For Online Publication Only

Appendix B. Asymptotic Normality with Links and Triangles

Proposition B.1. *Consider a links and triangles SUGM with associated parameters* $\beta^n_{0,L}, \beta^n_{0,T} = \left(\frac{b_{0,L}}{n^{h_L}}, \frac{b_{0,T}}{n^{h_T}}\right)$ *such that*[40]

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

or

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

Then the model satisfies the conditions of Theorem 1.

These conditions cover a range of sparse and dense graphs.

Proof of Proposition B.1.

Our proof uses Theorem 1 directly to establish that the conditions hold for the statistic $S_L(g)$, and then we use Corollary 1 to establish that the conditions hold for the statistic $S_T(g)$.

S_L case:

In this case, α refers to a generic link ij.

We first apply the theorem letting $\Delta\left(\alpha, \binom{n}{2}\right) = \{\eta : \eta \cap \alpha \neq \emptyset\}$, so $\Delta(ij, N) = \{ij\} \cup \{ik | k \neq i, j\} \cup \{jk | k \neq i, j\}$. After that we will apply the corollary. Each covers different parameter intervals.

Applying Theorem 1: So, for the application of the theorem, Δ includes the link ij itself, as well as adjacent links, ik and jk, for $k \neq i, j$. We show that the conditions for Theorem 1 hold.

(5.3) is obvious from the definition of $\Delta\left(ij, \binom{n}{2}\right)$, because if ij and kl do not share nodes then no triangle (nor any link) could generate both, and so they are independent and the left hand side term is then 0.

Next we verify (5.1). Consider the product

$$\mathrm{E}\,|X_\alpha X_\eta X_\gamma|$$

[40]The constants, $b_{0,L}, b_{0,T}$, can be allowed to vary with n within some compact set without consequence for the result or its proof.

for $\eta, \gamma \in \Delta(\alpha, N)$. There are several cases to consider. The first three are cases in which α, η, γ are distinct and the last two handle cases in which two or more of these are the same link. In each case the letters i, j, k, l are distinct generic nodes, and α could be any one of the links

[1]: ij, jk, il (a line) - there are order n^4 of these.
[2]: ij, ik, il (a star) - there are order n^4 of these.
[3]: ij, jk, ik (a triangle) - there are order n^3 of these.
[4]: ij, ij, jk or ij, jk, jk, two of the links repeat - there are order n^3 of these.
[5]: ij, ij, ij all of the links repeat - there are order n^2 of these.

Note that

$$\mathrm{E}\left|X_\alpha X_\eta X_\gamma\right| = \mathrm{P}\left(X_\alpha X_\eta X_\gamma = 1\right),$$

which we now bound in each case.

From the proof of Proposition 1, recall that q_L^n is the probability of a link forming in the graph which can be due to a link forming directly or as a part of a triangle, and $\widetilde{q}_L^n$ is the probability of a link forming if a particular triangle that it could be part of does not form. Let $\widetilde{q}'_L$ denote the probability that a link forms conditional on two triangles that it could be part of not forming.

It is useful to suppress the n subscripts unless explicitly needed. Then loose upper bounds on the probabilities of the various structures are

- for [1]: $\beta_{0,T}^2 + 2(1-\beta_{0,T})\beta_{0,T}\widetilde{q}_L + (1-\beta_{0,T})^2\widetilde{q}_L^2\widetilde{q}'_L \leq \beta_{0,T}^2 + 2\beta_{0,T}q_L + q_L^3$,
- for [2]: $\beta_{0,T}^3 + 3(1-\beta_{0,T})\beta_{0,T}^2 + 3(1-\beta_{0,T})^2\beta_{0,T}\widetilde{q}'_L + (1-\beta_{0,T})^3(\widetilde{q}'_L)^3 \leq 4\beta_{0,T}^2 + 3\beta_{0,T}q_L + q_L^3$,
- for [3]: $\beta_{0,T} + (1-\beta_{0,T})(\widetilde{q}'_L)^3 \leq \beta_{0,T} + q_L^3$,
- for [4]: $\beta_{0,T} + (1-\beta_{0,T})(\widetilde{q}'_L)^2 \leq \beta_{0,T} + q_L^2$
- for [5]: q_L

Thus, from the numbers and bounds on probabilities, it follows that

$$\mathrm{E}\left|X_\alpha X_\eta X_\gamma\right| \leq \Theta(n^4(\beta_{0,T}^2 + \beta_{0,T}q_L + q_L^3) + n^3(\beta_{0,T} + q_L^2) + n^2 q_L). \tag{B.1}$$

Next, note that straightforward calculations show that for $k \neq i$

$$\mathrm{cov}(X_{ij}, X_{jk}) = \beta_{0,T}(1-\beta_{0,T})(1-\widetilde{q}_L)^2 \approx \beta_{0,T}. \tag{B.2}$$

It then follows directly that

$$a_N = \Theta\left(n^2 q_L + n^3\beta_{0,T}\right).$$

Now we can compare our expression for $\mathrm{E}\left|X_\alpha X_\eta X_\gamma\right|$ from (B.1) to

$$o(a_N^{3/2}) = o\left(n^3 q_L^{3/2} + n^{9/2}\beta_{0,T}^{3/2}\right),$$

to check the condition.

