Graph Limits & Subgraph Counts

Econometric Methods for Networks,

SMU, May 29th & June 1st, 2017

Bryan S. Graham

University of California - Berkeley

Introduction

In 1970 Paul Holland and Samuel Leinhardt (1970, *AJS*) introduced the *triad census*.

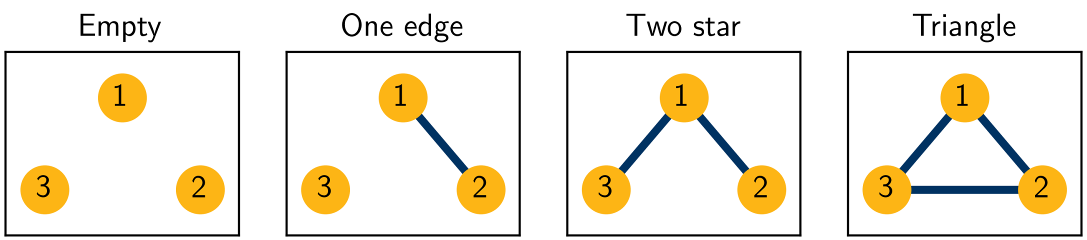
- counts of all 4 (16) unique triad isomorphisms in an undirected (directed) graph;

- can construct transitivity index (TI) from triad census...

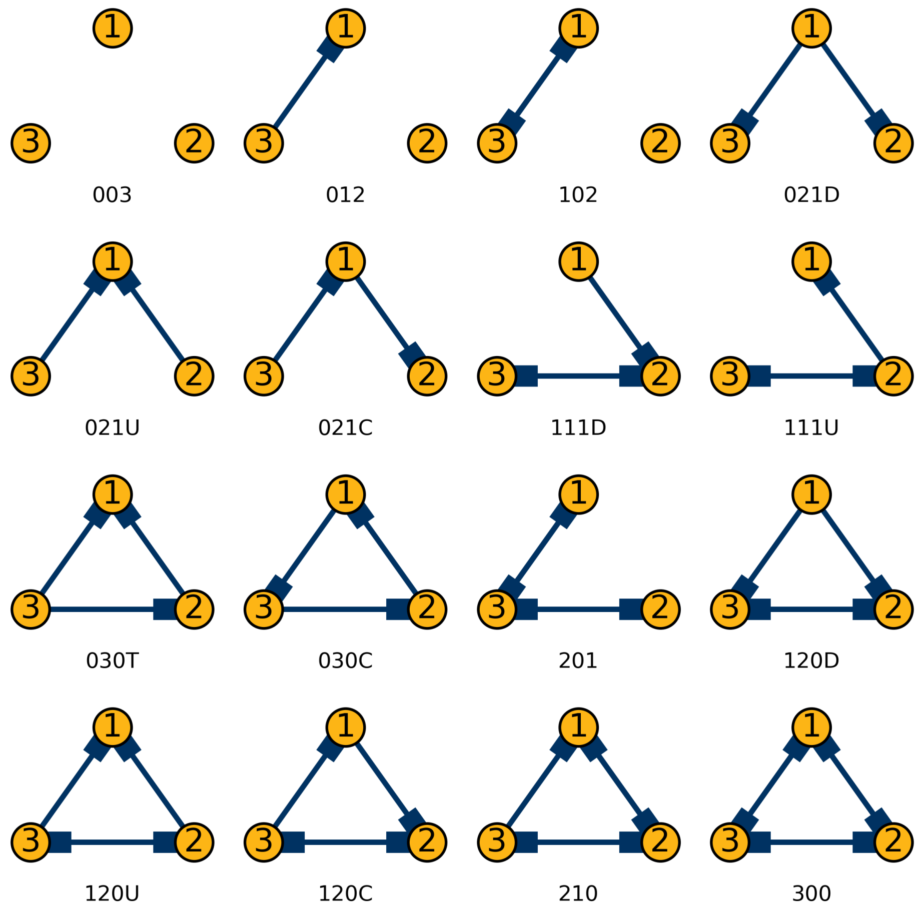
- ...as well as the mean and variance of the degree sequence.

Holland and Leinhardt (1976, SM) provided variance expressions for these counts (brute force).

Triads: Undirected Case



Triads: Directed Case



Introduction (continued)

In early work normality of these counts was assumed (w/o proof).

Nowicki (1989, 1991) showed asymptotic normality of counts for homogenous random graphs.

Bickel, Chen & Levina (2011, AS) demonstrated asymptotic normality in the "general" case under specific conditions.

Introduction (continued)

Subgraph counts called *network moments* by Bickel, Chen and Levina (2011); summarize average local properties of a network.

Large literature in sociology which uses triad counts to "test" various hypotheses

- see Holland and Leinhardt (1976, SM) and Wasserman and Faust (1994)

- cf., computational biology (e.g., Milo et al., 2002)

Asymptotic distribution theory puts these tests on firmer ground.

