

# Graph Limits & Subgraph Counts

Econometric Methods for Networks,

CORE, Dec 12th - 14th, 2016

*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) provide variance expressions for these counts

## 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

## Introduction (continued)

- 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:
  - use structural model to simulate networks...and count subgraphs
  - compare simulated counts with actual counts

## Setup

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

- agents/vertices  $\mathcal{V} = \{1, \dots, N\}$ ,
- links/edges  $\mathcal{E} = \{\{i, j\}, \{k, l\}, \dots\}$ , and
- adjacency matrix  $\mathbf{D} = [D_{ij}]$  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
  - $S = \text{triangle}$  is a partial subgraph of  $G = \text{square}$ , but not an induced subgraph





## Graph Isomorphism

- Consider two graphs,  $R$  and  $S$ , of the same order.
- Let  $\varphi : \mathcal{V}(R) \rightarrow \mathcal{V}(S)$  be a bijection from the nodes of  $R$  to those of  $S$ .
- The bijection  $\varphi : \mathcal{V}(R) \rightarrow \mathcal{V}(S)$ 
  - *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  $\varphi : \mathcal{V}(R) \rightarrow \mathcal{V}(S)$ .
- 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
- $\mathbf{i}_p \subseteq \{1, 2, \dots, N\}$  is a set of  $p$  integers with  $i_1 < i_2 < \dots < i_p$ 
  - $\mathcal{C}_{p,N}$  is set of all  $\binom{N}{p}$  such integer sets
  - $G[\mathbf{i}_p]$  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  $t_{\text{ind}}(S, G_N)$  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  $S$ :

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

## Induced Subgraph Density (Examples)

- $t_{\text{ind}}(\triangle, \square) = \frac{2}{4}$ ,  $t_{\text{ind}}(\wedge, \square) = \frac{2}{4}$  and  $t_{\text{ind}}(\cdot \diagup, \square) = \frac{0}{4}$
- $t_{\text{ind}}(\triangle, \blacksquare) = \frac{1}{4}$ ,  $t_{\text{ind}}(\wedge, \blacksquare) = \frac{2}{4}$  and  $t_{\text{ind}}(\cdot \diagup, \blacksquare) = \frac{1}{4}$

## Induced Subgraph Density: Graphon Case

- Let  $h(U_i, U_j)$  be a valid graphon.
- $\text{iso}(S)$  is the group of isomorphisms of  $S$ , and  $|\text{iso}(S)|$  its cardinality
- Under the “Aldous-Hoover DGP” the *ex ante* probability that an induced p-subgraph is isomorphic to  $S$  is given by

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

## Graph Limits

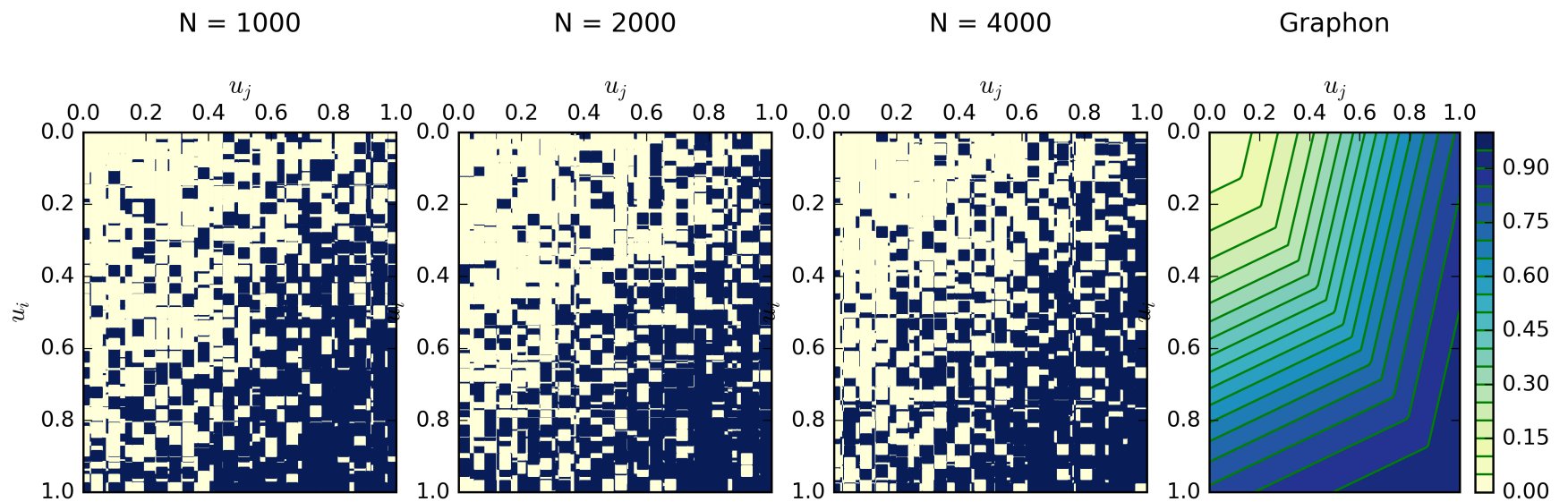
Let  $\{G_N\}_{N=1}^{\infty}$  be a sequence of networks. If