This imposes a number of constraints (omitting the ones that are obviously satisfied, such as $n^4\beta_{0,T}^2 + n^3 q_L^2 = o(n^3 q_L^{3/2} + n^{9/2}\beta_{0,T}^{3/2})$):

$$n^4\beta_{0,T}q_L + n^4 q_L^3 + n^3\beta_{0,T} + n^2 q_L = o(n^3 q_L^{3/2} + n^{9/2}\beta_{0,T}^{3/2}), \tag{B.3}$$

Noting, that as in the proof of Proposition 1, (working there with $\tilde{q}_L^n$, which is of the same order)

$$q_L^n = \Theta\left(\frac{1}{n^{h_L}} + \frac{1}{n^{h_T-1}}\right).$$

it follows that

$$o\left(n^3 q_L^{3/2} + n^{9/2}\beta_{0,T}^{3/2}\right) = o\left(n^{3-(3/2)h_L} + n^{9/2-(3/2)h_T}\right).$$

Thus, (B.3) becomes

$$\max[4 - h_T - \min[h_L, h_T - 1], 4 - 3\min[h_L, h_T - 1], 3 - h_T, 2 - \min[h_L, h_T - 1]]$$
$$< \max[3 - (3/2)h_L, 9/2 - (3/2)h_T],$$

We break this into two cases: If $h_L \geq h_T - 1$ then this is satisfied whenever $h_T \in (5/3, 3)$. If $h_L < h_T - 1$ then this is satisfied whenever $h_L \in (2/3, 2)$ and $h_T > \max[(3/2)h_L, h_L + 1]$.

Next, we turn to (5.2). We compute terms of the form

$$\operatorname{cov}\left(\left(g_{ij} - q_L\right)\left(g_{jk} - q_L\right), \left(g_{rs} - q_L\right)\left(g_{st} - q_L\right)\right).$$

since $\eta \in \Delta(\alpha, N)$ and $\eta' \in \Delta(\alpha', N)$, where here we allow for the cases that $k = i$ and $r = t$. Iterating on the expectations, one can show that

$$\operatorname{cov}\left(\left(g_{ij} - q_L\right)\left(g_{jk} - q_L\right), \left(g_{rs} - q_L\right)\left(g_{st} - q_L\right)\right) \leq \mathrm{E}\left(g_{ij}g_{jk}g_{rs}g_{st}\right).$$

It is easy to see that if $\{i, j, k\} \cap \{r, s, t\} = \emptyset$, then the covariance is zero since the events are independent. Thus, we are summing over cases in which the intersection is nonempty. The cases with intersection of two or more nodes are handled as we already did above, noting that the condition is less restrictive here (note that $a_N > 1$, so $a_N^2 > a_N^{3/2}$).

So, we restrict attention to the cases in which there is only one node of intersection. In this case the intersection could come from (1) $s = j$, so we are looking at two two-stars that are joined at the center, (2) $i = r$ so we are looking at a line, or (3) $s = i$, where the center of one star is attached to the leaf of the other. These exhaust all configurations up to a relabeling. Consider the event that $g_{ij}g_{jk}g_{rs}g_{st} = 1$ and say we are in case 1. This has the highest probability relative to the other two cases, so we can construct a crude bound using this. This probability is of order no more than:

$$\beta_{0,T}^3 + \beta_{0,T}^2 q_L + \beta_{0,T} q_L^2 + q_L^3$$

So, we need to check that

$$n^5(\beta_{0,T}^3 + \beta_{0,T}^2 q_L + \beta_{0,T} q_L^2 + q_L^3) = o\left(n^4 q_L^2 + n^6\beta_{0,T}^2\right)$$

It is sufficient that

$$n\beta_{0,T}^2 q_L = o\left(q_L^2 + n^2\beta_{0,T}^2\right), \quad n\beta_{0,T}q_L^2 = o\left(q_L^2 + n^2\beta_{0,T}^2\right) \quad nq_L^3 = o\left(q_L^2 + n^2\beta_{0,T}^2\right)$$

