Graph Limits & Subgraph Counts

Econometric Methods for Social Spillovers and Networks

University of St. Gallen, September 28th to October 6th, 2020

Bryan S. Graham

University of California - Berkeley

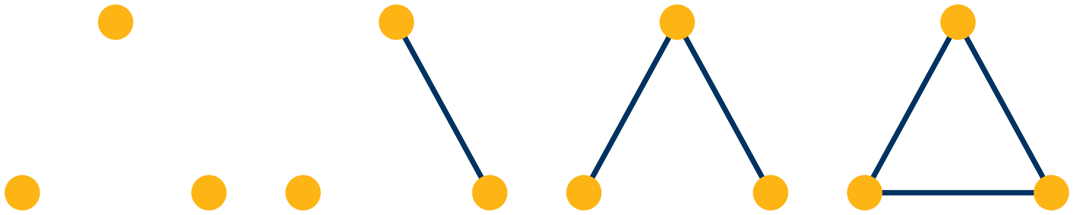
Introduction

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



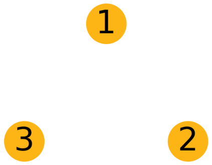
Empty

One Edge

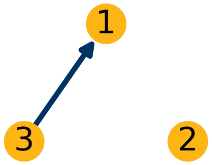
Two Star

Triangle

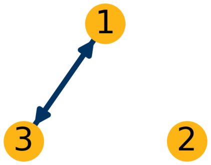
Triads: Directed Case



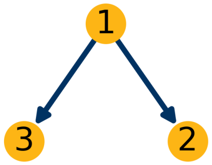
003



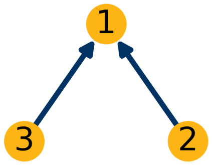
012



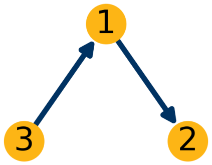
102



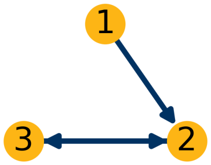
021D



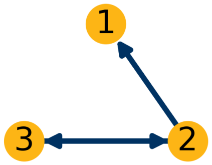
021U



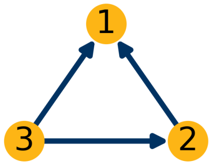
021C



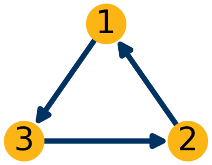
111D



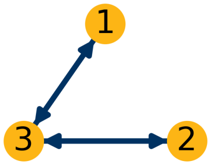
111U



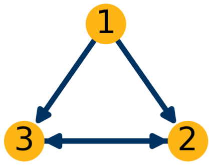
030T



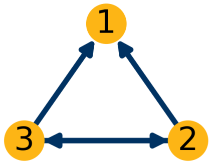
030C



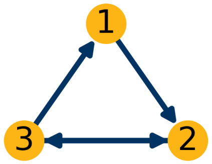
201



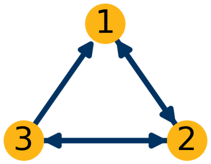
120D



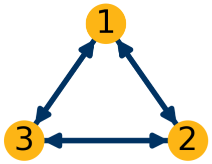
120U



120C



210



300

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., 2018).

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

- agents/vertices $\mathcal{V}=\{1,\ldots,N\}$,

- links/edges $\mathcal{E}=\left\{\{i,j\},\{k,l\},\ldots\right\}$, and

- adjacency matrix $\mathbf{D}=\left[D_{ij}\right]$ with

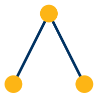
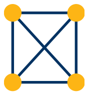
$$D_{ij}=\begin{cases}1 & \text{if } \{i,j\}\in\mathcal{E}\\ 0 & \text{otherwise}\end{cases}$$

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."

P-Cycles

A p-*cycle* is p^{th} order graphlet with nodes labeled (or relabeled) such that its edges form a cycle:

$$\mathcal{E}\left(S\right) = \left\{\left(i_1, i_2\right), \left(i_2, i_3\right), \ldots, \left(i_p, i_1\right)\right\}.$$

A p-cycle is a connected graphlet with p edges on p nodes.

Examples: triangles ($S =$ △) and 4-cycles ($S =$ □).

Trees

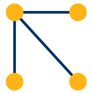
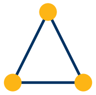
A *tree* is a connected graph with no cycles.