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

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

- Lovász (2012) for complete development
- Diaconis and Janson (2008) for connections with Aldous-Hoover Theorem

## 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  $\varphi : \mathcal{V}(S) \rightarrow \mathcal{V}(G_N)$  that preserve edge adjacency

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

## Homomorphism Density (continued)

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

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

- Can use  $t_{\text{hom}}(S, G_N)$  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 subgraph of  $G_N[\mathbf{i}_p]$ , for  $\mathbf{i}_p$  chosen uniformly at random from  $C_{p,N}$ , is isomorphic to  $S$
- If  $\lim_{N \rightarrow \infty} P_N(S) = t_{\text{ind}}(S, h)$  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:  $T \cup U = G(\mathcal{V}(T) \cup \mathcal{V}(U), \mathcal{E}(T) \cup \mathcal{E}(U))$
- Let  $W_{q,R,S}$  be a union of two isomorphisms, respectively  $T$  and  $U$ , of the graphlets  $R$  and  $S$  with
  1.  $|\mathcal{V}(R)| = |\mathcal{V}(S)| = p$
  2.  $|\mathcal{V}(R) \cap \mathcal{V}(S)| = q$  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  $\mathcal{W}_{q,S,S} = \mathcal{W}_{q,S}$ )

## Graphlet Stitching: Example #1

- Let the graphlets  $R = \text{---}$  and  $S = \text{---}$  share one vertex in common.
  - There is just one possible way to join them:  $R \cup S \cong \text{^}$
- We therefore have that  $\mathcal{W}_{1, \text{---}} = \{ \text{^} \}$

## Graphlet Stitching: Example #1 (continued)

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

$$\begin{aligned} Q\left(\mathcal{W}_{1, \text{---}}\right) &= \sum_{W \in \mathcal{W}_{1, \text{---}}} Q(W) \\ &= Q\left(\text{---}\right) \\ &= \mathbb{E}[D_{12}D_{13}] \end{aligned}$$

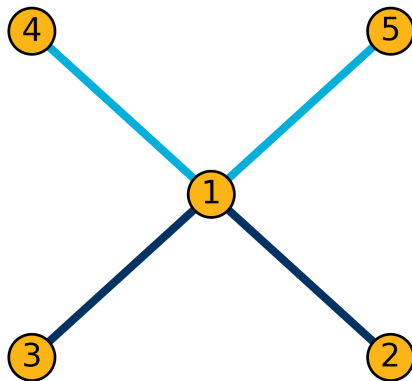
with  $Q(W)$  the homomorphism density introduced above

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

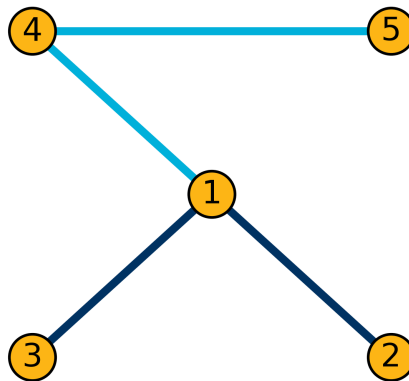
## Graphlet Stitching: Example #2

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

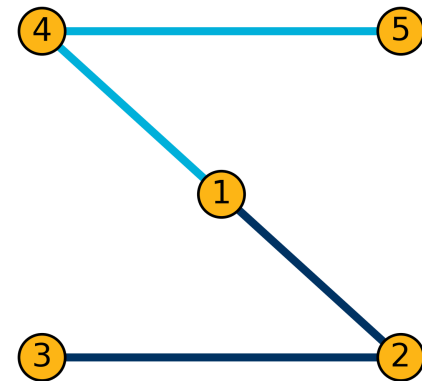
4-Star (1)