These conditions become:

$$\max[1-2h_T-\min[h_L, h_T-1], 1-h_T-2\min[h_L, h_T-1], 1-3h_L, 4-3h_T] < \max[-2h_L, 2-2h_T]$$

We break this into two cases: If $h_L \geq h_T - 1$ then this is satisfied whenever

$$h_L \geq h_T - 1, \qquad h_T \in (2, 1/2 + (3/2)h_L).$$

If $h_L < h_T - 1$ then this is satisfied whenever

$$h_L > 1, \qquad h_T > \max[(2/3)h_L + 4/3, h_L + 1].$$

Summarizing all of the conditions together (noting some redundancies), it is sufficient that either

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

or

$$h_L > 1 \quad \text{and} \quad h_T > \max[(3/2)h_L, h_L + 1, (2/3)h_L + 4/3].$$

Applying Corollary 1: Next, let us analyze the conditions imposed by the corollary.

Condition (i) is straightforward, since $\text{var}(X_\alpha) = \mu(1-\mu)$ for a Bernoulli random variable.

Condition (iii) is verified by checking that (noting our expression for $\text{cov}(X_{ij}, X_{jk})$ above and that $\text{var}(X_\alpha) = \mu(1-\mu) \approx q_L$):

$$\sum_{\alpha \neq \eta} \text{cov}(X_\alpha, X_\eta) \approx n^3\beta_T = o(n^2 q_L)$$

which is satisfied whenever $h_L < h_T - 1$.

It is also easy to check that (ii) holds whenever (iii) does. In particular, note that

$$\sum_{\alpha \neq \eta} \text{cov}((X_\alpha - \mu)^2, (X_\eta - \mu)^2) = \text{E}[(X_\alpha^2 - 2X_\alpha\mu + \mu^2)(X_\eta^2 - 2X_\eta\mu + \mu^2)]$$

and, since for a Bernoulli random variable $X_\alpha^2 = X_\alpha$, this becomes

$$\sum_{\alpha \neq \eta} \text{cov}((X_\alpha - \mu)^2, (X_\eta - \mu)^2) = \text{E}[(X_\alpha(1-2\mu) + \mu^2)(X_\eta(1-2\mu) + \mu^2)]$$

$$= \text{E}[X_\alpha X_\eta - \mu^2] + O(\mu^3) \approx \text{cov}(X_\alpha, X_\eta).$$

Thus, (ii) is satisfied whenever (iii) is (and note that $a_N \geq 1$ so $a_N^2 \geq a_N$).

Therefore, again, re-summarizing all of the conditions together, it is sufficient that either

$$h_L > 0 \quad \text{and} \quad h_T > h_L + 1$$

or

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

S_T case: Set $\Delta\left(ijk, \binom{n}{2}\right) = \{ijk\}$ and we apply Corollary 1.

Condition (i) follows from the fact that since $\mu \leq 1/2$ for large enough N, $\text{var}(X_\alpha) = \mu(1-\mu) \geq \mu^2 \geq \mu^2 N^{-1/3+\varepsilon}$.

Condition (ii) is implied by condition (iii), below, just as argued above.

So we check condition (iii):

$$\sum_{\alpha \neq \eta} \text{cov}\left(X_\alpha, X_\eta\right) = o\left(N \cdot \text{var}\left(X_\alpha\right)\right).$$

To see this first observe that (again, noting that X_α is Bernoulli):

$$\text{var}\left(X_\alpha\right) = q_T(1-q_T) = \beta_{0,T}(1+o(1)),$$

which follows from the proof of Proposition 1.

We compute $\text{cov}\left(X_\alpha, X_\eta\right)$ for various cases of α, η as a function of how many nodes the two triangles have in their intersection:

- $|\alpha \cap \eta| = 0$: $\text{cov}\left(X_\alpha, X_\eta\right) = 0$ by independence.
- $|\alpha \cap \eta| = 1$: $\text{cov}\left(X_\alpha, X_\eta\right) = O\left(\beta_{0,T} \cdot q_L^4\right)$. There are $O\left(n^5\right)$ of these.
  - This is because we need at least one link from each triangle to have formed together and not have already formed independently, which can happen only if the joint node is part of a triangle, and neither of the triangles formed directly. This gives us $\beta_{0,T}\widetilde{q}_L^4 \leq \beta_{0,T} q_L^4$.
- $|\alpha \cap \eta| = 2$: $\text{cov}\left(X_\alpha, X_\eta\right) = O\left(\beta_{0,T} q_L^2 + q_L^5\right)$.
  - This is because we need the common link from each triangle to have formed together and not have already formed independently in both cases, which can happen only if exactly one of the triangles formed directly and the other did not, or else neither triangle to have formed and all of the links to have formed. This is of order $\beta_{0,T}\widetilde{q}_L^2 + (\widetilde{q}_L')^5 \leq \beta_{0,T} q_L^2 + q_L^5$.