Introduction (continued)

Subgraph frequencies might be used to (partially) identify structural models of network formation (e.g., de Paula et al., 2015).

indirect inference approach:

1. use structural model to simulate networks...and count subgraphs;

2. compare simulated counts with actual counts;

3. estimate structural parameters by minimum distance.

Setup

Let $G(\mathcal{V}, \mathcal{E})$ be a finite undirected random graph with

- agents/vertices $\mathcal{V} = \{1, \ldots, N\}$,
- links/edges $\mathcal{E} = \{\{i, j\}, \{k, l\}, \ldots\}$, and
- adjacency matrix $\mathbf{D} = \left[D_{ij}\right]$ with

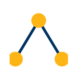
$$D_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in \mathcal{E} \\ 0 & \text{otherwise} \end{cases}$$

Subgraphs

- (Partial Subgraph) Let $\mathcal{V}(S) \subseteq \mathcal{V}(G)$ be any subset of the vertices of G and $\mathcal{E}(S) \subseteq \mathcal{E}(G) \cap \mathcal{V}(S) \times \mathcal{V}(S)$, then $S = (\mathcal{V}(S), \mathcal{E}(S))$ is an *partial subgraph* of G.

- (Induced Subgraph) Let $\mathcal{V}(S) \subseteq \mathcal{V}(G)$ be any subset of the vertices of G and $\mathcal{E}(S) = \mathcal{E}(G) \cap \mathcal{V}(S) \times \mathcal{V}(S)$, then $S = (\mathcal{V}(S), \mathcal{E}(S))$ is an *induced subgraph* of G.

Subgraphs (continued)

- The induced subgraph S includes *all* edges in G connecting any two agents in $\mathcal{V}(S)$
  - a (partial) subgraph may include only a subset of such edges
  - $S =$  is a partial subgraph of $G =$ , but not an induced subgraph

Graph Isomorphism

- Consider two graphs, R and S, of the same order.

- Let $\varphi : \mathcal{V}(R) \to \mathcal{V}(S)$ be a bijection from the nodes of R to those of S.

- The bijection $\varphi : \mathcal{V}(R) \to \mathcal{V}(S)$

  - *maintains adjacency* if for every dyad $i, j \in \mathcal{V}(R)$ if $\{i, j\} \in \mathcal{E}(R)$, then $\{\varphi(i), \varphi(j)\} \in \mathcal{E}(S)$;

  - *maintains non-adjacency* if for every dyad $i, j \in \mathcal{V}(R)$ if $\{i, j\} \notin \mathcal{E}(R)$, then $\{\varphi(i), \varphi(j)\} \notin \mathcal{E}(S)$.

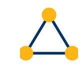
Graph Isomorphism (continued)

- If the bijection maintains both adjacency and non-adjacency we say it *maintains structure*.

- (Graph Isomorphism) The graphs R and S are *isomorphic* if there exists a structure-maintaining bijection $\varphi : \mathcal{V}(R) \to \mathcal{V}(S)$.

- Notation: $R \cong S$ means "R is isomorphic to S."

Induced Subgraph Density

- S is a p^{th}-order graphlet of interest (e.g., $S =$  or $S =$ )

- G_N is the network/graph under study

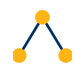
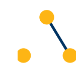
- $\mathbf{i}_p \subseteq \{1, 2, \ldots, N\}$ is a set of p integers with $i_1 < i_2 < \cdots < i_p$
  - $\mathcal{C}_{p,N}$ is set of all $\binom{N}{p}$ such integer sets
  - $G\left[\mathbf{i}_p\right]$ is the induced subgraph of G associated with vertex set $\mathbf{i}_p$

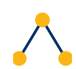
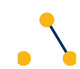
Induced Subgraph Density (continued)

- The *induced subgraph density* of S in G_N, denoted by $t_{\mathsf{ind}}(S, G_N)$ or $P_N(S)$ equals the probability that $G_N[\mathbf{i}_p]$, for $\mathbf{i}_p$ chosen uniformly at random from $C_{p,N}$, is isomorphic to S:

$$\begin{aligned} t_{\mathsf{ind}}(S, G_N) &= \binom{N}{p}^{-1} \sum_{\mathbf{i}_p \in C_{p,N}} \mathbf{1}\left(S \cong G_N[\mathbf{i}_p]\right) \\ &= \Pr\left(S \cong G_N[\mathbf{i}_p]\right) \\ &= P_N(S) \end{aligned}$$

Induced Subgraph Density (Examples)

- $t_{\text{ind}}($  $,$  $) = \frac{2}{4}$, $t_{\text{ind}}($  $,$  $) = \frac{2}{4}$ and $t_{\text{ind}}($  $,$  $) = \frac{0}{4}$

- $t_{\text{ind}}($  $,$  $) = \frac{1}{4}$, $t_{\text{ind}}($  $,$  $) = \frac{2}{4}$ and $t_{\text{ind}}($  $,$  $) = \frac{1}{4}$

Induced Subgraph Density: Graphon Case

Let $h\left(U_i, U_j\right)$ be a valid graphon.

