

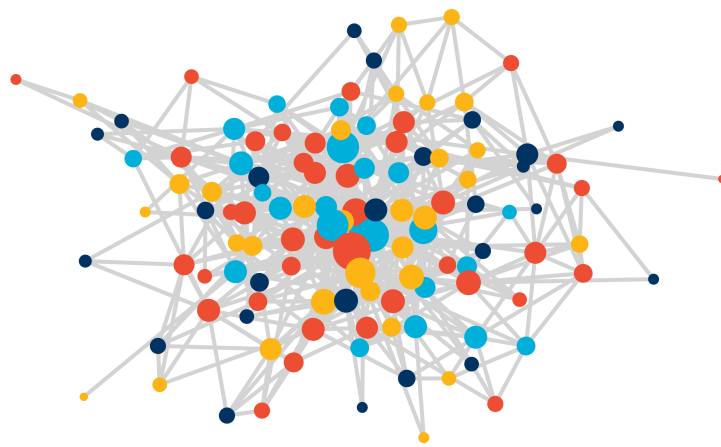
# **Describing Social Networks**

**Econometric Methods for Networks,  
GCEP, May 8th & 9th, 2017**

*Bryan S. Graham*

University of California - Berkeley

Nyakatoke Risk-Sharing Network

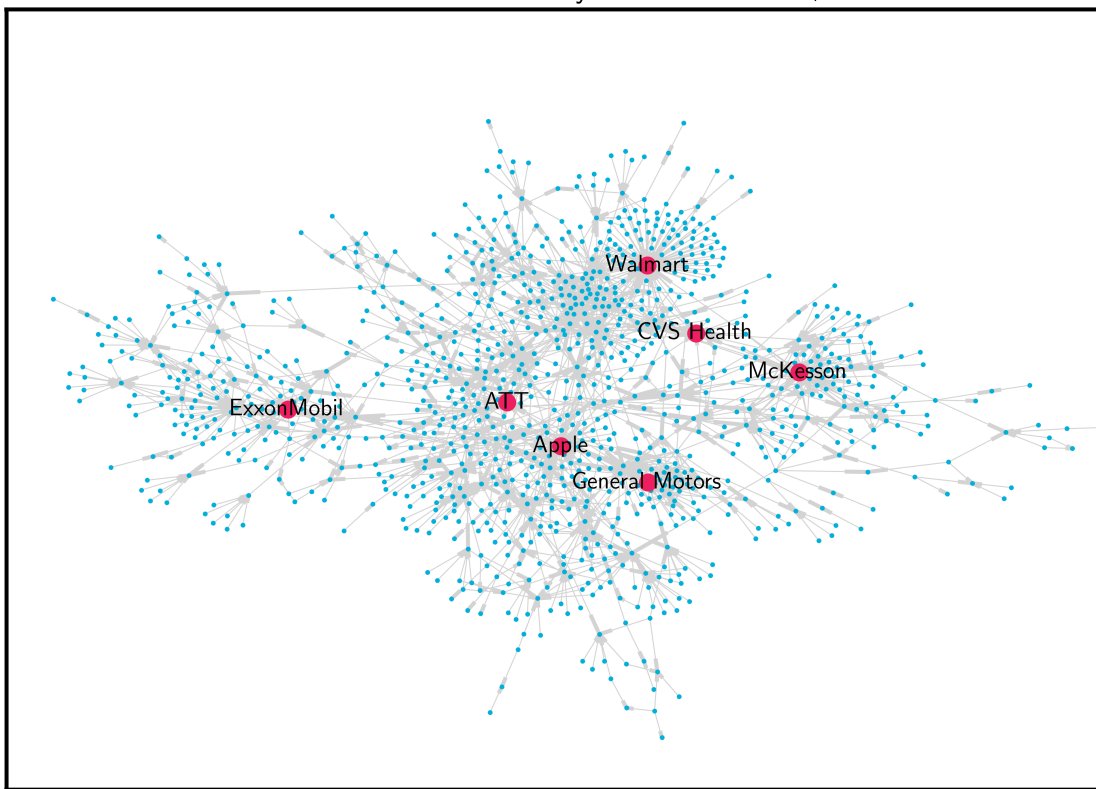


Note: node sizes are proportional to household degree



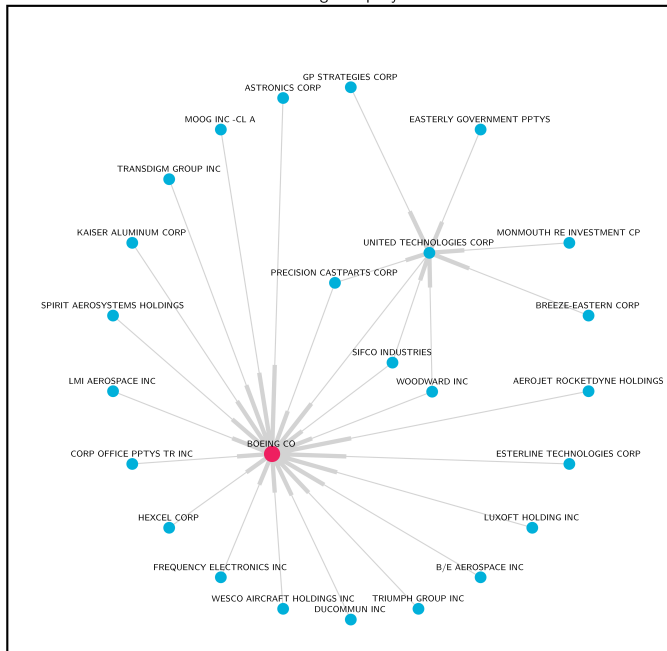
**(N = 119, n = 7,021)**

United States Inter-Firm Buyer-Seller Network, 2015

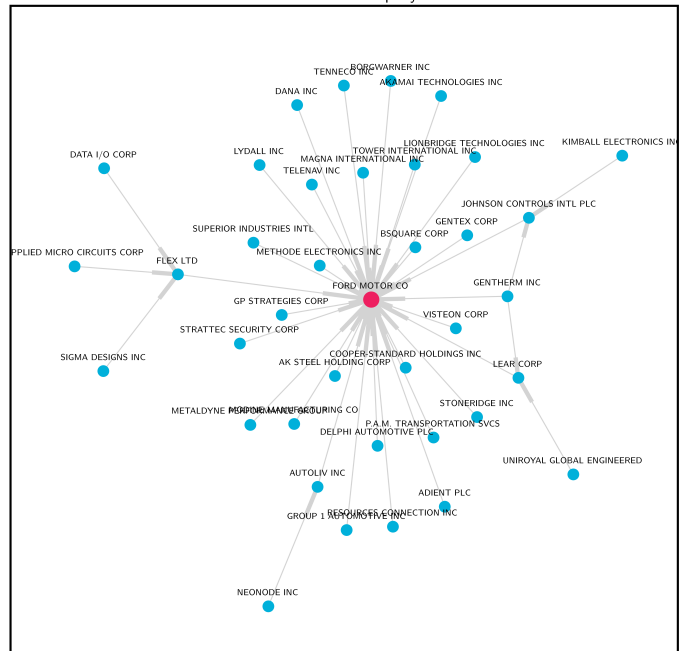


Source: Compustat - Capital IQ and author's calculations.  
Raw data available at <https://wrds-web.wharton.upenn.edu/wrds/> (Accessed January 2017)

Boeing Company



Ford Motor Company



## Questions

- How do the number, structure and characteristics of an agent's ties influence her behaviors and outcomes?
- How are ties formed? Are externalities involved?
- What configuration of ties would a social planner choose?
  - How does this idealized network compare with the observed one?
  - Are observed networks efficient?

## Questions (continued)

- Can we identify “important” agents in the network? Why is this interesting?
- What policies influence network structure (and outcomes)?
- How does network structure influence the diffusion of disease, ideas and new technologies?
- Are there optimal locations on a network in which to intervene?

## **Applications...**

- Buyer-supplier networks (Industrial Organization)
- Friendship networks (Education, Labor)
- Criminal networks (Urban)
- Trading networks (Industrial Organization and International Trade)