  There are $O\left(n^4\right)$ of these.

Thus, it must be that

$$n^5 \beta_{0,T} q_L^4 + n^4(\beta_{0,T} q_L^2 + q_L^5) = o\left(n^3 \beta_{0,T}\right).$$

These conditions become $h_L > 1/2$ and $h_T < 5h_L - 1$.

If we put all of the conditions together from both the links and the triangles, we end up with

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

or

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

which completes the proof. ■

Appendix C. Identification in an Example with Links and K-stars

Note that K-stars are somewhat thorny as adding one link from the center of a K-star to another node now results in K incidental K-stars. Nonetheless, we can still identify them from counts of simple statistics.

It is useful to work with a parameter

$$\overline{\beta}^n_{0,K} = 1 - (1 - \beta^n_{0,K})^{2\binom{n-2}{K-1}}$$

which is the probability that a link ij forms incidentally via on of $2\binom{n-2}{K-1}$ possible K-stars of which ij could potentially be a part.

Again, we say that a sequence of SUGMs is *well-balanced* if the probability that a link is formed directly or formed as part of some K-star are of the same order: $\overline{\beta}^n_{0,K} = \Theta(\beta^n_{0,L})$.

Proposition **C.1.**

- *A SUGM based on links and K-stars is identified by $S_L, \frac{S_d}{S_{d-1}}$, where S_d is the fraction of nodes that have degree d, for any choice of $0 < d < K$. That is, if $\beta'_L, \beta'_K \neq \beta_L, \beta_K$ then*

$$\left(\mathrm{E}_{\beta'_L,\beta'_K}[S_L(g)], \frac{\mathrm{E}_{\beta'_L,\beta'_K}[S_d(g)]}{\mathrm{E}_{\beta'_L,\beta'_K}[S_{d-1}(g)]}\right) \neq \left(\mathrm{E}_{\beta_L,\beta_K}[S_L(g)], \frac{\mathrm{E}_{\beta_L,\beta_K}[S_d(g)]}{\mathrm{E}_{\beta_L,\beta_K}[S_{d-1}(g)]}\right).$$

- *Moreover, if a sequence of SUGMs is well-balanced and $\beta^n_{0,L}$ and $\overline{\beta}^n_{0,K}$ are bounded away from 1, then $\beta^n_{0,L}, \overline{\beta}^n_{0,K}$ are identifiably unique.*[41]

In Proposition C.1 the identification is based on a comparison between the relative frequency of two degrees: d and $d-1$, for any $0 < d < K$. For many choices of β, identification can be achieved with just $\mathrm{E}_{\beta_L,\beta_K}[S_d(g)]$ rather than $\frac{\mathrm{E}_{\beta_L,\beta_K}[S_d(g)]}{\mathrm{E}_{\beta_L,\beta_K}[S_{d-1}(g)]}$, but to get identification across all parameter values requires comparison of two degrees.

Proof of Proposition C.1.

[41] Here the r^n_L, r^n_K are both set to be the order of $\beta^n_{0,L}$, and the identifiable uniqueness is achieved via $\left(\frac{\mathrm{E}_{\beta_L,\beta_K}[S_L(g)]}{r^n}, \frac{\mathrm{E}_{\beta_L,\beta_K}[S_d(g)]}{r^n \mathrm{E}_{\beta_L,\beta_K}[S_{d-1}(g)]}\right)$.

First, note that

$$q_L := \mathrm{E}_{\beta_L,\beta_K}\left[S_L(g)\right] = \beta_L + (1-\beta_L)\left(1-(1-\beta_K)^{2\binom{n-2}{K-1}}\right), \tag{C.1}$$

where the $(1-\beta_K)^{2\binom{n-2}{K-1}}$ represents the probability that none of the $2\binom{n-2}{K-1}$ possible K-stars, of which ij could potentially be a part, form.

The expected fraction of nodes having degree $d < K$ is

$$q_d := \mathrm{E}_{\beta_L,\beta_K}\left[S_d(g)\right] = \binom{n-1}{d}\pi^d(1-\pi)^{n-1-d}(1-\beta_K)^{\binom{n-1}{K}}, \tag{C.2}$$

where

$$\pi = \beta_L + (1-\beta_L)\left(1-(1-\beta_K)^{\binom{n-2}{K-1}}\right)$$

is the probability that some link ij forms without it being part of a K-star centered at i.