Let $\mathsf{iso}(S)$ be the group of isomorphisms of S, and $|\mathsf{iso}(S)|$ its cardinality.

Under the "Aldous-Hoover DGP" the *ex ante* probability that an induced p-subgraph is isomorphic to S is given by

$$\begin{aligned} t_{\text{ind}}(S,h) &= |\mathsf{iso}(S)| \\ &\times \mathbb{E}\left[\prod_{\{i,j\}\in\mathcal{E}(S)} h\left(U_i,U_j\right) \prod_{\{i,j\}\in\mathcal{E}(\bar{S})} \left[1-h\left(U_i,U_j\right)\right]\right] \\ &= P(S). \end{aligned}$$

Graph Limits

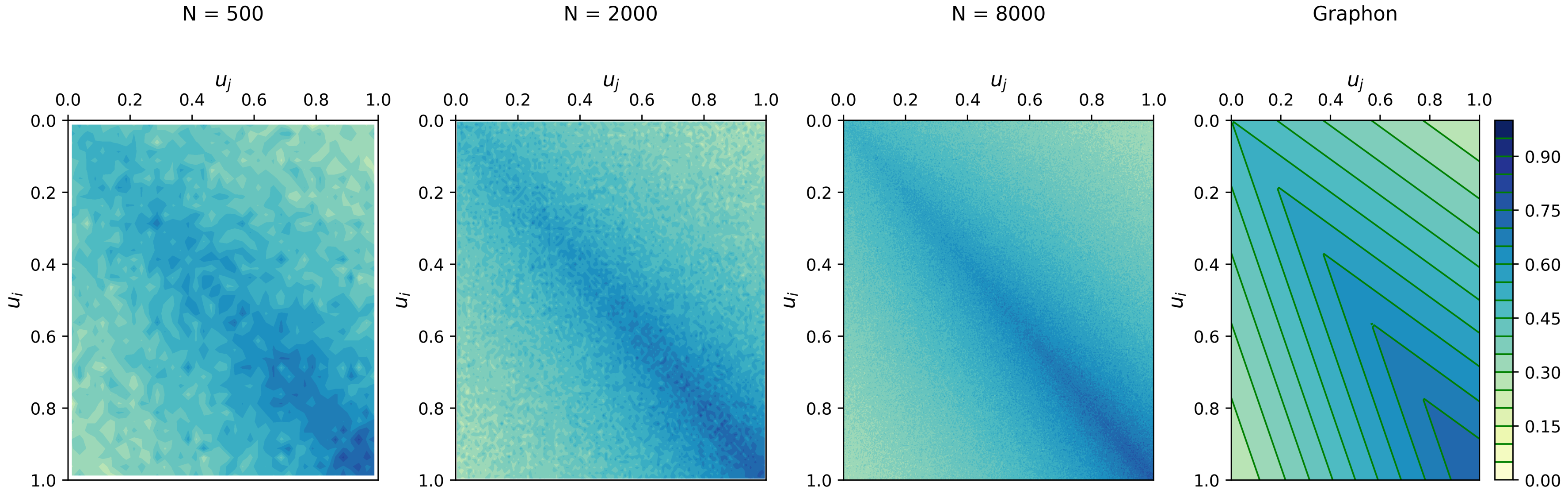
Let $\{G_N\}_{N=1}^{\infty}$ be a sequence of networks. If

$$\lim_{N\to\infty} t_{\text{ind}}\left(S, G_N\right) = t_{\text{ind}}\left(S, h\right)$$

for some graphon $h\left(\cdot,\cdot\right)$ and *all* fixed subgraphs S, then we say that G_N converges to $h\left(\cdot,\cdot\right)$.

- Lovász (2012) for complete development.
- Diaconis and Janson (2008) for connections with Aldous-Hoover Theorem.
- Result establishes a connection between subgraph counts and the graphon.

