manuscript*. 3.1

Baldi, P. and Y. Rinott (1989): "On normal approximations of distributions in terms of dependency graphs," *The Annals of Probability*, 1646–1650. 6

Banerjee, A., A. Chandrasekhar, E. Duflo, and M. Jackson (2013): "Diffusion of Microfinance," *Science*, 341, DOI: 10.1126/science.1236498, July 26 2013. 1, 3.1, 5.1, 26

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2019): "Using gossips to spread information: Theory and evidence from two randomized controlled trials," *The Review of Economic Studies*, 86, 2453–2490. 1.2, 5.1

Barabasi, A. and R. Albert (1999): "Emergence of scaling in random networks," *Science*, 286, 509. 1.1, 5.2

Bester, C. A., T. G. Conley, and C. B. Hansen (2011): "Inference with dependent data using cluster covariance estimators," *Journal of Econometrics*, 165, 137–151. 13

Bhamidi, S., G. Bresler, and A. Sly (2008): "Mixing time of exponential random graphs," *Arxiv preprint arXiv:0812.2265*. 1.1

Bhattacharyya, S., P. J. Bickel, et al. (2015): "Subsampling bootstrap of count features of networks," *Annals of Statistics*, 43, 2384–2411. 12

Bickel, P., A. Chen, and E. Levina (2011): "The method of moments and degree distributions for network models," *Annals of Statistics*, 39, 2280–2301. 2

Biscio, C. A. N., A. Poinas, and R. Waagepetersen (2018): "A note on gaps in proofs of central limit theorems," *Statistics & Probability Letters*, 135, 7–10. C.3.2, C.3.2

Blitzstein, J. and P. Diaconis (2011): "A sequential importance sampling algorithm for generating random graphs with prescribed degrees," *Internet mathematics*, 6, 489–522. 3

Blumenstock, J., N. Eagle, and M. Fafchamps (2011): "Charity and Reciprocity in Mobile Phone-Based Giving: Evidence from Rwanda," *mimeo: University of California, Berkeley*. 3.1

Bollobas, B. (2001): *Random Graphs*, Cambridge University Press. 4.3.1, 4.3.3, 5.2

Bollobás, B., C. Borgs, J. Chayes, and O. Riordan (2010): "Percolation on dense graph sequences," *The Annals of Probability*, 38, 150–183. 5.2

Bollobás, B., S. Janson, and O. Riordan (2011): "Sparse random graphs with clustering," *Random Structures & Algorithms*, 38, 269–323. 11

Bolthausen, E. (1982): "On the central limit theorem for stationary mixing random fields," *Annals of Probability*, 10, 1047–1050. 13, 6

Boucher, V. and I. Mourifié (2017): "My friend far, far away: a random field approach to exponential random graph models," *The econometrics journal*, 20, S14–S46. 1.1, 6

Bramoullé, Y., S. Currarini, M. Jackson, P. Pin, and B. Rogers (2012): "Homophily and Long-Run Integration in Social Networks," *Journal of Economic Theory*, 147, 1754–1786. 1.1, 37

Breza, E., A. G. Chandrasekhar, T. H. McCormick, and M. Pan (2020): "Using aggregated relational data to feasibly identify network structure without network data," Tech. Rep. 8. 7

Butts, C. (2009): "Using potential games to parameterize erg models," *University of California, Irvine working paper*. 1.1

Cai, J., A. deJanvry, and E. Sadoulet (2015): "Social Networks and the Decision to Insure," *American Economic Journal: Applied Economics*, 7:2, 81–108. 1

Calvo-Armengol, A. and M. Jackson (2007): "Networks in labor markets: Wage and employment dynamics and inequality," *Journal of Economic Theory*, 132, 27–46. 5.2

Calvo-Armengol, A., E. Patacchini, and Y. Zenou (2009): "Peer Effects and Social Networks in Education," *The Review of Economic Studies*, 76, 1239–1267. 1

Carrell, S. E., B. I. Sacerdote, and J. E. West (2013): "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, 81, 855–882. 1

Chandrasekhar, A. and M. Jackson (2012): "Tractable and Consistent Random Graph Models," *SSRN Working Paper: http://ssrn.com/abstract=2150428*. 1.1

Chandrasekhar, A. G. (2016): "Econometrics of network formation," *The Oxford Handbook of the Economics of Networks*, 303–357. 20

Chandrasekhar, A. G., M. O. Jackson, T. H. McCormick, and V. Thiyageswaran (2023): "General Covariance-Based Conditions for Central Limit Theorems with Dependent Triangular Arrays," *arXiv preprint arXiv:2308.12506*. 6, 6.2

Chaney, T. (2016): "Networks in international trade," *The Oxford Handbook on the Economics of Networks, edited by Bramoullé, Yann, Andrea Galeotti and Brian Rogers, Oxford: Oxford University Press*. 1

Charbonneau, K. B. (2017): "Multiple fixed effects in binary response panel data models," *The Econometrics Journal*, 20, S1–S13. 3

Chatterjee, S., P. Diaconis, and A. Sly (2010): "Random graphs with a given degree sequence," *Arxiv preprint arXiv:1005.1136*. 1.1, 3, 30

Chen, L. H. and Q.-M. Shao (2004): "Normal approximation under local dependence," *The Annals of Probability*, 32, 1985–2028. 6, 39, 40

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebl, N. Hendren, R. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, D. Johnston, M. Koenen, E. Laguna-Muggenburg, F. Mudekereza, T. Rutter, N. Thor, W. Townsend, R. Zhang, M. Bailey, P. Barbera, M. Bhole, and N. Wernerfelt (2022): "Social Capital in the United States I: Measurement and Associations with Economic Mobility," *Nature*, 608. 3.1

Christakis, N., J. Fowler, G. W. Imbens, and K. Kalyanaraman (2020): "An empirical model for strategic network formation," in *The Econometric Analysis of Network Data, edited by A. de Paula and B. Graham*, Elsevier, 123–148. 1.1

Coleman, J. (1988): "Social Capital in the Creation of Human Capital," *American Journal of Sociology*, 94, S95–S120. 5

Conley, T. G. (1999): "GMM estimation with cross sectional dependence," *Journal of econometrics*, 92, 1–45. 13, 6

Currarini, S., M. Jackson, and P. Pin (2009): "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 77, 1003–1045. 1.1, 3.1, 4.3.1, 5.2, 36, 37