Tailed 3-Star (4)



5-Path (4)

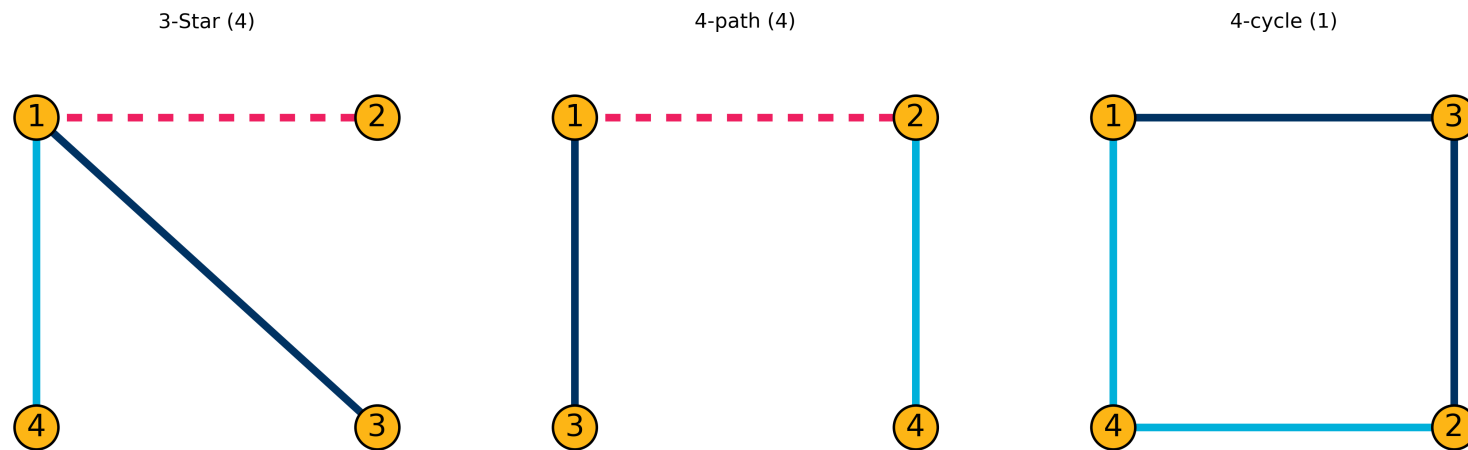


**Notes:** Number of isomorphisms of each graphlet in  $\mathcal{W}_{1,q}$  given in parentheses.





## Graphlet Stitching: Example #2 (continued)

There are nine ways (three up to isomorphisms) to join the graphlets  $R = \text{triangle}$  and  $S = \text{triangle}$ , 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

## Density

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

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

- Projecting onto  $U_1, \dots, U_N$  yields the decomposition:

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

- Observe that  $T_N$  is mean independent of  $U_N$

## Density: Variance Calculation

We have

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

A Hoeffding (1948) variance decomposition gives

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

for

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

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

## Density: Variance Calculation (continued)

Evaluating  $\Omega_1$  yields

$$\begin{aligned}
 \Omega_1 &= \mathbb{E} [h_N (U_1, U_2) h_N (U_1, U_3)] - \mathbb{E} [h_N (U_1, U_2)] \mathbb{E} [h_N (U_1, U_3)] \\
 &= Q \left( \mathcal{W}_{1, \text{---}} \right) - P \left( \text{---} \right) P \left( \text{---} \right) \\
 &= Q \left( \text{^} \right) - P \left( \text{---} \right) P \left( \text{---} \right)
 \end{aligned}$$