Graph Limits: Example



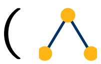
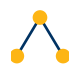
(Injective) Homomorphism Density

The homomorphism density gives the probability that S is (isomorphic to) a subgraph of a randomly selected induced subgraph of G_N of order $p = |\mathcal{V}(S)|$

Alternatively the homomorphism density equals fraction of injective mappings $\varphi : \mathcal{V}(S) \to \mathcal{V}(G_N)$ that preserve edge adjacency

$$
\begin{aligned}
t_{\mathsf{hom}}(S, G_N) &= \frac{1}{\binom{N}{p}\,|\mathsf{iso}(S)|} \sum_{R \subseteq K_N, R \cong S} \mathbf{1}(R \subseteq G_N) \\
&= \frac{1}{\binom{N}{p}\,|\mathsf{iso}(S)|} \sum_{R \subseteq K_N, |V(R)| = p} \mathbf{1}(R \cong S) \prod_{\{i,j\} \in \mathcal{E}(R)} D_{ij} \\
&= Q_N(S)
\end{aligned}
$$

Homomorphism Density (continued)

Summation in $t_{\text{hom}}\,(S, G_N) = Q_N\,(S)$ is over the $\binom{N}{3}\left|\text{iso}(\right.$  $\left.)\right| = \frac{3}{6}N\,(N-1)\,(N-2)$ (partial) subgraphs of K_N (the complete graph) which are isomorphic to $S =$ ) .

We count the number of these subgraphs which are also *partial* subgraphs of G_N

Homomorphism Density (continued)

The expected value of $Q_N(S)$ is:

$$
\begin{aligned}
\mathbb{E}\left[Q_N(S)\right] &= \frac{1}{\binom{N}{p}\left|\mathrm{iso}(S)\right|} \sum_{R \subseteq K_N, |V(R)|=p} \left\{ \mathbf{1}\left(R \cong S\right) \right. \\
&\left. \times \mathbb{E}\left[\mathbb{E}\left[\prod_{\{i,j\} \in \mathcal{E}(R)} D_{ij} \middle| U_1, \ldots, U_N \right]\right] \right\} \\
&= \mathbb{E}\left[\prod_{\{i,j\} \in \mathcal{E}(S)} h\left(U_i, U_j\right) \right] \\
&= Q(S) \stackrel{def}{\equiv} t_{\mathrm{hom}}(S, h)
\end{aligned}
$$

Can also use $t_{\mathrm{hom}}(S, G_N)$ to define graph convergence.

Recap

Induced subgraph density, $P_N(S)$: probability that $G_N[\mathbf{i}_p]$, for $\mathbf{i}_p$ chosen uniformly at random from $C_{p,N}$, is isomorphic to S.

Homomorphism density, $Q_N(S)$: probability that *a (partial) subgraph of* $G_N[\mathbf{i}_p]$, for $\mathbf{i}_p$ chosen uniformly at random from $C_{p,N}$, is isomorphic to S.

If $\lim_{N\to\infty} P_N(S) = t_{\text{ind}}(S,h)$ for some graphon $h(\cdot,\cdot)$ and all fixed subgraphs S, then we say that G_N converges to $h(\cdot,\cdot)$.

One more tool! Graphlet Stitchings

Graph union: $T \cup U = G\left(\mathcal{V}\left(T\right) \cup \mathcal{V}\left(U\right), \mathcal{E}\left(T\right) \cup \mathcal{E}\left(U\right)\right)$.

Let $W_{q,R,S}$ be a union of two isomorphisms, respectively T and U, of the graphlets R and S with

1. $\left|\mathcal{V}\left(R\right)\right| = \left|\mathcal{V}\left(S\right)\right| = p$;

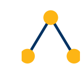
2. $\left|\mathcal{V}\left(R\right) \cap \mathcal{V}\left(S\right)\right| = q$ vertices in common;

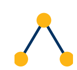
3. identical structures across all vertices in common.

The multiset of all such graphlet stitchings (including isomorphisms) is denoted by $\mathcal{W}_{q,R,S}$ (with $\mathcal{W}_{q,S,S} = W_{q,S}$).

Graphlet Stitching: Example #1

Let the graphlets $R =$  and $S =$  share one vertex in common.

- There is just one possible way to join them: $R \cup S \cong$  .

We therefore have that $\mathcal{W}_{1,}$  $= \{$  $\}$.

Graphlet Stitching: Example #1 (continued)

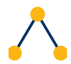
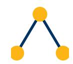
Define the probability of observing an element of $\mathcal{W}_{1,}$ as a subgraph of a randomly sampled triad as