The number of edges on a p^{th} order tree is $p-1$; a feature which will prove highly convenient.

Examples: p-star graphlets, such as two-stars ($S =$) and three-stars ($S =$).

Also called connected acyclic graphs.

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
  - $G[\mathbf{i}_p]$ is the induced subgraph of G associated with vertex set $\mathbf{i}_p$

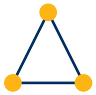
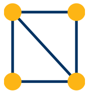
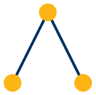
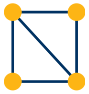
Induced Subgraph Density (continued)

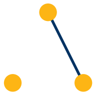
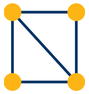
- The *induced subgraph density* of S in G_N, denoted by $t_{\text{ind}}(S, G_N)$ or $P_N(S)$ equals the probability that $G_N[\mathbf{i}_p]$, for $\mathbf{i}_p$ chosen uniformly at random from $C_{p,N}$, is isomorphic to S:

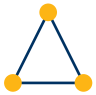
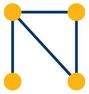
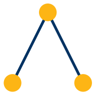
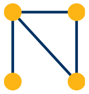
$$\begin{aligned} t_{\text{ind}}(S, G_N) &= \binom{N}{p}^{-1} \sum_{\mathbf{i}_p \in C_{p,N}} \mathbf{1}(S \cong G_N[\mathbf{i}_p]) \\ &= \Pr(S = G_N[\mathbf{i}_p]) \\ &= P_N(S) \end{aligned}$$

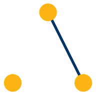
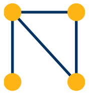
- Slightly different definition used in some of the technical literature...(see *Handbook* chapter)

Induced Subgraph Density (Examples)

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

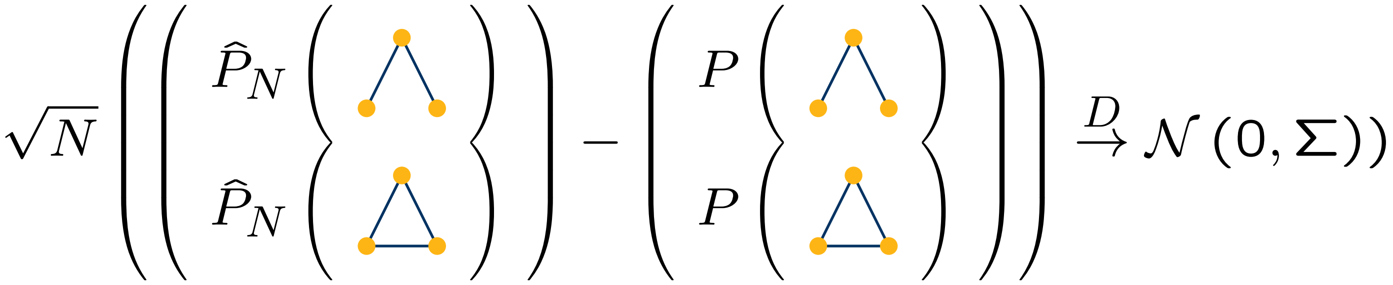
  and $t_{\text{ind}}($  ,  $) = \frac{0}{4}$

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

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

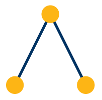
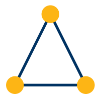
Goal

We would like a result of the form...



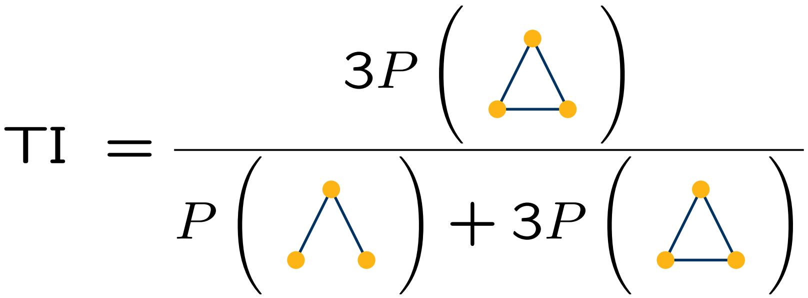
...under conditions we can understand

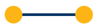
...with a covariance Σ we can estimate

...and interpretable limit values P () and P ()

Goal (continued)

With this result we can conduct inference on *transitivity*...