Evaluating  $\Omega_2$  yields

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

## Density: Variance Calculation (continued)

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

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

## Density: Variance Calculation (continued)

Collecting terms we have:

$$\begin{aligned}
 \mathbb{V}(\hat{\rho}_N) &= \frac{4(N-2)}{N(N-1)} \left[ Q \left( \text{---} \begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \bullet \quad \bullet \end{array} \right) - P \left( \text{---} \bullet \text{---} \bullet \right) P \left( \text{---} \bullet \text{---} \bullet \right) \right] \\
 &\quad + \frac{2}{N(N-1)} \mathbb{V}(\mathbb{E}[D_{12} | \mathbf{U}]) + \frac{2}{N(N-1)} \mathbb{E}[\mathbb{V}(D_{12} | \mathbf{U})] \\
 &= \frac{4(N-2)}{N(N-1)} \left[ Q \left( \text{---} \begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \bullet \quad \bullet \end{array} \right) - P \left( \text{---} \bullet \text{---} \bullet \right) P \left( \text{---} \bullet \text{---} \bullet \right) \right] \\
 &\quad + \frac{2}{N(N-1)} P \left( \text{---} \bullet \text{---} \bullet \right) \left( 1 - P \left( \text{---} \bullet \text{---} \bullet \right) \right)
 \end{aligned}$$

## Density: Variance Calculation (continued)

- To allow for graph sequences where  $\rho_N \rightarrow 0$  as  $N \rightarrow \infty$  we normalize

– Let  $\tilde{Q}(\text{triangle}) = \frac{Q(\text{triangle})}{\rho^2}$  and  $\tilde{P}(\text{edge}) = \frac{P(\text{edge})}{\rho_N}$

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



## Density: Variance Calculation (continued)

- After normalization:

$$\begin{aligned}\mathbb{V}\left(\frac{\hat{\rho}_N}{\rho_N}\right) &= \frac{4(N-2)}{N(N-1)} \left[ \tilde{Q}\left(\begin{array}{c} \bullet \\ \diagup \quad \diagdown \\ \bullet \quad \bullet \end{array}\right) - \tilde{P}\left(\begin{array}{c} \bullet \text{---} \bullet \end{array}\right) \tilde{P}\left(\begin{array}{c} \bullet \text{---} \bullet \end{array}\right) \right] \\ &\quad + \frac{2}{N\lambda_N} \tilde{P}\left(\begin{array}{c} \bullet \text{---} \bullet \end{array}\right) - \frac{2}{N(N-1)} \tilde{P}\left(\begin{array}{c} \bullet \text{---} \bullet \end{array}\right)^2 \\ &= O\left(\frac{1}{N}\right) + O\left(\frac{1}{N\lambda_N}\right) + O\left(\frac{1}{N^2}\right)\end{aligned}$$

- If  $\lambda_N \rightarrow \infty$  first term dominates
- If  $\lambda_N \rightarrow \lambda_0 > 0$ , first two terms dominate

## Asymptotic Inference

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

$$\sqrt{N} \left( \frac{\hat{\rho}_N}{\rho_N} - 1 \right) \xrightarrow{D} \mathcal{N} \left( 0, 4 \left[ \tilde{Q} \left( \text{triangle} \right) - \tilde{P} \left( \text{edge} \right) \tilde{P} \left( \text{edge} \right) \right] \right)$$

- Result due to Bickel, Chen and Levina (2011, *Annals of Statistics*)
- Comment: Under Erdos-Renyi  $\tilde{Q} \left( \text{triangle} \right) = \tilde{P} \left( \text{edge} \right) \tilde{P} \left( \text{edge} \right)$

## Variance Estimation

We can estimate the asymptotic variance using the analog estimators:

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

and

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

## Variance Estimation for $\hat{P}(\text{---})$ : Nyakatoke

For Nyakatoke we have

$$\hat{Q}(\text{^V}) \cong 0.006105$$

and

$$\hat{P}(\text{---}) \simeq 0.0698$$

which gives

$$\begin{matrix} \hat{\rho}_N \\ \text{(a.s.e)} \end{matrix} = \begin{matrix} 0.0698 \\ (0.0064) \end{matrix}$$

Note: Estimate of excluded higher order variance term is  $4.6 \times 10^{-6}$

## Limit Distribution of $\hat{P}(\text{triangle})$

- Define the multiset of graphlet stitchings:

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

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

- Normalize graphlet according to number of edges in it.

$$- \tilde{P}(\text{triangle}) = \frac{P(\text{triangle})}{\rho_N^2} \text{ and } \tilde{Q}(\mathcal{W}_{1, \text{triangle}}) = \frac{Q(\mathcal{W}_{1, \text{triangle}})}{\rho_N^4}$$

## Limit Distribution of $\hat{P}(\text{triangle})$

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

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

- 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., almost none) empirical work using these results
- Tremendous scope for using these methods in empirical analysis; but not easy!