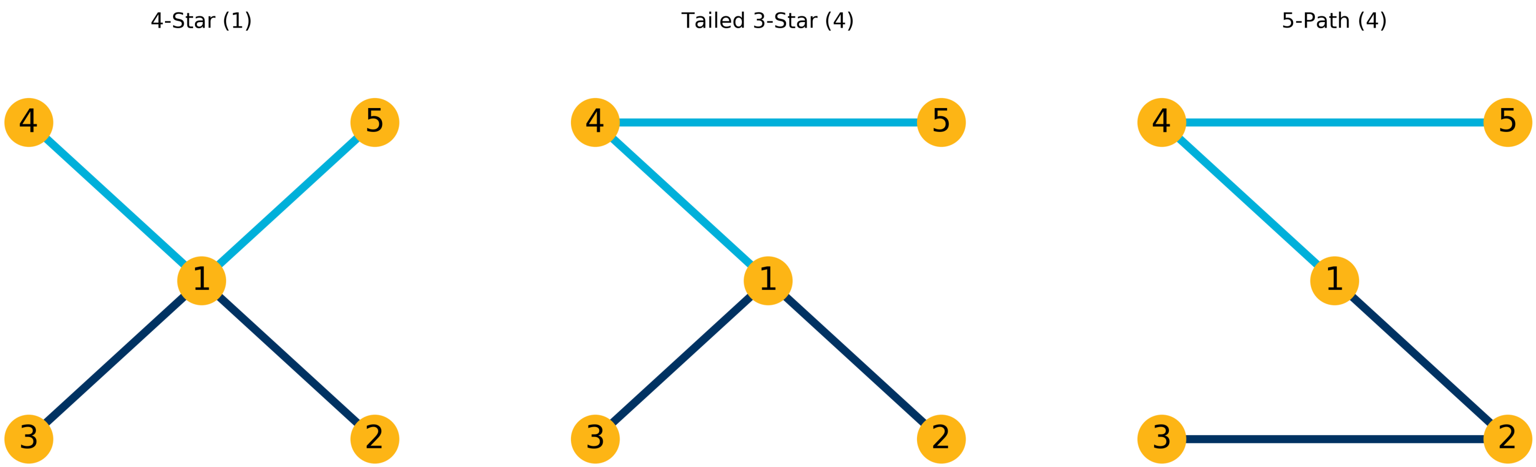
$$Q\left(\mathcal{W}_{1,}\right) = \sum_{W \in \mathcal{W}_{1,}} Q\left(W\right)$$
$$= Q\left(\quad\right)$$
$$= \mathbb{E}\left[D_{12}D_{13}\right]$$

with $Q\left(W\right)$ the homomorphism density introduced above.

For $q = 2$ (two nodes in common) we have, of course, $\mathcal{W}_{2,} = \left\{ \quad \right\}$.

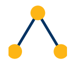
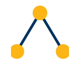
Graphlet Stitching: Example #2

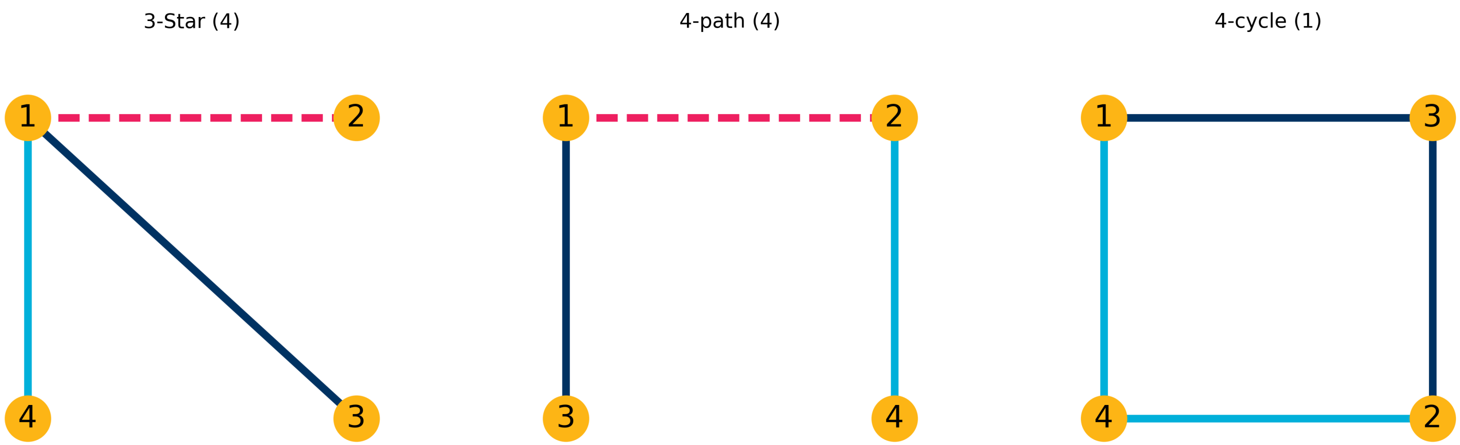
There are nine ways (three up to isomorphisms) to join the graphlets $R =$  and $S =$ , sharing one vertex in common.



Notes: Number of isomorphisms of each graphlet in $\mathcal{W}_{1,q}$ given in parentheses.

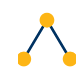
Graphlet Stitching: Example #2 (continued)

There are nine ways (three up to isomorphisms) to join the graphlets $R =$  and $S =$ , sharing two vertices in common.



Notes: Number of isomorphisms of each graphlet in $\mathcal{W}_{2,q}$ given in parentheses.