Is TI $> P\Big($  $\Big)$ (see Jackson et al. (2012) for some motivation)?

cf., Blitzstein and Diaconis (2011)

Induced Subgraph Density: Graphon Case

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

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

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

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

Graph Limits

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

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

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.

- Result establishes a connection between subgraph counts and the graphon.

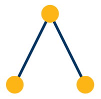
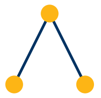
(Injective) Homomorphism Density

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 $\binom{N}{3}\Big|$iso()$\Big| = \frac{3}{6}N(N-1)(N-2)$ (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:

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

Can also use $t_{\mathsf{hom}}(S,G_N)$ 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 $\lim\limits_{N\to\infty} P_N(S) = t_{\text{ind}}(S,h)$ for some graphon $h(\cdot,\cdot)$ and all fixed subgraphs S, then we say that G_N converges to $h(\cdot,\cdot)$.

Computation

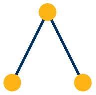
Useful to reformulate definition of $\widehat{P}_N(S)$.

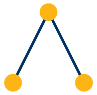
Let $\mathbf{D}_{[\mathbf{i}_p,\mathbf{i}_p]}$ be the $p \times p$ sub-adjacency matrix constructed by removing all rows and columns of $\mathbf{D}$ except those in $\mathbf{i}_p = \{i_1, \ldots, i_p\}$.

Let S be a graphlet of interest.

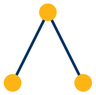
We can check for whether S is an isomorphism of $G[\mathbf{i}_p]$ by inspecting the elements of the $\mathbf{D}_{[\mathbf{i}_p,\mathbf{i}_p]}$ sub-adjacency matrix.

Computation (continued)

Consider the two star triad $S =$  , we can express $\mathbf{1}\left(S \cong G_N\left[\mathbf{i}_p\right]\right)$ in terms of $\mathbf{D}_{[\mathbf{i}_p,\mathbf{i}_p]}$ as

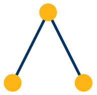
$\mathbf{1}\Bigg($  $\cong G_N\left[\mathbf{i}_3\right]\Bigg) = D_{i_1i_2}D_{i_1i_3}\left(1 - D_{i_2i_3}\right) + D_{i_1i_2}\left(1 - D_{i_1i_3}\right)D_{i_2i_3}$

$+ \left(1 - D_{i_1i_2}\right)D_{i_1i_3}D_{i_2i_3}$

$\overset{def}{\equiv} V$  $_{,\mathbf{i}_3}$

Computation (continued)

Let iso(S) be the group of isomorphisms of S, and $|\text{iso}(S)|$ its cardinality (i.e., number of subgraphs of K_p that are isomorphic to S).

We have $|\text{iso}($ $)| = 3$; three terms to the right of the (first) equality are indicators for three isomorphisms of  on $\{i_1, i_2, i_3\}$.

Computation (continued)

In general $\mathbf{1}\left(S \cong G_N\left[\mathbf{i}_p\right]\right)$ may be defined in terms of $\mathbf{D}_{[\mathbf{i}_p,\mathbf{i}_p]}$ with number of components equal to the number of possible isomorphisms of S.

There is only one isomorphism of the $\triangle$ configuration, yielding a second example of

$$\mathbf{1}\left(\triangle \cong G_N\left[\mathbf{i}_3\right]\right) = D_{i_1 i_2} D_{i_1 i_3} D_{i_2 i_3} \overset{def}{\equiv} V_{\triangle,\mathbf{i}_3}$$

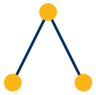
Unbiasedness

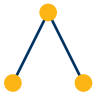
Two star configuration; iterated expectations and conditional independence of edges given $\mathbf{U} = (U_1, \ldots, U_N)'$ yields

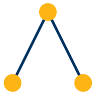
$$\begin{aligned}
\mathbb{E}\left[D_{i_1 i_2} D_{i_1 i_3}\left(1 - D_{i_2 i_3}\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[D_{i_1 i_2} D_{i_1 i_3}\left(1 - D_{i_2 i_3}\right)\middle| \mathbf{U}\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[D_{i_1 i_2} D_{i_1 i_3}\left(1 - D_{i_2 i_3}\right)\middle| U_{i_i}, U_{i_2}, U_{i_3}\right]\right] \\
&= \mathbb{E}\left[h\left(U_{i_1}, U_{i_2}\right) h\left(U_{i_1}, U_{i_3}\right)\left[1 - h\left(U_{i_2}, U_{i_3}\right)\right]\right]
\end{aligned}$$

Unbiasedness (continued)

Value of $\mathbb{E}\left[D_{i_1i_2}D_{i_1i_3}\left(1-D_{i_2i_3}\right)\right]$ is invariant to permutations of its indices.