We show the first part of proposition by showing that if

$$(q_L, q_d/q_{d-1}) = \left(\mathrm{E}_{\beta'_L,\beta'_K}\left[S_L(g)\right], \frac{\mathrm{E}_{\beta'_L,\beta'_K}\left[S_d(g)\right]}{\mathrm{E}_{\beta'_L,\beta'_K}\left[S_{d-1}(g)\right]}\right) = \left(\mathrm{E}_{\beta_L,\beta_K}\left[S_L(g)\right], \frac{\mathrm{E}_{\beta_L,\beta_K}\left[S_d(g)\right]}{\mathrm{E}_{\beta_L,\beta_K}\left[S_{d-1}(g)\right]}\right) \tag{C.3}$$

then $\beta_L, \beta_K = \beta'_L, \beta'_K$.

If $\beta_L, \beta_K = (0,0)$ then $q_L = 0$ which implies directly from (C.1) that $\beta'_L, \beta'_K = (0,0)$, and so that case is straightforward. So, we concentrate on the case in which $\beta_L, \beta_K \neq (0,0)$, which then implies that $q_L > 0$. Similarly, we consider the case in which $\beta_L < 1, \beta_K < 1$ (as otherwise the network is complete), which implies that $q_L < 1$ and then that $\beta'_L < 1, \beta'_K < 1$, which then also implies that $0 < q_d < 1$.

By (C.1):

$$q_L = \beta_L + (1-\beta_L)\left(1-(1-\beta_K)^{2\binom{n-2}{K-1}}\right).$$

which implies that

$$(1-\beta_K)^{2\binom{n-2}{K-1}} = \left(\frac{1-q_L}{1-\beta_L}\right)^{1/2}.$$

This implies that

$$q_d = \binom{n-1}{d}\pi(\beta_L)^d(1-\pi(\beta_L))^{n-1-d}\left(\frac{1-q_L}{1-\beta_L}\right)^{(n-1)/(2K)},$$

where

$$\pi(\beta_L) = 1-(1-\beta_L)^{1/2}(1-q_L)^{1/2}.$$

This implies that

$$q_d/q_{d-1} = \pi(\beta_L)/(1-\pi(\beta_L)). \tag{C.4}$$

Thus, repeating the argument for β' and then employing (C.3):

$$\frac{1-(1-\beta_L)^{1/2}(1-q_L)^{1/2}}{(1-\beta_L)^{1/2}(1-q_L)^{1/2}} = \frac{1-(1-\beta'_L)^{1/2}(1-q_L)^{1/2}}{(1-\beta'_L)^{1/2}(1-q_L)^{1/2}}.$$

Thus

$$\frac{1}{(1-\beta_L)^{1/2}} = \frac{1}{(1-\beta'_L)^{1/2}},$$

which implies that $\beta_L = \beta'_L$. This then implies that $\beta_K = \beta'_K$ via (C.1), and so we have established the result.

To establish identifiable uniqueness we argue that for any $\varepsilon > 0$ there exists $\phi > 0$ such that for large enough n, if $\delta((\beta_L^n, \overline{\beta}_K^n), (\beta_{0,L}^n, \overline{\beta}_{0,K}^n)) > \varepsilon$, then at least one of the following inequalities holds:

$$\left|\frac{\mathrm{E}_{\beta^n}[S_L(g)] - \mathrm{E}_{\beta_0^n}[S_L(g)]}{r^n}\right| > \phi$$

or

$$\left|\frac{\mathrm{E}_{\beta^n}[S_d(g)]}{r^n \mathrm{E}_{\beta^n}[S_{d-1}(g)]} - \frac{\mathrm{E}_{\beta_0^n}[S_d(g)]}{r^n E_{\beta_0^n}[S_{d-1}(g)]}\right| > \phi.$$

So, suppose the contrary: there exists a subsequence of $(\beta_L^n, \overline{\beta}_K^n), (\beta_{0,L}^n, \overline{\beta}_{0,K}^n)$ and $\phi^n \to 0$ such that for every n $\delta((\beta_L^n, \overline{\beta}_K^n), (\beta_{0,L}^n, \overline{\beta}_{0,K}^n)) > \varepsilon$ and yet

$$\left|\frac{\mathrm{E}_{\beta^n}[S_L(g)] - \mathrm{E}_{\beta_0^n}[S_L(g)]}{r^n}\right| \leq \phi^n$$

and

$$\left|\frac{\mathrm{E}_{\beta^n}[S_d(g)]}{r^n \mathrm{E}_{\beta^n}[S_{d-1}(g)]} - \frac{\mathrm{E}_{\beta_0^n}[S_d(g)]}{r^n E_{\beta_0^n}[S_{d-1}(g)]}\right| \leq \phi^n.$$