——— (2010): "Identifying the roles of race-based choice and chance in high school friendship network formation," *Proceedings of the National Academy of Sciences*, 107, 4857–4861. 1.1, 4.3.1, 5.2, 36, 37

De Paula, A. (2017): "Econometrics of network models," in *Advances in economics and econometrics: Theory and applications, eleventh world congress*, Cambridge University Press Cambridge, 268–323. 9

De Paula, Á., S. Richards-Shubik, and E. Tamer (2018): "Identifying preferences in networks with bounded degree," *Econometrica*, 86, 263–288. 1.1

DeMarzo, P., D. Vayanos, and J. Zwiebel (2003): "Persuasion Bias, Social Influence, and Unidimensional Opinions*," *Quarterly journal of economics*, 118, 909–968. 5.2

Driscoll, J. C. and A. C. Kraay (1998): "Consistent covariance matrix estimation with spatially dependent panel data," *Review of Economics and Statistics*, 80, 549–560. 6

Elliott, M., B. Golub, and M. O. Jackson (2014): "Financial Networks and Contagion," *American Economic Review*, 104(10), 3115–3153. 1

Frank, O. and D. Strauss (1986): "Markov graphs," *Journal of the American Statistical Association*, 832–842. 4

Froot, K. A. (1989): "Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data," *Journal of Financial and Quantitative analysis*, 24, 333–355. 6

Gai, P. and S. Kapadia (2010): "Contagion in financial networks," *Proceedings of the Royal Society A*, 466, 2401–2423. 1

Glaeser, E. L., B. Sacerdote, and J. A. Scheinkman (1996): "Crime and Social Interactions," *The Quarterly Journal of Economics*, 111(2), 507–548. 1

Goldstein, L. and Y. Rinott (1996): "Multivariate normal approximations by Stein's method and size bias couplings," *Journal of Applied Probability*, 1–17. 6

Golub, B. and M. Jackson (2012): "How Homophily Affects the Speed of Learning and Best-Response Dynamics," *Quarterly Journal of Economics*, 127, 1287–1338. 5.2

Graham, B. S. (2017): "An econometric model of network formation with degree heterogeneity," *Econometrica*, 85, 1033–1063. 1.1, 3, 4.3.3, 5.2, 30

Greene, W. H. (2000): "Econometric analysis 4th edition," *International edition, New Jersey: Prentice Hall*, 201–215. 45

Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002): "Latent space approaches to social network analysis," *Journal of the american Statistical association*, 97, 1090–1098. 1.1, 5.2, 6

Holland, P. W. and S. Leinhardt (1981): "An exponential family of probability distributions for directed graphs: Rejoinder," *Journal of the American Statistical Association*, 76, 62–65. 3

Jackson, M. (2008): *Social and economic networks*, Princeton: Princeton University Press. 1.1, 4.3, 4.3.3, 5.2

Jackson, M. and B. Rogers (2007): "Meeting strangers and friends of friends: How random are social networks?" *The American economic review*, 97, 890–915. 1.1, 6, 5.2

Jackson, M. and A. Watts (2001): "The Existence of Pairwise Stable Networks," *Seoul Journal of Economics*, 14(3), 299–321. 1.1

Jackson, M. and A. Wolinsky (1996): "A Strategic Model of Social and Economic Networks," *Journal of Economic Theory*, 71, 44–74. 1.1, 5.3.1

Jackson, M. O. (2005): "A Survey of Models of Network Formation: Stability and Efficiency," *Group Formation in Economics; Networks, Clubs and Coalitions, ed. G Demange, M Wooders. Cambridge, UK: Cambridge Univ. Press.* 1.1

——— (2019): *The Human Network: How Your Social Position Determines Your Power, Beliefs, and Behaviors*, Pantheon. 1

——— (2023): "Inequality's Economic and Social Roots: The Role of Social Networks and Homophily," *in Advances in Economics and Econometrics, Theory and Applications: Twelfth World Congress of the Econometric Society, Cambridge University Press, https://dx.doi.org/10.2139/ssrn.3795626.* 5.2

Jackson, M. O. and S. Nei (2015): "Networks of Military Alliances, Wars, and International Trade," *Proceedings of the National Academy of Sciences*, 112(50), 15277–15284. 1

Jackson, M. O., T. Rodriguez-Barraquer, and X. Tan (2012): "Social Capital and Social Quilts: Network Patterns of Favor Exchange," *American Economic Review*, 102, 1857–1897. 1.2, 5, 5.3.2

Jackson, M. O., B. W. Rogers, and Y. Zenou (2016): "The Economic Consequences of Social Network Structure," *Journal of Economic Literature (forthcoming).* 1

Jackson, M. O. and E. C. Storms (2017): "Behavioral Communities and the Atomic Structure of Networks," *Arxiv: https://arxiv.org/abs/1710.04656.* 2

Jenish, N. and I. R. Prucha (2009): "Central limit theorems and uniform laws of large numbers for arrays of random fields," *Journal of Econometrics*, 150, 86–98. 13, 6, C.2

Karlan, D., M. Mobius, T. Rosenblat, and A. Szeidl (2009): "Trust and Social Collateral*," *Quarterly Journal of Economics*, 124, 1307–1361. 5.2

König, M. D., D. Rohner, M. Thoenig, and F. Zilibotti (2017): "Networks in conflict: Theory and evidence from the great war of Africa," *Econometrica*, 85, 1093–1132. 1

Krackhardt, D. (1988): "Predicting with Networks: Nonparameteric Multiple Regression Analysis of Dyadic Data," *Social Networks*, 10, 359–381. 4

Kuersteiner, G. M. (2019): "Limit theorems for data with network structure," *arXiv preprint arXiv:1908.02375.* 38

Leung, J. and H. Moon (2022): "A central limit theorem for latent networks," *arXiv preprint arXiv:2204.05885.* 39

Leung, M. (2014): "A Random-Field Approach to Inference in Large Models of Network Formation," *Stanford Working Paper.* 1.1, 6

Lubold, S., A. G. Chandrasekhar, and T. H. McCormick (2023): "Identifying the latent space geometry of network models through analysis of curvature," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85, 240–292. 8, 31

McCormick, T. H. and T. Zheng (2015): "Latent surface models for networks using Aggregated Relational Data," *Journal of the American Statistical Association*, 110, 1684–1695. 7

McFadden, D. (1973): "Conditional logit analysis of qualitative choice behavior," *Institute of Urban and Regional Development, University of California*. 5.3.2