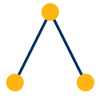
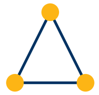
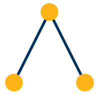
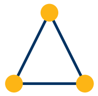
Recalling that $\Bigg|\text{iso}\Bigg($  $\Bigg)\Bigg| = 3$ we have

$\mathbb{E}\Bigg[\mathbf{1}\Bigg($  $\cong G_N\left[\mathbf{i}_p\right]\Bigg)\Bigg] = 3\cdot\int\int\int h\left(t,u\right)h\left(t,v\right)\left[1-h\left(u,v\right)\right]\mathrm{d}t\mathrm{d}u\mathrm{d}v$

$\overset{def}{\equiv} P\Bigg($  $\Bigg)$

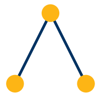
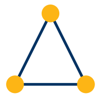
Large Sample Properties

Our estimator is

$$\left(\begin{array}{l} \widehat{P}_N \Big(\end{array}$$

$$\begin{array}{l} \Big) \\ \widehat{P}_N \Big(\end{array}$$

$$\begin{array}{l} \Big) \end{array} \right) = \binom{N}{3}^{-1} \sum_{i_1<i_2<i_3} \left(\begin{array}{l} V \end{array}$$

$$\begin{array}{l} {}_{,\mathbf{i}_3} \\ V \end{array}$$

$$\begin{array}{l} {}_{,\mathbf{i}_3} \end{array} \right).$$

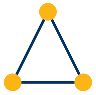
It is not a U-Statistics, but has many U-Statistic-like properties.

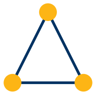
Large Sample Properties (continued)

It is unbiased for $\Bigg($ $P\Big($  $\Big)$ $P\Big($  $\Big)$ $\Bigg)$ under joint exchangeability (iterated expectations).

Can use Hoeffding (1948) arguments to study variance-covariance (cf., Holland and Leinhardt, 1976).

Network moments: Large N behavior

Projecting $\widehat{P}_N\Bigg($  $\Bigg)$ on $\mathbf{U} = (U_1, \ldots, U_N)'$ gives:

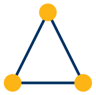
$\widehat{P}_N\Bigg($  $\Bigg) =$

$$
\begin{aligned}
&\binom{N}{3}^{-1} \sum_{i_1<i_2<i_3} h\left(U_{i_1}, U_{i_2}\right) h\left(U_{i_1}, U_{i_3}\right) h\left(U_{i_2}, U_{i_3}\right) \\
&+ \binom{N}{3}^{-1} \sum_{i_1<i_2<i_3} \left\{D_{i_1 i_2} D_{i_1 i_3} D_{i_2 i_3}\right. \\
&\left. -h\left(U_{i_1}, U_{i_2}\right) h\left(U_{i_1}, U_{i_3}\right) h\left(U_{i_2}, U_{i_3}\right)\right\}.
\end{aligned}
$$

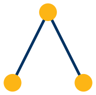
Second term is mean independent of first with conditionally independent summands.

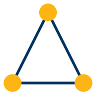
First term is a 3^{rd} order U-Statistic (large sample properties well-understood).

Network moments: Large N behavior (continued)

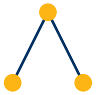
Under some conditions (most important of which is that average degree grows with N) $\widehat{P}_N\Big($  $\Big)$ behaves like a U-Statistic s.t.

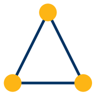
$\sqrt{N}\Bigg(\Bigg($

$\widehat{P}_N\Big($  $\Big)$

$\widehat{P}_N\Big($  $\Big)$

$\Bigg) - \Bigg($

$P\Big($  $\Big)$

$P\Big($  $\Big)$

$\Bigg)\Bigg) \overset{D}{\rightarrow} \mathcal{N}\left(0, 9\Sigma_1\right)$

...with Σ_1 estimable (analog estimate involves $O\left(N^5\right)$ operations!).

Use delta method to conduct inference on transitivity.

Intellectual history

Some basic ideas (e.g., use of Hoeffding-like variance decompositions) go back (at least) to Holland and Leinhardt (1976).