By (C.4), it follows that[42]

$$q_d^n / q_{d-1}^n = 1 - (1-\beta_L^n)^{1/2}(1-q_L^n)^{1/2}$$

and

$$q_{0,d}^n / q_{0,d-1}^n = 1 - (1-\beta_{0,L}^n)^{1/2}\left(1-q_{0,L}^n\right)^{1/2}.$$

Suppose that $|q_L^n - q_{0,L}^n| \leq r^n \phi^n$. For any $\gamma > 0$, if

$$\frac{|\beta_L^n - \beta_{0,L}^n|}{\max(\beta_L^n, \beta_{0,L}^n)} > \gamma$$

then

$$|\beta_L^n - \beta_{0,L}^n| > c_1 r^n \gamma,$$

and so it follows that for small enough ϕ^n, $|q_d^n/q_{d-1}^n - q_{0,d}^n/q_{0,d-1}^n| > \phi^n$, which is a contradiction. Thus, it must also be that there is a sequence $\gamma^n \to 0$, and a further subsequence of

[42] We use the obvious notation for q_d^n and $q_{0,d}^n$, and so forth following (C.1) and (C.2), to indicate the expected counts associated with the parameters β^n and β_0^n.

our parameters such that

$$|\beta_L^n - \beta_{0,L}^n| \leq \gamma^n$$

along the subsequence.

Along the subsequence it must then be that

$$\frac{|\overline{\beta}_K^n - \overline{\beta}_{0,K}^n|}{\max(\overline{\beta}_K^n, \overline{\beta}_{0,K}^n)} > \varepsilon,$$

and so

$$|\overline{\beta}_K^n - \overline{\beta}_{0,K}^n| > c_1 r^n \varepsilon.$$

But now, from (C.1)

$$q_L^n = \beta_L^n + (1 - \beta_L^n)\overline{\beta}_K^n$$

and

$$q_{0,L}^n = \beta_{0,L}^n + (1 - \beta_{0,L}^n)\overline{\beta}_{0,K}^n.$$

Then given that $\beta_{0,L}^n, \overline{\beta}_{0,K}^n$ are bounded away from 1, and that $|\beta_L^n - \beta_{0,L}^n| \leq \gamma^n \to 0$, while $|\overline{\beta}_K^n - \overline{\beta}_{0,K}^n| > c_1 r^n \varepsilon$ implies the the difference in the two equations above cannot tend to 0, which is a contradiction. Thus, our supposition was incorrect. ■

Appendix D. Identification with Dense Links and Triangles

In the case from Section 4.4 in which $h_L \leq 1/2$ we can identify links and triangles.

In particular, it is useful to count 'almost' triangles, or two-stars: triples ijk for which exactly two out of three links are present. Let

$$S_2(g) = \frac{\#\{ijk : g_{ij} = 1, g_{ik} = 1, g_{jk} = 0\}}{3\binom{n}{3}}$$

be the fraction of triples for which two of the links are present and the third is missing.

Note then that the probability that two out of three links are present in a triple is

$$\mathrm{E}[S_2(g)] = (1 - \beta_{0,T})\widetilde{q}_L^2(1 - \widetilde{q}_L),$$

while the probability of a triangle is

$$\mathrm{E}[S_T(g)] = \beta_{0,T} + (1 - \beta_{0,T})\widetilde{q}_L^3.$$

Thus,

$$\beta_{0,T} = \mathrm{E}[S_T(g)] - \mathrm{E}[S_2(g)]\frac{\widetilde{q}_L}{(1 - \widetilde{q}_L)}.$$

In a case in which $h_T \geq h_L + 1$ and $h_L \in (0, 1/2]$, it then follows that $\frac{\widetilde{q}_L}{(1-\widetilde{q}_L)} \approx q_L$, and so we can estimate β_T by $S_T(g) - S_2(g)S_L(g)$. Then having identified β_T, one can estimate β_L

from $S_L(g)$. We omit the details, but the logic of the approach is analgous to the technique used to prove Proposition 4.4.

Appendix E. An Extension to Continuous and Interdependent Covariates

For simplicity, we have focused on models in which covariates are captured by indexing subgraphs by covariates. This encompasses covariates that take on a finite set of values or are approximated by a finite set of values, and is a flexible approach, although it may not work as well with fully continuous data that take on a wide range of values. Such continuous covariates can also easily be handled, as our models and results have natural extensions.