Estimation of Subgraph Frequencies

- We will develop explicit results for two subgraph frequencies
  - the frequency of connected dyads: $S =$ 
  - the frequency of two star triads: $S =$ 
- General case involves no new ideas...
  - ...but can be *very* tedious in practice
  - good software would be a real help

Density

We estimate $\rho_N = \Pr\left(D_{ij} = 1\right)$ by

$$\hat{\rho}_N = \frac{2}{N(N-1)} \sum_{i<j} D_{ij}.$$

Projecting onto $U_1, ..., U_N$ yields the decomposition:

$$\hat{\rho}_N = \underbrace{\frac{2}{N(N-1)} \sum_{i<j} h_N\left(U_i, U_j\right)}_{\text{U-Statistic}} + \underbrace{\frac{2}{N(N-1)} \sum_{i<j} \left(D_{ij} - h_N\left(U_i, U_j\right)\right)}_{\text{"Poisson Binomial R.V''}}$$
$$= U_N + T_N.$$

Observe that T_N is mean independent of U_N.

Density: Variance Calculation

We have

$$\begin{aligned}\mathbb{V}\left(\hat{\rho}_N\right) &= \mathbb{V}\left(U_N\right) + \mathbb{V}\left(T_N\right) + 2\mathbb{C}\left(U_N, T_N\right) \\ &= \mathbb{V}\left(U_N\right) + \mathbb{V}\left(T_N\right).\end{aligned}$$

A Hoeffding (1948) variance decomposition gives

$$\mathbb{V}\left(U_N\right) = \binom{N}{2}^{-2} \sum_{q=1}^{2} \binom{N}{2}\binom{2}{q}\binom{N-2}{2-q}\Omega_q$$

for

$$\Omega_q = \mathbb{C}\left(h_N\left(U_{i_1}, U_{i_2}\right), h_N\left(U_{j_1}, U_{j_2}\right)\right)$$

with $\{i_1, i_2\}$ and $\{j_1, j_2\}$ sharing $q = 1, 2$ indices in common.

Density: Variance Calculation (continued)

Evaluating Ω_1 yields

$$
\begin{aligned}
\Omega_1 &= \mathbb{E}\left[h_N\left(U_1,U_2\right)h_N\left(U_1,U_3\right)\right] - \mathbb{E}\left[h_N\left(U_1,U_2\right)\right]\mathbb{E}\left[h_N\left(U_1,U_3\right)\right] \\
&= Q\left(\mathcal{W}_{1,\,\bullet\!\!-\!\!\bullet}\right) - P\left(\bullet\!\!-\!\!\bullet\right)P\left(\bullet\!\!-\!\!\bullet\right) \\
&= Q\left(\overset{\bullet}{\diagup\diagdown}\right) - P\left(\bullet\!\!-\!\!\bullet\right)P\left(\bullet\!\!-\!\!\bullet\right).
\end{aligned}
$$

Evaluating Ω_2 yields

$$
\begin{aligned}
\Omega_2 &= \mathbb{E}\left[h_N\left(U_1,U_2\right)^2\right] - \mathbb{E}\left[h_N\left(U_1,U_2\right)\right]\mathbb{E}\left[h_N\left(U_1,U_2\right)\right] \\
&= \mathbb{V}\left(\mathbb{E}\left[D_{12}\middle|\,\mathbf{U}\right]\right).
\end{aligned}
$$

Density: Variance Calculation (continued)

Evaluating the variance of $\mathbb{V}(T_N)$ we get

$$\begin{aligned}
\mathbb{V}(T_N) &= \mathbb{V}(\mathbb{E}[T_N|\,\mathbf{U}]) + \mathbb{E}[\mathbb{V}(T_N|\,\mathbf{U})] \\
&= 0 + \left(\frac{2}{N(N-1)}\right)^2 \mathbb{E}\left[\mathbb{V}\left(\sum_{i<j}\Big(D_{ij} - h_N\big(U_i, U_j\big)\Big)\Bigg|\,\mathbf{U}\right)\right] \\
&= \left(\frac{2}{N(N-1)}\right)^2 \mathbb{E}\left[\sum_{i<j}\mathbb{V}\Big(D_{ij} - h_N\big(U_i, U_j\big)\Big|\,\mathbf{U}\Big)\right] \\
&= \frac{2}{N(N-1)}\mathbb{E}[\mathbb{V}(D_{12}|\,\mathbf{U})].
\end{aligned}$$

Density: Variance Calculation (continued)