McPherson, M., L. Smith-Lovin, and J. Cook (2001): "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, 27, 415–444. 5.2

Mele, A. (2017): "A structural model of dense network formation," *Econometrica*, 85, 825–850. 5, 1.1

——— (2022): "A structural model of homophily and clustering in social networks," *Journal of Business & Economic Statistics*, 40, 1377–1389. 5

Mele, A. and L. Zhu (2023): "Approximate variational estimation for a model of network formation," *Review of Economics and Statistics*, 105, 113–124. 5

Munshi, K. and M. Rosenzweig (2006): "Traditional institutions meet the modern world: Caste, gender, and schooling choice in a globalizing economy," *The American Economic Review*, 96, 1225–1252. 5.4

Newey, W. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245. 4.2, A, C.3.3

Newman, M., S. Strogatz, and D. Watts (2001): "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, 64, 26118. 4.3.3

Newman, M. E. (2003): "The Structure and Function of Complex Networks," *SIAM Review*, 45:2, 167–256. 4.3

Park, J. and M. E. Newman (2004): "Statistical mechanics of networks," *Physical Review E*, 70, 066117. 3

Patacchini, E. and Y. Zenou (2008): "The strength of weak ties in crime," *European Economic Review*, 52, 209–236. 1

Penrose, M. (2003): *Random Geometric Graphs*, Oxford University Press. 1.1, 6

Pötscher, B. and I. Prucha (1997): *Dynamic Nonlinear Econometric Models*, Springer-Verlag, New York. C.1, C.2

Pötscher, B. M. and I. R. Prucha (1991): "Basic structure of the asymptotic theory in dynamic nonlinear econometric models," *Econometric Reviews*, 10, 253–325. C.1

Ross, N. (2011): "Fundamentals of Stein's method," *Probab. Surv*, 8, 210–293. 6, B.1, B.2

Shalizi, C. and A. Rinaldo (2012): "Consistency under Sampling of Exponential Random Graph Models," *ArXiv 1111.3054v3*. 1.1

Sheng, S. (2020): "A structural econometric analysis of network formation games through subnetworks," *Econometrica*, 88, 1829–1858. 1.1

Stein, C. (1972): "A bound for the error in the normal approximation of a sum of dependent random variables," *In: Proceedings of Berkeley Symposium M.S.P.*, 2, 583–603. 13

——— (1986): "Approximate computation of expectations," *Lecture Notes-Monograph Series*, 7, i–164. 1.2, 6, B.1, B.1

Wasserman, S. and P. Pattison (1996): "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp," *Psychometrika*, 61, 401–425. 4

Watts, D. and S. Strogatz (1998): "Collective dynamics of small-world networks," *Nature*, 393, 440–442. 5.2

Appendix A. Proofs

Proof of Lemma 1. Note that for $z \in \{favor, info\}$, $\frac{\gamma_{T,z}}{\gamma_{L,z}} = \frac{\beta_{T,z}^{1/3}}{\beta_{L,z}^{1/2}}$ and so the condition becomes $\frac{\beta_{T,favor}^{1/3}}{\beta_{L,favor}^{1/2}} > \frac{\beta_{T,info}^{1/3}}{\beta_{L,info}^{1/2}}$ from which the result directly follows. ■

Proof of Lemma 2. 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.[43] Thus, the relative meeting frequency of different type links compared same type links is

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

For triangles, picking three individuals out of the community at any point in time would lead to a $f^3 + (1-f)^3$ probability that all three are of the same type, and $1 - (f^2 + (1-f)^2)$ of them being of mixed types, and so

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

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

$$\frac{\pi_T(same)}{\pi_T(diff)} < \frac{\pi_L(same)}{\pi_L(diff)}. \tag{A.1}$$

So different type triangles are more likely to have opportunities to form under this random mixing model than different type links. In particular, note that $\frac{p_{T,diff}}{p_{T,same}} < \frac{p_{L,diff}}{p_{L,same}}$ if and only if $\left(\frac{\beta_{T,diff}}{\beta_{T,same}} \frac{\pi_T(diff)}{\pi_T(same)}\right)^{1/3} < \left(\frac{\beta_{L,diff}}{\beta_{L,same}} \frac{\pi_L(same)}{\pi_L(diff)}\right)^{1/2}$. In summary, given (A.1), sufficient condition for $\frac{p_{T,diff}}{p_{T,same}} < \frac{p_{L,diff}}{p_{L,same}}$ is that $\frac{\beta_{T,diff}}{\beta_{T,same}} < \left(\frac{\beta_{L,diff}}{\beta_{L,same}}\right)^{3/2}$. ■

Proof of Theorem 1. Order subgraph types so that the number of links in a subgraph of type ℓ is nondecreasing in ℓ. Let ℓ^* be the smallest ℓ for which $\beta_\ell \neq \beta'_\ell$.

Consider a particular subgraph $g' \in G_{\ell^*}$ with labeled nodes. Let $p_\beta(g')$ denote the probability that the subgraph g' (without any extra links within the subgraph nor any links anywhere else in the network) forms from some collection of subgraphs in G_ℓ for $\ell < \ell^*$. We can then write the probability of forming the subgraph g' (and no other links anywhere) as

$$p_\beta(g') + (1 - p_\beta(g'))\beta_{\ell^*},$$

where recall that β_{ℓ^*} is the probability that g' forms directly. Let $no_\beta(g')$ denote the probability that all $g'' \in G_\ell$ for $\ell < \ell^*$ such that $g'' \subset g'$ *do not* form. Then the probability that none of the links in g' are present as parts of subgraphs that do not extend beyond g' is then

$$no_\beta(g')(1 - \beta_{\ell^*}).$$

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

Let $\emptyset$ denote the empty network. It then follows that

$$\frac{\mathrm{P}_\beta(g')}{\mathrm{P}_\beta(\emptyset)} = \frac{p_\beta(g') + (1 - p_\beta(g'))\beta_{\ell^*}}{no_\beta(g')(1 - \beta_{\ell^*})}.$$

So the probability that the realized network is exactly g' compared to the probability that it is the empty network (which has positive probability given that each $\beta_\ell < 1$), depends only on the probability that g' forms directly or incidentally from subgraphs of it, over the probability that no subgraph of g' (including itself) forms.