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

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

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

Environment. Every node $i \in \{1,...,n\}$ draws a d-dimensional covariate vector $x^n_i \in \mathcal{X}$. For simplicity we let $\mathcal{X} = \prod_{k=1}^d [x_{L,k}, x_{H,k}]$ be a d-dimensional product of intervals of $\mathbb{R}$.[43] Letting x_ℓ denote the $d \times n$ matrix of data, we assume $x_\ell x'_\ell$ has full rank along the sequence. For expositional simplicity in our proofs we considering a sequence of fixed-regressors, $x_{\ell,n}$ where n indexes the sequence. Clearly stochastic regressors can be accommodated.

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

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

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

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

A group of size m_ℓ forms with a probability $p_\ell^n(x_{\ell,j}; \gamma_\ell)$ which depends on some function of the m_ℓ individuals' characteristics and a parameter γ_ℓ, whose value in theory may depend on n.[44]

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

(1) a linear probability model with uniform link function $p_\ell^n(x_{\ell,j}; \gamma_\ell) = \gamma'_{\ell,n} x_{j,\ell}$,
(2) a logistic regression model $p_\ell^n(x_{\ell,j}; \gamma_\ell) = \frac{\exp(\gamma'_{\ell,n} x_{j,\ell})}{1+\exp(\gamma_{\ell,n} x_{j,\ell})}$,

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

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

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

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

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

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

A link forms with probability

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

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

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

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

Estimation. The above defines a well-defined network-generation process. As before, we need a relative sparsity condition to hold so that when we count a structure, with probability approaching one it was not incidentally generated. Here we provide a sufficient condition for relative sparsity hold as the continuous covariates vary. The condition is that given m_ℓ nodes, no matter what the value of each covariate is among these nodes, the probability of forming the subgraph isomorphic to H_ℓ shrinks at the same as n grows to infinity. This ensures that the relative rate of incidentally generated objects is unaffected by the particular values of the covariates.[45]

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

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

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

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

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

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

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

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

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

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

THEOREM **E.1.** *Assume* $\|\gamma_{0,n,\ell}\|_1 = \Theta(1/n^{m_\ell - h_\ell})$[46] *with* $0 < h_\ell < m_\ell$ *and the* h_ℓ *are such that relative sparsity condition is satisfied. Then*

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

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

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

Logistic Regression. We turn to our main example where $p^n_\ell(x_{\ell,j}; \gamma_\ell)$ is given by a logistic link function. In all that follows $\gamma_{0,n}$ consists of elements that are either order constant or tending to $-\infty$, as discussed below.

THEOREM **E.2.** *If* $\|\gamma_{0,n}\|_1 \cdot \sup_{x\in\mathcal{X}} \|x\|_\infty \lesssim h_\ell \cdot \log n^{m_\ell}$ *for* $0 \leq h_\ell < m_\ell$*, and relative sparsity holds, then*

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

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

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

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

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

[46] $f_n \in \Theta(g_n)$ means $\exists k_1 > 0, \exists k_2 > 0, \exists n_0 > 0, \forall n > n_0$ such that $g_n k_1 < f_n < g_n k_2$.

some constant $C > 0$. It is easy to show by the Lindeberg-Feller central limit theorem for triangular arrays that in this case

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

provided $p_n n \to \infty$. This implies that $\sqrt{np_n}\,(\widehat{\alpha} - \alpha_n) = \sqrt{n^{1-h}}\,(\widehat{\alpha} - \alpha_n) \rightsquigarrow \mathcal{N}(0,1)$. This follows from observing that α_n will be consistent[47] and by the delta method.

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

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

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

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

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

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

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

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

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

This can be expressed as

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

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

where[48]

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

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

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

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

Because of relative sparsity, incidental generation is rare. Therefore, for most of the data the preceding result directly applies. However, for a vanishing proportion of m_ℓ-tuples, the structures are present due to incidental generation. We only need to show that this happens for a vanishing proportion of data and is asymptotically negligible.

To make this argument, out of the n observations, we say that each observation is "invalid" (i.e., observed with measurement error such as $y_{in} = 1$ when the true value is 0) with some probability. Our claim can be written in the notation of the preceding lemma by saying that some of our n data points are "invalid" and we show that the probability that an observation is invalid is bounded by $z_n \downarrow 0$ at a fast enough rate. Our relative sparsity assumption directly implies that $z_n \downarrow 0$.

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

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

Then the conclusion of Lemma E.2 holds.