- Political networks (Political Economy)
- Bank networks (Finance)
- Online networks

### **...and Funding!**

- SBE Directorate of the National Science Foundation (NSF) recently identified network analysis as one of five key “cross cutting themes” with special grant opportunities.



## Literatures

- Psychology, sociology, anthropology, political science and economics all have empirical and theoretical literatures on “networks”.
  - Wasserman & Faust (1994)
  - Jackson (2008)
- Networks are widely-studied in Physics.
  - Newman (2010)

## Literatures

- The mathematical representation of networks as graphs makes discrete math (esp. graph theory), matrix analysis, and computer science highly useful.
- The statistical/econometric literature *very* underdeveloped (cf., Goldenberg *et al.* 2009)
- ...but growing rapidly (e.g., Bickel & Chen, 2009; Bickel, Chen & Levina, 2011; Graham, 2017; de Paula, 2016).
- A older and rich applied probability literature
  - Lovasz (2012)

## Outline of Course

- Lecture 1a (5/8/17, AM): Describing social networks
  - introduction to network data
  - definition and computation of basic summary network statistics
- Lecture 1b (5/8/17, AM): Shock propagation
  - centrality, PageRank, social multiplier
  - network structure and aggregate volatility

## Outline of Course (continued)

- Lecture 2a (5/8/17, PM): Nonparametrics
  - graphons, graph limits
  - nonparametric estimation of link probabilities
- Lecture 2b (5/8/17, PM): Inference
  - network moments
  - network subsampling/bootstrap