Note that this expression is strictly increasing in β_{ℓ^*} since $p_\beta(g') < 1$ and $no_\beta(g') > 0$. By the definition of ℓ^*: $p_\beta(g') = p_{\beta'}(g') < 1$ and $no_\beta(g') = no_{\beta'}(g')$. It then follows that

$$\frac{\mathrm{P}_\beta(g')}{\mathrm{P}_\beta(\emptyset)} \neq \frac{\mathrm{P}_{\beta'}(g')}{\mathrm{P}_{\beta'}(\emptyset)},$$

which establishes the claim. ■

Proof of Proposition 1. First, note that $1 - (1 - \beta_T)^x$ 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[44]

$$\widetilde{q}_L(\beta_L, \beta_T) = \beta_L + (1 - \beta_L)\left(1 - (1 - \beta_T)^{n-3}\right). \tag{A.2}$$

To see the derivation, note that:

$$\begin{aligned}
\widetilde{q}_L(\beta_L, \beta_T) = \mathrm{P}(g_{ij=1} \mid T_{ijk} = 0) &= \frac{\mathrm{P}(g_{ij} = 1 \cap T_{ijk} = 0)}{\mathrm{P}(T_{ijk} = 0)} \\
&= \frac{\mathrm{P}(g_{ij} = 1 \cap T_{ijk} = 0)}{1 - \beta_T} \text{ by definition} \\
&= \frac{\underbrace{\beta_L(1 - \beta_T)}_{(i)} + \underbrace{(1 - \beta_L)(1 - (1 - \beta_T)^{n-3})(1 - \beta_T)}_{(ii)}}{1 - \beta_T}
\end{aligned}$$

where T_{ijk} is an indicator for the direct triangle on i, j, k forming. Here (i) captures when the link forms directly and the indicated direct triangle does not and (ii) captures when the link does not form directly and some other triangle generates it and the indicated directed triangle does not.

Given this, note that the probability that a link forms can be written as

$$\mathrm{E}_\beta(S_L(g)) = q_L = \beta_T + (1 - \beta_T)\widetilde{q}_L(\beta_L, \beta_T), \tag{A.3}$$

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. Note that this is equivalent to saying a link can be formed

[44] That 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.

as part of some triangle, or else that it could form by other means that exclude that triangle (but then are either direct or involve other triangles, which is the $\widetilde{q}_L(\beta_L, \beta_T)$.

We can write the probability of some triangle forming as

$$\mathrm{E}_\beta(S_T(g)) = q_T = \beta_T + (1-\beta_T)(\widetilde{q}_L(\beta_L, \beta_T))^3, \tag{A.4}$$

where the first expression β_T is the probability that the triangle is directly generated, and then the second expression $(1-\beta_T)(\widetilde{q}_L(\beta_L, \beta_T))^3$ 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 $(\widetilde{q}_L(\beta_L, \beta_T))^3$). The result follows from Lemma A.1, with $x_1 = \beta_L$, $x_2 = \beta_T$, $q_L = a_1(x)$, $q_T = a_2(x)$ and $\widetilde{q}_L(\beta_L, \beta_T) = f(x)$. ■

LEMMA **A.1.** *Let $x = (x_1, x_2) \in [0,1)^2$ and $a(x) = (a_1(x), a_2(x))$ be two real-valued functions*

$$\begin{aligned} a_1(x) &= x_2 + (1-x_2) f(x) \\ a_2(x) &= x_2 + (1-x_2) f(x)^3, \end{aligned}$$

with

$$f(x) = x_1 + (1-x_1)\left[1 - (1-x_2)^N\right] = 1 - (1-x_1)(1-x_2)^N$$

for some integer $N \geq 0$. Then $x \neq x' \implies a(x) \neq a(x')$.

Proof. Suppose the contrary. Then

$$x_2' + (1-x_2') f(x') = x_2 + (1-x_2) f(x) \text{ and } x_2' + (1-x_2') f(x')^3 = x_2 + (1-x_2) f(x)^3.$$

First, note that if $x_2' = x_2$, then since these are both less than one, the first equation above implies that $f(x') = f(x)$. However, that is not possible since f is increasing in x_1 and $x_1' \neq x_1$ - recalling that $x' \neq x$ and $x_2' = x_2$. Thus, $x_2' \neq x_2$, and so without loss of generality consider the case in which $x_2' < x_2$. This implies that both

$$f(x') = bf(x) + c$$

and

$$f(x')^3 = bf(x)^3 + c,$$

where $b = \frac{1-x_2}{1-x_2'} \in (0,1)$ and $c = \frac{x_2 - x_2'}{1-x_2'} \in (0,1)$, and $b + c = 1$.

This implies that

$$bf(x)^3 + 1 - b = (bf(x) + 1 - b)^3.$$

This as an equation of the form

$$by^3 + 1 - b = (by + 1 - b)^3$$

where $b \in (0,1)$ and $y \in [0,1)$. Note that the left hand side is larger when $y = 0$ and the two are equal when $y = 1$, and that the derivative of the difference is

$$3by^2 - 3b(by + 1 - b)^2 = 3b\left[y^2 - (by + 1 - b)^2\right] < 0.$$

The difference is decreasing over the entire interval, and hits 0 at the end. Thus, the difference is always positive in $[0, 1)$ and there is no solution, meaning our supposition was incorrect. ■

LEMMA **A.2.** *Any event (in the discrete σ-algebra generated by all possible realizations of all subgraphs) associated with any SUGM has a probability that is an analytic function (and so it is in C^∞), and has derivatives and cross partials at all levels being uniformly continuous and bounded on the whole parameter space of $[0, 1]^k$.*

Proof. An 'outcome' is a specification of exactly which subgraphs form and which do not - so a complete specification of what happens. Any event then corresponds to a set of outcomes, and so its probability is a sum of probabilities of the outcomes. Each outcome's probability is of the form

$$\prod_{\ell} \beta_\ell^{z_\ell} (1 - \beta_\ell)^{m_\ell - z_\ell}$$

where z_ℓ indicates how many subgraphs of type ℓ are present in the outcome. As each of these functions is analytic (and hence in C^∞), all of the derivatives and partials, cross partials, etc., are continuous and bounded on $[0, 1]^k$ and hence uniformly continuous on $[0, 1]^k$. Any event is then a finite sum of analytic functions and so the result follows directly. ■

Proof of Theorem 2. We verify the conditions of Theorem 2.5 of Newey and McFadden (1994) for consistency. Assumption (i) holds by Theorem 1 and we assume compactness of the parameter space (Assumption (ii)). Continuity of $\log \mathrm{P}_\beta(g)$ at each β with probability one is mechanical in our model since subgraph probabilities are analytic functions of the parameters (Lemma A.2). Finally, the uniform bound of assumption (iv) holds since n is fixed, there are only a finite number of graphs in consideration, each with assigned probabilities in a compact set of parameters, and there is a positive probability of seeing any graph in $\mathcal{G}^n$. Therefore, the supremum must be finite.