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

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

Observe that

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

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

This implies

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

Thus

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

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

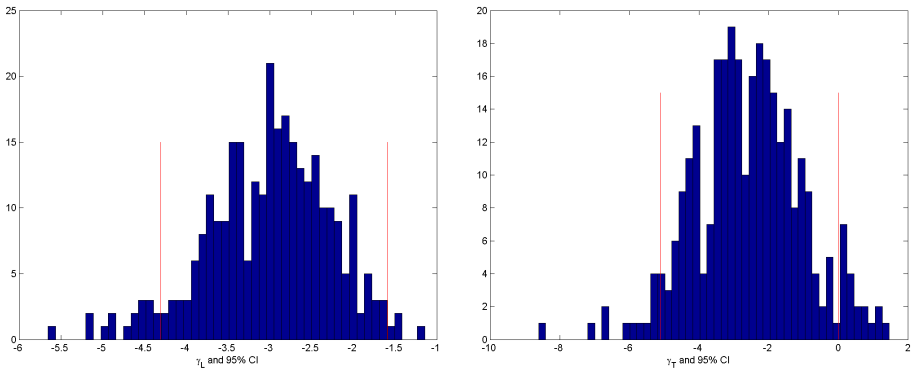
$$\begin{aligned}\frac{1}{\sqrt{z_n n}}\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} &= \frac{1}{\sqrt{n^{1-k}}}\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in}=O_p(1),\ \text{where } k=h+\delta\\ \implies O\left(\frac{1}{\sqrt{n^{1-h}}}\right)\sum_{n^*<i\leq n}(y_{in}-q_{in})\,x_{in} &= O\left(\frac{1}{\sqrt{n^{1-k+\delta}}}\right)\sum_{n^*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

showing the result. ■

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

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

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



(A) Links parameter (B) Triads parameter

FIGURE E.1. Displays the distribution of estimated parameter value as well as the median 95% confidence interval from a simple logistic regression.

APPENDIX F. EXTENSIONS OF TABLE 1

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

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

We use a logistic regression for this estimation.

The estimated a vector of regression coefficients for each village capture how characteristics of a dyad correspond to linking probabilities. This gives a predicted probability that each

household is linked to each of the other households in the village. We use these predicted probabilities to generate 100 simulated networks per village and study the characteristics of the resulting networks. These are presented in column [3] of Table F.1.

TABLE F.1. Estimation of Additional Models: Extension of Table 1

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

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

Column [3] contains the statistics from the enriched link-based model, while the remainder of the table is exactly the same as what is presented in the body of the paper. Adding over 12 parameters to flexibly control for demographic attributes makes almost no difference in generating network characteristics that match the observed data, providing very small improvements, and still not coming close to doing as well as the simple SUGMs. Moreover, since the specification developed here makes use of considerably richer data than those used in the two candidate SUGM models, it suggests that by decomposing a network into a tapestry of random structures (triangles, links and even isolates), considerable value is added in modeling higher order features of networks in a parsimonious way.

In Table F.2, we show the results of Table 1 adding standard errors, to show that the SUGM models better replicate patterns in the data.

TABLE F.2. Network Properties: Extended

		Data	Link-based model with covariates	SUGM with links and triangles	SUGM with isolates, links and triangles
		[1]	[2]	[3]	[4]
Models are fit to different combinations of these statistics.	Number of Unsupported Links	160.8	236.2	161.2	161.8
		(9.8536)	(13.7273)	(10.4048)	(10.9867)
	Number of Triangles	39.2	3.1	39.7	39.5
		(3.9425)	(0.3257)	(5.8249)	(3.7884)
	Average Degree	2.3243	2.3260	2.5916	2.5219
		(0.0555)	(0.0569)	(0.1019)	(0.0880)
	Number of Isolates	54.9722	25.7222	31.4444	65.9167
		(3.4599)	(1.6322)	(4.1232)	(4.0961)
None of the models are directly fit to any of these statistics.	Average Clustering	0.0895	0.0105	0.1268	0.0829
		(0.0042)	(0.0014)	(0.0074)	(0.0072)
	Fraction in Giant Component	0.7061	0.8315	0.7982	0.6718
		(0.0138)	(0.0148)	(0.0136)	(0.0133)
	First Eigenvalue	5.5446	3.8578	4.6762	5.3025
		(0.1649)	(0.0737)	(0.0970)	(0.1166)
	Spectral Gap	0.9550	0.3354	0.6684	1.0617
		(0.0889)	(0.0338)	(0.0502)	(0.0646)
	Second Eigenvalue of Stochastized Matrix	0.9573	0.9632	0.9559	0.9069
		(0.0035)	(0.0037)	(0.0038)	(0.0034)
	Average Path Length	4.6921	5.6565	5.1215	4.1180
		(0.0986)	(0.1887)	(0.1106)	(0.1212)

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