Collecting terms we have:

$$\begin{aligned}
\mathbb{V}\left(\widehat{\rho}_N\right) &= \frac{4\left(N-2\right)}{N\left(N-1\right)}\left[Q\left(\wedge\right) - P\left(\bullet\!-\!\bullet\right)P\left(\bullet\!-\!\bullet\right)\right] \\
&+ \frac{2}{N\left(N-1\right)}\mathbb{V}\left(\mathbb{E}\left[\left.D_{12}\right|\mathbf{U}\right]\right) + \frac{2}{N\left(N-1\right)}\mathbb{E}\left[\mathbb{V}\left(\left.D_{12}\right|\mathbf{U}\right)\right] \\
&= \frac{4\left(N-2\right)}{N\left(N-1\right)}\left[Q\left(\wedge\right) - P\left(\bullet\!-\!\bullet\right)P\left(\bullet\!-\!\bullet\right)\right] \\
&+ \frac{2}{N\left(N-1\right)}P\left(\bullet\!-\!\bullet\right)\left(1 - P\left(\bullet\!-\!\bullet\right)\right).
\end{aligned}$$

Density: Variance Calculation (continued)

To allow for graph sequences where $\rho_N \to 0$ as $N \to \infty$ we normalize''

- Let $\tilde{Q}\left(\wedge\right) = \dfrac{Q\left(\wedge\right)}{\rho^2}$ and $\tilde{P}\left(\bullet\!-\!\bullet\right) = \dfrac{P\left(\bullet\!-\!\bullet\right)}{\rho_N}$.

- Recall that $\lambda_N = (N-1)\,\rho_N$.

Density: Variance Calculation (continued)

After normalization:

$$\begin{aligned}\mathbb{V}\left(\frac{\widehat{\rho}_N}{\rho_N}\right) &= \frac{4\,(N-2)}{N\,(N-1)}\left[\tilde{Q}\left(\wedge\right) - \tilde{P}\left(\bullet\!-\!\bullet\right)\tilde{P}\left(\bullet\!-\!\bullet\right)\right] \\ &\quad + \frac{2}{N\lambda_N}\tilde{P}\left(\bullet\!-\!\bullet\right) - \frac{2}{N\,(N-1)}\tilde{P}\left(\bullet\!-\!\bullet\right)^2 \\ &= O\left(\frac{1}{N}\right) + O\left(\frac{1}{N\lambda_N}\right) + O\left(\frac{1}{N^2}\right).\end{aligned}$$

- If $\lambda_N \to \infty$ first term dominates.

- If $\lambda_N \to \lambda_0 > 0$, first two terms dominate.

Asymptotic Inference

Asymptotic theory for U-Statistics gives, for $\lambda_N \to \infty$ as $N \to \infty$

$$\sqrt{N}\left(\frac{\widehat{\rho}_N}{\rho_N} - 1\right) \xrightarrow{D} \mathcal{N}\left(0, 4\left[\tilde{Q}\left(\wedge\right) - \tilde{P}\left(\bullet\!-\!\bullet\right)\tilde{P}\left(\bullet\!-\!\bullet\right)\right]\right).$$

Result (in high level form) due to Bickel, Chen and Levina (2011, *Annals of Statistics*).

Comment: Under Erdos-Renyi $\tilde{Q}\left(\wedge\right) = \tilde{P}\left(\bullet\!-\!\bullet\right)\tilde{P}\left(\bullet\!-\!\bullet\right)$.

Variance Estimation

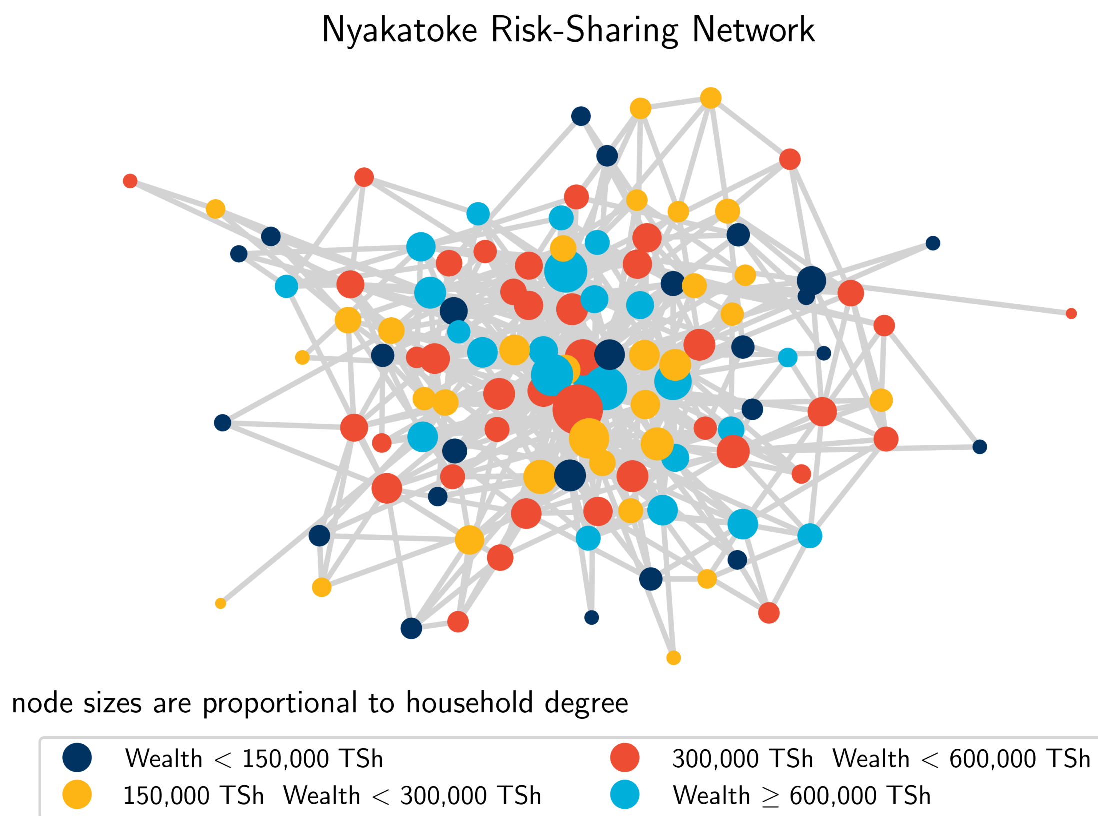
We can estimate the asymptotic variance using the analog estimators:

$$\begin{aligned}\hat{Q}\left(\wedge\right) &= \binom{N}{3}^{-1} \sum_{i<j<k} \frac{1}{3}\left\{D_{ij}D_{ik} + D_{ij}D_{jk} + D_{ik}D_{jk}\right\} \\ &= \binom{N}{3}^{-1} \frac{1}{3}\left[T_{\mathsf{TS}} + 3T_{\mathsf{T}}\right]\end{aligned}$$

and

$$\hat{P}\left(\bullet\!-\!\bullet\right) = \binom{N}{2}^{-1} \sum_{i<j} D_{ij}$$

Nyakatoke


Nyakatoke Risk-Sharing Network
node sizes are proportional to household degree
Wealth < 150,000 TSh
150,000 TSh Wealth < 300,000 TSh
300,000 TSh Wealth < 600,000 TSh
Wealth ≥ 600,000 TSh

Variance Estimation for $\widehat{P}\left(\bullet\!-\!\bullet\right)$: Nyakatoke

For Nyakatoke we have

$$\widehat{Q}\left(\overset{\bullet}{\bullet\!\bigwedge\!\bullet}\right) \cong 0.006105$$

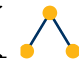
and

$$\widehat{P}\left(\bullet\!-\!\bullet\right) \simeq 0.0698$$

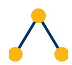
which gives

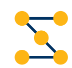
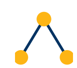
$$\underset{\text{(a.s.e)}}{\widehat{\rho}_N} = \underset{(0.0072)}{0.0698}, \quad \underset{\text{(a.s.e)}}{\widehat{\lambda}_N} = \underset{(0.8459)}{8.2364}$$

Note: Estimate above includes first two terms.

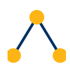
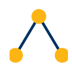
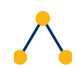
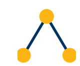
Limit Distribution of $\widehat{P}($  $)$

Define the multiset of graphlet stitchings:

- $\mathcal{W}_{1,}$  $= (\{$  $,$  $,$  $\}, m)$

- Here $m = \{($  $, 1), ($  $, 4), ($  $, 4)\}$ gives the multiplicity of each unique graphlet in $\mathcal{W}_{1,}$  .

Normalize graphlet according to number of edges in it:

- $\tilde{P}($  $) =$ $P($  $)$ / ρ_N^2 and $\tilde{Q}\left(\mathcal{W}_{1,}\right.$  $\left.\right) =$ $Q\left(\mathcal{W}_{1,}\right.$  $\left.\right)$ / ρ_N^4

Limit Distribution of $\widehat{P}(\wedge)$

If $\lambda_N \to \infty$ as $N \to \infty$, then

$$\sqrt{N}\left(\frac{\widehat{P}(\wedge)}{\rho_N^2} - \tilde{P}(\wedge)\right) \xrightarrow{D} \mathcal{N}\left(0, 9\left[\tilde{Q}\left(\mathcal{W}_{1,\wedge}\right) - \tilde{P}\left(\wedge\right)\tilde{P}\left(\wedge\right)\right]\right).$$

- Analysis involves a variance calculation along the lines outlined above.

- And the characterization of the limiting variance of a 3rd order U-Statistics.

Wrapping Up

In large graphs subgraph counting is computationally challenging

- implications for feasibility of both estimation and inference.
- see Bhattacharya and Bickel (2015) for a subsampling approach.

Very little (i.e., essentially none) empirical work using these results.

Tremendous scope for using these methods in empirical analysis; but not easy!