We verify the conditions of Theorem 3.3 of Newey and McFadden (1994) for asymptotic normality. We have assumed (i), interiority of the parameter, and our model by construction places positive mass on all of $\mathcal{G}^n$. We have assumed (iv). Lemma A.2 implies Assumptions (ii), (iii), and (v). Because all events have probabilities that are analytic functions of parameters, with all derivatives and cross-partials being uniformly continuous and bounded in the parameter space, the norms of the maximal derivative ($\|\nabla_\beta \mathrm{P}_\beta(g)\|$) and second derivatives ($\|\nabla_{\beta\beta} \mathrm{P}_\beta(g)\|$) of the probability functions, as well as the log likelihood ($\|\nabla_\beta \log \mathrm{P}_\beta(g)\|$), have uniform and finite upper bounds. ■

Proof of Proposition 2. First we check consistency by the conditions of Theorem 2.6 of Newey and McFadden (1994). Here each observation is an independently drawn network. For Assumption (i) let $\widehat{W}$ be the identity matrix and then apply Proposition 1. For (ii), we have assumed that the parameter space $\mathcal{B}$ is compact. (iii) follows from the fact that $\mathrm{E}_\beta [S(g_r)]$ is continuous at each β with probability one since it composes continuous functions

of parameter entries. Finally (iv) follows from the fact that since both S_ℓ are shares, they are strictly less than 1.

Next we check asymptotic normality by the conditions of Theorem 3.4 of Newey and McFadden (1994).[45]

Since β_0 is in the interior of the compact parameter space, so (i) is met. We see (iii) holds since by definition the subgraph counts are fractions between 0 and 1. Both (ii) (that the empirical moment function is continuously differentiable in a neighborhood of the true parameter) and (iv) (that the gradient of the moment function is continuous at the true parameter and that it satisfies a ULLN) follow from Lemma A.2. Analytic functions are C^∞, so there are arbitrarily many derivatives. The bounds follow the expressions in (C.6) and following, which provide the rows of H, and the expressions are bounded in magnitude by $3(n-3)+1$ (and note the n is fixed and it is R that is growing). Finally, for (v), that HH' is non-singular follows from the linear independence of rows of H (see the expressions in (C.6) and following, which provide the rows of H, which dividing through by $(1-\beta_T)^{n-2}$ are clearly independent for interior parameters). ■

Proof of Theorem 3. When obvious, we omit superscript n's to simplify notation, but they are implicit. It follows that,

$$\breve{\beta}^{\mathrm{DC}}_{n,\ell} = \left(\frac{S^{true}_\ell}{\kappa_\ell \binom{n}{m_\ell}} + \frac{\widetilde{S}^{true}_\ell - S^{true}_\ell}{\kappa_\ell \binom{n}{m_\ell}} + \frac{\widetilde{S}_\ell(g) - \widetilde{S}^{true}_\ell}{\kappa_\ell \binom{n}{m_\ell}} \right) \tag{A.5}$$

where S^{true}_ℓ is the number of truly generated such subgraphs (unobserved) on the whole network, and $\widetilde{S}^{true}_\ell$ is the number of truly generated such subgraphs (unobserved) on the networks that the after removing the links in $D_\ell(g) = \{ij : ij \in g', g' \subset g, g' \in G_{\ell'}, \ell' < \ell\}$, and $\binom{n}{m_\ell}$ counts the number of ways to pick m_ℓ nodes out of n.

We show below that $|\widetilde{S}^{true}_\ell - S^{true}_\ell| = o_p(S^{true}_\ell)$ and $|\widetilde{S}_\ell(g) - \widetilde{S}^{true}_\ell| = o_p(\widetilde{S}^{true}_\ell)$; which then also implies that $\widetilde{S}_\ell(g) - \widetilde{S}^{true}_\ell = o_p(S^{true}_\ell)$. Together with (A.5), these tell us that

$$\breve{\beta}^{\mathrm{DC}}_{n,\ell} = \left(\frac{S^{true}_\ell}{\kappa_\ell \binom{n}{m_\ell}} \right) (1 + o_p(1)). \tag{A.6}$$

Note that $S^{true}_\ell(g)$ has a binomial distribution with parameter $\beta^n_{0,\ell}$. From this and (A.6), it then follows that

$$\frac{\breve{\beta}^{\mathrm{DC}}_{n,\ell} - \beta^n_{0,\ell}}{\sigma_{n,\ell}} \rightsquigarrow \mathcal{N}(0,1)$$

where $\sigma_{n,\ell} = \left(\frac{\beta^n_{0,\ell}(1-\beta^n_{0,\ell})}{\kappa_\ell \binom{n}{m_\ell}} \right)^{1/2}$. This is because $\frac{S^{true}_\ell}{\kappa_\ell \binom{n}{m_\ell}}$ is a self-normalized sum of $\kappa_\ell \binom{n}{m_\ell}$ Bernoulli($\beta^n_{0,\ell}$) independent random variables with standard deviation $\sigma_{n,\ell}$ by definition. Convergence in distribution in the sense of the Central Limit Theorem is a consequence of our

[45] See also Greene (2000) for the simplification of the variance term in the case in which there is just-identification as we have.

own Corollary 2, (though one could appeal to other CLTs on triangular arrays for independent random variables). So the result follows from Slutsky's Theorem since $1 + o_p(1) \xrightarrow{\text{P}} 1$.

Next, note that the $S_\ell^{true}(g)$ are independent across ℓ. From (A.6) it then follows that

$$\sum_\ell \alpha_\ell \breve{\beta}_{n,\ell}^{\text{DC}} = \sum_\ell \alpha_\ell \frac{S_\ell^{true}}{\kappa_\ell \binom{n}{m_\ell}} (1 + o_p(1))$$

for any $\alpha \in [0,1]^k$, with $\sum_\ell \alpha_\ell = 1$. Given the independence of S_ℓ^{true} across ℓ, it then follows that the random variable on the right hand side converges to being normal.[46] Then, by the Cramér-Wold Theorem, this implies that the $\breve{\beta}_{n,\ell}^{\text{DC}}$ are jointly normally distributed in the limit, and so

$$\Sigma_n^{-1/2}(\breve{\beta}_n^{\text{DC}} - \beta_0^n) \rightsquigarrow \mathcal{N}(0, I)$$

where $\Sigma_{n,\ell\ell} = \frac{\beta_{0,\ell}^n(1-\beta_{0,\ell}^n)}{\kappa_\ell \binom{n}{m_\ell}}$ and the off-diagonals are all 0.