Subsequent work by Nowicki (1991), Picard et al. (2008) and others.

Big breakthrough by Bickel et al. (2011) – abstract (proof uses lots of "tricks") and limiting variance is not characterized.

Bhattacharya and Bickel (2015) – explicit characterization of variance and an estimator (cf., Menzel, 2017).

Some (interesting and empirically-relevant) subtleties ignored today.

Intellectual history (continued)

My exposition (anchored in textbook U-Statistic theory) is based on basic approach of Graham (2017).

Challenge is finding a notation that can neatly handle all cases.

Some open questions regarding sparse graph sequences.

Second (Simple) Example Density

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

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

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

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

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

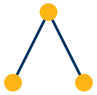
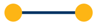
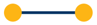
$$\Omega_q = \mathbb{C}\left(h_N\left(U_{i_1}, U_{i_2}\right), h_N\left(U_{j_1}, U_{j_2}\right)\right)$$

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

$$\Omega_1 = \mathbb{E}\left[h_N\left(U_1, U_2\right) h_N\left(U_1, U_3\right)\right] - \mathbb{E}\left[h_N\left(U_1, U_2\right)\right] \mathbb{E}\left[h_N\left(U_1, U_3\right)\right]$$

$= Q\Bigg($  $\Bigg) - P\Bigg($  $\Bigg) P\Bigg($  $\Bigg).$

Evaluating Ω_2 yields

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

Density: Variance Calculation (continued)

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

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

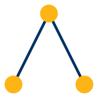
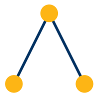
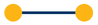
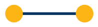
Density: Variance Calculation (continued)

Collecting terms we have:

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

Density: Variance Calculation (continued)

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

- Let $\tilde{Q}\Big($  $\Big) =$ $Q\Big($  $\Big) \Big/ \rho_N^2$ and $\tilde{P}\Big($  $\Big) =$ $P\Big($  $\Big) \Big/ \rho_N$.

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

Density: Variance Calculation (continued)

After normalization:

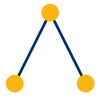
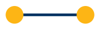
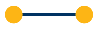
$$\begin{aligned}\mathbb{V}\left(\frac{\widehat{\rho}_N}{\rho_N}\right) &= \frac{4\,(N-2)}{N\,(N-1)}\left[\tilde{Q}\left(\bullet\!\!\overset{\bullet}{\wedge}\!\!\bullet\right)-\tilde{P}\left(\bullet\!-\!\bullet\right)\tilde{P}\left(\bullet\!-\!\bullet\right)\right]\\ &+\frac{2}{N\lambda_N}\tilde{P}\left(\bullet\!-\!\bullet\right)-\frac{2}{N\,(N-1)}\tilde{P}\left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

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

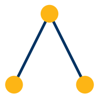
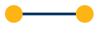
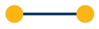
- If $\lambda_N \to \lambda_0 > 0$, first two terms dominate.

Asymptotic Inference

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

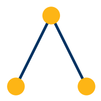
$\sqrt{N}\left(\frac{\widehat{\rho}_N}{\rho_N} - 1\right) \overset{D}{\to} \mathcal{N}\Bigg(0, 4\Bigg[\tilde{Q}\Bigg($  $\Bigg) - \tilde{P}\Bigg($  $\Bigg)\tilde{P}\Bigg($  $\Bigg)\Bigg]\Bigg).$

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

Comment: Under Erdos-Renyi $\tilde{Q}\Bigg($  $\Bigg) = \tilde{P}\Bigg($  $\Bigg)\tilde{P}\Bigg($  $\Bigg).$

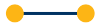
Variance Estimation

We can estimate the asymptotic variance using the analog estimators:

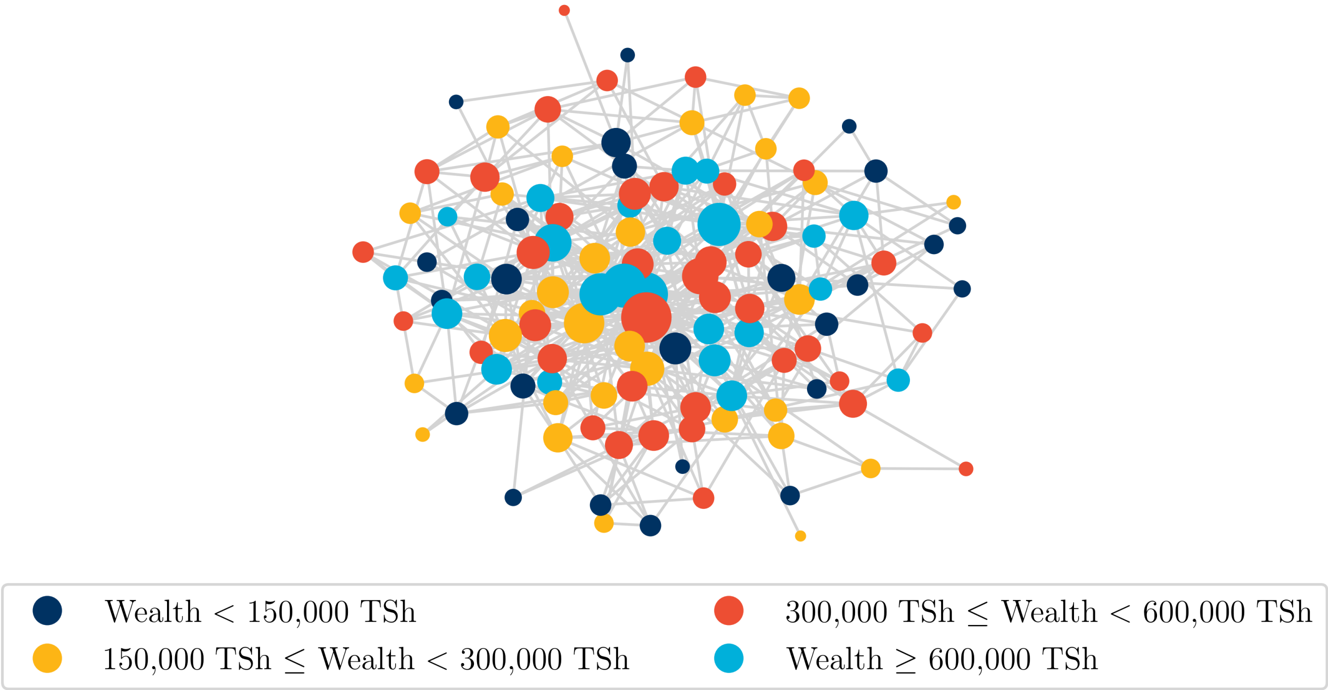
$\hat{Q}\Bigg($  $\Bigg) = \binom{N}{3}^{-1} \sum_{i<j<k} \frac{1}{3} \left\{ D_{ij}D_{ik} + D_{ij}D_{jk} + D_{ik}D_{jk} \right\}$

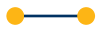
$$= \binom{N}{3}^{-1} \frac{1}{3} \left[T_{\mathsf{TS}} + 3T_{\mathsf{T}} \right]$$

and

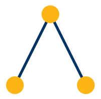
$\hat{P}\Bigg($  $\Bigg) = \binom{N}{2}^{-1} \sum_{i<j} D_{ij}$

Nyakatoke

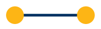

● Wealth < 150,000 TSh
● 150,000 TSh ≤ Wealth < 300,000 TSh
● 300,000 TSh ≤ Wealth < 600,000 TSh
● Wealth ≥ 600,000 TSh

Variance Estimation for $\widehat{P}\Bigg($  $\Bigg)$: Nyakatoke

For Nyakatoke we have

$\widehat{Q}\Bigg($  $\Bigg) \cong 0.006105$

and

$\widehat{P}\Bigg($  $\Bigg) \simeq 0.0698$

which gives

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

Note: Estimate above includes first two terms.

Standard Error Estimation for $\widehat{\mathrm{TI}}$: Nyakatoke

In Nyakatoke there are $\binom{119}{3} = 273,819$ triad configurations to count and a total of $\binom{119}{5} = 182,637,273$ pentads that need to be inspected in order to calculate variances.

Direct calculation gives

$$P_N(\triangle) = \begin{array}{c} 0.00115 \\ (0.00030) \end{array}, \; P_N(\wedge) = \begin{array}{c} 0.00496 \\ (0.00100) \end{array}$$

Standard Error Estimation for $\hat{\mathrm{TI}}$: Nyakatoke (continued)

Applying the delta method we get

$$\hat{\mathrm{TI}} = \begin{array}{c} 0.188 \\ (0.011) \end{array}$$

which suggests that transitivity is greater than what we would expect to observe under the Erdös-Renyi random graph null.

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!