Thus, to complete the proof we show that $|\widetilde{S}_\ell^{true} - S_\ell^{true}| = o_p(S_\ell^{true})$ and $|\widetilde{S}_\ell(g) - \widetilde{S}_\ell^{true}| = o_p(\widetilde{S}_\ell^{true})$.

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

Let z_ℓ^n denote the probability that any given $g' \in G_\ell^n$ is incidentally generated. We now show that $z_\ell^n / \beta_{0,\ell}^n = o(1)$, which establishes the first claim. Consider $g_\ell \in G_\ell^n$ and a (minimal, ordered) generating subclass $\mathcal{C} = (\ell_j, c_j)_{j \in J}$, and for which $\ell_j \geq \ell$ for all j.

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

Consider a subnetwork in $G_{\ell_j}^n$. The probability of getting at least one such network that has the c_j nodes out of the m_ℓ in g_ℓ is no more than

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

Then, we can bound the desired ratio by

$$\frac{z_\ell^n}{\beta_{0,\ell}^n} \leq \frac{\prod_{j \in J} n^{m_{\ell_j} - c_j} \kappa_{\ell_j} \beta_{0,\ell_j}^n}{\beta_{0,\ell}^n} \leq \frac{n^{\sum_{j \in J} m_{\ell_j} - h_{\ell_j} - c_j} \prod_{j \in J} \kappa_{\ell_j} b_{0,\ell_j}}{n^{-h_\ell} b_{0,\ell}} \to 0,$$

where the last convergence is guaranteed by (4.3).

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

[46] Note that under the assumption that $m_\ell > h_\ell$ there are a growing number of observations of each subgraph.

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 ℓ from this type of generating class, while the denominator is the expected number of the subgraphs of type ℓ.[47]

$$\frac{\kappa_\ell \binom{n}{m_\ell} \prod_{j\in J} n^{m_{\ell_j}-c_{\ell_j}} \kappa_{\ell_j} \beta_{0,\ell_j}^n}{\kappa_{\ell_{j'}} \binom{n}{m_{\ell_{j'}}} \beta_{0,\ell_{j'}}^n} = \Theta\left(\frac{n^{m_\ell} n^{\sum_{j\in J} m_{\ell_j}-c_{\ell_j}-h_{\ell_j}}}{n^{m_{\ell_{j'}}-h_{\ell_{j'}}}}\right) \to 0,$$

where the convergence to 0 follows from (4.4).

Finally, by multiplying and dividing by n^{h_ℓ} and collecting terms, it follows that $|\breve{b}_n^{\mathrm{DC}} - b_0| \xrightarrow{\mathrm{P}} 0$ and $V_n^{-1/2}\left(\breve{b}_n^{\mathrm{DC}} - b_0\right) \rightsquigarrow \mathcal{N}(0, I)$. To see this, observe that $\Sigma_n^{-1/2}(\breve{\beta}_n^{\mathrm{DC}} - \beta_0^n) = V_n^{-1/2}\left(\breve{b}_n^{\mathrm{DC}} - b_0\right)$. ■

Proof of Corollary 1. Note that $\sum_j c_j \geq m_\ell + (|C| - 1)z$ for some $z \geq 1$, where $z \geq 2$ if subgraphs are acyclic (each subgraph in the incidental set overlaps the others with at least one node, and at least two if the subgraphs are acyclic). The conditions then simplify directly. ■

Appendix B. Proof of Central Limit Theorem 4 and Corollary 2

B.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

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

for every $f(\cdot)$ that is 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 Q, denoted $d_K(Y, Q)$. We can bound this from above by (a constant times) the Wasserstein distance, $d_W(Y, Q)$, which itself is bounded by the below expression. Convergence in Wasserstein distance implies convergence in distribution. Let $\|f\|$ denote the sup norm over the domain of f.

Lemma **B.1** (Stein (1986); Ross (2011))**.** *If Y is a random variable and Q has the standard normal distribution, then*

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

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

Define

$$\mathbf{Z}^N := \sum_{\alpha\in\Lambda^N} Z_\alpha^N \text{ and } \overline{\mathbf{Z}}^N = \mathbf{Z}^N / a_N^{1/2}.$$

[47] We use Bachmann-Landau notation so $f(n) = \Theta(g(n))$ means that f is bounded above and below asymptotically by g. That is, $\exists k_1 > 0, \exists k_2 > 0, \exists n_0$ such that $\forall n > n_0$, $k_1 g(n) \leq f(n) \leq k_2 g(n)$.

For ease of notation, we omit the superscript Ns below. Recall that

$$\mathbf{Z}_{-\mathcal{A}(\alpha,N)} = \sum_{\eta \notin \mathcal{A}(\alpha,N)} Z_\eta$$

and let

$$\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)} := \mathbf{Z}_{-\mathcal{A}(\alpha,N)}/a^{1/2}.$$

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

$$\sup_{\{f:||f||,||f''||\leq 2,||f'||\leq\sqrt{2/\pi}\}} \left|\mathrm{E}[f'(\overline{\mathbf{Z}}^N) - \overline{\mathbf{Z}}^N f(\overline{\mathbf{Z}}^N)]\right| \to 0,$$

then $d_W(\overline{\mathbf{Z}}^N, Q) \to 0$, and so it must be asymptotically normally distributed.

B.2. **Proofs of Theorem 4 and Corollary 2.** The following lemmas are useful in the proof.

LEMMA **B.2.** *A solution to* $\max_h \mathrm{E}[Zh(Y)]$ *s.t.* $|h(\cdot)| \leq 1$ *(so the absolute value of* h *is bounded by 1, where* h *is measurable) is* $h(Y) = \mathrm{sign}(\mathrm{E}[Z|Y])$*, where we break ties, setting* $\mathrm{sign}(\mathrm{E}[Z|Y]) = 1$ *when* $\mathrm{E}[Z|Y] = 0$*.*

Proof. This can be seen from direct calculation:

$$\mathrm{E}[Zh(Y)] = \int_Y \mathrm{E}[Z|Y]h(Y)d\mathrm{P}(Y)$$

Maximizing $\mathrm{E}[Z|Y]h(Y)$ pointwise when $|h| \leq 1$ is achieved by setting $h(Y) = \mathrm{sign}(\mathrm{E}[Z|Y])$, and we break ties by setting $\mathrm{sign}(\mathrm{E}[Z|Y]) = 1$ when $\mathrm{E}[Z|Y] = 0$, as that makes no difference in the integral. ∎

LEMMA **B.3.** $\mathrm{E}[XYh(Y)]$ *when* $h(\cdot)$ *is measurable and bounded by* $\sqrt{\frac{2}{\pi}}$ *satisfies*

$$\mathrm{E}[XYh(Y)] \leq \sqrt{\frac{2}{\pi}}\mathrm{E}\left[XY \cdot \mathrm{sign}(\mathrm{E}[X|Y]Y)\right].$$

Proof. This follows from Lemma B.2, setting $Z = XY$. ∎

Proof of Theorem 4. By Lemma B.1, it is sufficient to show that the appropriate sequence of random variables $\overline{\mathbf{Z}}^N$ satisfies

$$\sup_{\{f:||f||,||f''||\leq 2,||f'||\leq\sqrt{2/\pi}\}} \left|\mathrm{E}[f'(\overline{\mathbf{Z}}^N) - \overline{\mathbf{Z}}^N f(\overline{\mathbf{Z}}^N)]\right| \to 0.$$

Observe that

$$\begin{aligned}
\mathrm{E}\left[\overline{\mathbf{Z}} f\left(\overline{\mathbf{Z}}\right)\right] &= \mathrm{E}\left[\frac{1}{a^{1/2}}\sum_\alpha Z_\alpha \cdot f\left(\overline{\mathbf{Z}}\right)\right] \\
&= \mathrm{E}\left[\frac{1}{a^{1/2}}\sum_\alpha Z_\alpha \left(f\left(\overline{\mathbf{Z}}\right) - f\left(\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}\right)\right)\right] \\
&\quad + \mathrm{E}\left[\frac{1}{a^{1/2}}\sum_\alpha Z_\alpha \cdot f\left(\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}\right)\right].
\end{aligned}$$

The first step is to show that

$$\left|\mathrm{E}\left[\frac{1}{a^{1/2}}\sum_{\alpha} Z_\alpha \cdot f\left(\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}\right)\right]\right| = o(1),$$

by employing condition (6.3).

In order to do this, we can expand the term to

$$\begin{aligned}\left|\mathrm{E}\left[\frac{1}{a_N^{1/2}}\sum_{\alpha\in\Lambda} Z_\alpha \cdot f\left(\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}\right)\right]\right| &= \left|\mathrm{E}\left[\frac{1}{a_N^{1/2}}\sum_{\alpha\in\Lambda} Z_\alpha \cdot f\left(\frac{1}{a_N^{1/2}}\sum_{\eta\notin\mathcal{A}(\alpha,N)} Z_\eta\right)\right]\right| \\ &\leq \underbrace{\left|\mathrm{E}\left[\frac{1}{a_N^{1/2}}\sum_{\alpha\in\Lambda} Z_\alpha \cdot f\left(0\right)\right]\right|}_{=0 \text{ since } \mathrm{E}[Z_\alpha]=0.} \\ &\quad + \left|\mathrm{E}\left[\frac{1}{a_N^{1/2}}\sum_{\alpha\in\Lambda} Z_\alpha \cdot \left(\frac{1}{a_N^{1/2}}\sum_{\eta\notin\mathcal{A}(\alpha,N)} Z_\eta\right)\cdot f'\left(\widehat{\overline{\mathbf{Z}}}_{-\mathcal{A}(\alpha,N)}\right)\right]\right|\end{aligned}$$

where $\widehat{\overline{\mathbf{Z}}}_{-\mathcal{A}(\alpha,N)}$ is an intermediate value between $\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}$ and 0.

To bound the second term, we apply Lemma B.3 to conclude that

$$\left|\frac{\mathrm{E}\left[\sum_{\alpha\in\Lambda;\eta\notin\mathcal{A}(\alpha,N)} Z_\alpha Z_\eta f'\left(\widehat{\overline{\mathbf{Z}}}_{-\mathcal{A}(\alpha,N)}\right)\right]}{a_N}\right| \leq \sqrt{\frac{2}{\pi}}\left|\frac{\mathrm{E}\left[\sum_{\alpha\in\Lambda;\eta\notin\mathcal{A}(\alpha,N)} Z_\alpha Z_\eta \cdot \mathrm{sign}\left(\mathrm{E}\left[Z_\alpha Z_\eta|\ Z_\eta\right]\right)\right]}{a_N}\right|.$$

Thus, it is sufficient that

$$\mathrm{E}\left[\sum_{\alpha\in\Lambda;\eta\notin\mathcal{A}(\alpha,N)} Z_\alpha Z_\eta \cdot \mathrm{sign}\left(\mathrm{E}\left[Z_\alpha Z_\eta|Z_\eta\right]\right)\right] = o(a_N) \tag{B.1}$$

or

$$\mathrm{E}\left[\sum_{\alpha\in\Lambda;\eta\notin\mathcal{A}(\alpha,N)} \left|\mathrm{E}\left[Z_\alpha Z_\eta|Z_\eta\right]\right|\right] = o(a_N)$$

to ensure that

$$\left|\frac{\mathrm{E}\left[\sum_{\alpha\in\Lambda;\eta\notin\mathcal{A}(\alpha,N)} Z_\alpha \cdot Z_\eta \cdot f'\left(\widehat{\overline{\mathbf{Z}}}_{-\mathcal{A}(\alpha,N)}\right)\right]}{a_N}\right| = o(1),$$

which is ensured by (6.3) (noting that $\widehat{\overline{\mathbf{Z}}}_{-\mathcal{A}(\alpha,N)}$ is a function of $\mathbf{Z}_{-\mathcal{A}(\alpha,N))}$).

Next, the second step of the proof is to apply a similar reasoning as in Ross (2011) with an $o(1)$ adjustment (from the first step above), to write

$$\begin{aligned}\left|\mathrm{E}\left[f'(\overline{\mathbf{Z}}) - \overline{\mathbf{Z}}f(\overline{\mathbf{Z}})\right]\right| &\leq \left|\mathrm{E}\left[\frac{1}{a^{1/2}}\sum_{\alpha} Z_\alpha(f(\overline{\mathbf{Z}}) - f(\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}) - (\overline{\mathbf{Z}} - \overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)})f'(\overline{\mathbf{Z}})\right]\right| \\ &\quad + \left|\mathrm{E}\left[f'(\overline{\mathbf{Z}})\left(1 - \frac{1}{a^{1/2}}\sum_{\alpha} Z_\alpha(\overline{\mathbf{Z}} - \overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)})\right)\right]\right| + o(1),\end{aligned}$$

and then to show that the right hand side of this expression goes to 0.

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

$$\left|\mathrm{E}\left[f'(\overline{\mathbf{Z}})-\overline{\mathbf{Z}}f(\overline{\mathbf{Z}})\right]\right| \leq \frac{||f''||}{2a^{1/2}}\sum_{\alpha}\mathrm{E}\left[|Z_\alpha|\left(\overline{\mathbf{Z}}-\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)}\right)^2\right] + \left|\mathrm{E}\left[f'(\overline{\mathbf{Z}})\left(1-\frac{1}{a^{1/2}}\sum_{\alpha}Z_\alpha(\overline{\mathbf{Z}}-\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)})\right)\right]\right| + o(1).$$

Let us denote the first two terms on the right hand side as A_1 and A_2 respectively. We bound each, and show that each is $o(1)$, which then completes the proof.

$$A_1 = \frac{||f''||}{2a^{3/2}}\sum_{\alpha}\mathrm{E}\left[|Z_\alpha|\left(\sum_{\eta\in\mathcal{A}(\alpha,N)}Z_\eta\right)^2\right] = \frac{||f''||}{2a^{3/2}}\sum_{\alpha;\eta\in\mathcal{A}(\alpha,N),\gamma\in\mathcal{A}(\alpha,N)}\mathrm{E}\left[|Z_\alpha|Z_\eta Z_\gamma\right] = o(1),$$

where the last equality follows from (6.1).

Next,

$$\begin{aligned}A_2 &= \left|\mathrm{E}\left[f'(\overline{\mathbf{Z}})\left(1-\frac{1}{a^{1/2}}\sum_{\alpha}Z_\alpha(\overline{\mathbf{Z}}-\overline{\mathbf{Z}}_{-\mathcal{A}(\alpha,N)})\right)\right]\right| = \frac{1}{a}\left|\mathrm{E}\left[f'(\overline{\mathbf{Z}})\left(a-\sum_{\alpha,\eta\in\mathcal{A}(\alpha,N)}Z_\alpha Z_\eta\right)\right]\right| \\ &\leq \frac{||f'||}{a}\mathrm{E}\left|\left(a-\sum_{\alpha,\eta\in\mathcal{A}(\alpha,N)}Z_\alpha Z_\eta\right)\right| = \frac{||f'||}{a}\mathrm{E}\left|\left(\sum_{\alpha,\eta\in\mathcal{A}(\alpha,N)}Z_\alpha Z_\eta-\mathrm{E}\left[Z_\alpha Z_\eta\right]\right)\right| \\ &\leq \frac{\sqrt{2}}{a\sqrt{\pi}}\left(\mathrm{var}\left[\sum_{\alpha,\eta\in\mathcal{A}(\alpha,N)}Z_\alpha Z_\eta\right]\right)^{1/2} = \frac{\sqrt{2}}{a\sqrt{\pi}}\left(\sum_{\alpha,\alpha',\eta\in\mathcal{A}(\alpha,N),\eta'\in\mathcal{A}(\alpha',N)}\mathrm{cov}\left(Z_\alpha Z_\eta, Z_{\alpha'}Z_{\eta'}\right)\right)^{1/2},\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz. The final expression is $o(1)$ by (6.2). ■

Proof of Corollary 2. We apply Theorem 4 to the case in which $\mathcal{A}(\alpha,N)=\{\alpha\}$. (6.1) becomes

$$\sum_{\alpha}\mathrm{E}\left[|Z_\alpha|^3\right] = o\left(\left(\sum_{\alpha}\mathrm{var}\left(Z_\alpha\right)\right)^{3/2}\right)$$

which becomes[48]

$$\sum_{\alpha}\mathrm{var}\left(Z_\alpha\right)^{3/2} = o\left(\left(\sum_{\alpha}\mathrm{var}\left(Z_\alpha\right)\right)^{3/2}\right),$$

which is satisfied directly, given that $\sum_\alpha \mathrm{var}\left(Z_\alpha\right)$ is growing without bound.

(i) and (ii) imply (6.2) and (6.3), respectively, (noting that the sign is always nonnegative by the supposition of the corollary).

We now show that for Bernoulli random variables with uniformly vanishing means, (i) holds whenever (ii) holds. Observe that

$$\begin{aligned}\mathrm{cov}\left(Z_\alpha^2, Z_\eta^2\right) &= \mathrm{cov}\left(\left(X_\alpha-\mu_\alpha\right)^2,\left(X_\eta-\mu_\eta\right)^2\right) \\ &= \mathrm{cov}\left(X_\alpha^2-2X_\alpha\mu_\alpha+\mu_\alpha^2, X_\eta^2-2X_\eta\mu_\eta+\mu_\eta^2\right)\end{aligned}$$

[48]Recall that it is assumed that $\mathrm{E}\left[|Z_\alpha|^3\right]/\mathrm{E}\left[Z_\alpha^2\right]^{3/2}$ is bounded above (and necessarily below via Jensen's Inequality).

$$= \text{cov}\left(X_\alpha^2, X_\eta^2\right) - 2\mu_\alpha \text{cov}\left(X_\alpha, X_\eta^2\right) - 2\mu_\eta \text{cov}\left(X_\alpha^2, X_\eta\right) + 4\mu_\alpha\mu_\eta \text{cov}\left(X_\alpha, X_\eta\right).$$

Because they are Bernoulli, $\text{cov}\left(X_\alpha^k, X_\eta^{k'}\right) = \text{cov}\left(X_\alpha, X_\eta\right)$ for any $k, k' > 0$. Since the means tend to zero, this means

$$\text{cov}\left(Z_\alpha^2, Z_\eta^2\right) = \text{cov}\left(X_\alpha, X_\eta\right)\left(1 + o\left(1\right)\right).$$

Therefore satisfying (ii) implies (i) (noting also that $a_N \geq 1$ so $a_N^2 \geq a_N$). ■