## Outline of Course (continued)

- Lecture 3a (5/9/17, AM): Link formation #1
  - importance sampling from networks w/ fixed degree
  - dyadic models of link formation
- Lecture 3b (5/9/17, AM): Link formation #2
  - network formation w/ interdependencies
  - strategic models

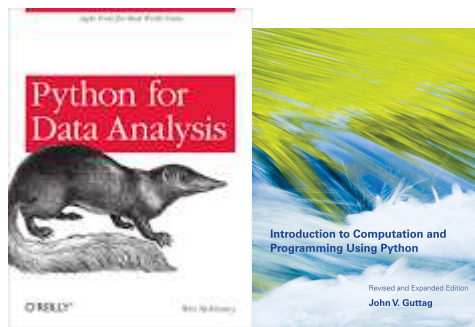
## Outline of Course (continued)

- Lecture 4a (5/9/17, PM): Peer effects
  - network structure & peer effects
  - neighborhood effects
- Lecture 4b (5/9/17, PM): Catch-up, general questions & answers

## Computation

- Some computational illustrations in class
- Most code is available on the course GitHub repository ([https://github.com/bryangraham/short\\_courses](https://github.com/bryangraham/short_courses))
- If you want to follow along (recommend, but not required) use the *Anaconda* distribution of Python v 2.7.12 <https://www.continuum.io/downloads>
- Includes key packages for data analysis & scientific computing (e.g., numpy, scipy, pandas, networkx)
- Also useful: Graphviz (visualization), Yhat Rodeo (IDE)

## Computation (continued)



<https://github.com/wesm/pydata-book>



<http://quant-econ.net/>



## Basic Terms & Notation

- An **undirected graph**  $G(\mathcal{N}, \mathcal{E})$  consists of a set of **nodes**  $\mathcal{N} = \{1, \dots, N\}$  and a list of unordered pairs of nodes called **edges**  $\mathcal{E} = \{\{i, j\}, \{k, l\}, \dots\}$  for  $i, j, k, l \in \mathcal{N}$ .
- A graph is conveniently represented by its **adjacency matrix**  $\mathbf{D} = [D_{ij}]$  where

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

- No self-ties & unordered edges  $\Rightarrow \mathbf{D}$  is a symmetric binary matrix with a diagonal of so-called structural zeros.
- vertex: node, agent or player.
- edges: links, friendships, connections or ties.

## Basic Terms & Notation (continued)

$$D = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

- Agent 1 is connected to agents 2 and 5.
- Agent 2 is connected to agent 1.
- Agent 3 is connected to no one.
- Agent 4 is connected to agent 5.

## Basic Terms & Notation (continued)

$$D = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

- Agent 5 is connected to agents 1 and 4.
- Agents 2 and 5 are indirectly connected through agent 1 (i.e., share her as a common friend).
- Agents 2 and 4 are indirectly connected through agents 1 and 5.
- 3 out of 10 possible ties are present in the network.

## Agents, Dyads and Triads

- A network consists of
  - $N$  agents
  - $\binom{N}{2} = \frac{1}{2}N(N-1) = O(N^2)$  pairs of agents or **dyads**.
  - $\binom{N}{3} = \frac{1}{6}N(N-1)(N-2) = O(N^3)$  triples of agents of **triads**.
  - $\binom{N}{4} = \frac{1}{24}N(N-1)(N-2)(N-3) = O(N^4)$  quadruples of agents of **tetrads**.
- In summarizing a network adjacency matrix it is convenient to conceptualize statistics as measures of agent-, dyad-, triad- or p-subgraph-level attributes.

## Agent-level Statistics: Degree

- The total number of links belonging to agent  $i$ , or her **degree** is  $D_{i+} = \sum_j D_{ij}$ .
- The **degree sequence** of a network is  $\mathbf{D}_+ = (D_{1+}, \dots, D_{N+})'$ .
- The **degree distribution** gives the frequency of each possible agent-level degree count  $\{0, 1, \dots, N\}$  in the network.

## Degree (continued)

- Some researchers take the degree distribution as their primary object of interest (e.g., Barabási and Albert, 1999).
  - Other key topological features of a network are fundamentally constrained by its degree distribution.
- Some datasets report agent degrees with no other network information

## Dyad-level Statistics: Density

- Dyads are either linked or unlinked.
- The **density** of a network equals the frequency with which any randomly drawn dyad is linked:

$$P_N = \binom{N}{2}^{-1} \sum_{i=1}^N \sum_{j<i} D_{ij}. \quad (2)$$

- Note that  $\lambda_N = (N - 1) P_N$  coincides with average degree.
- The density of the Nyakatoke network is 0.0698.
- Low density and skewed degree distributions (with fat tails) are common features of real world social networks.

## Paths

$$\mathbf{D}^2 = \begin{pmatrix} D_{1+} & \sum_i D_{1i}D_{2i} & \cdots & \sum_i D_{1i}D_{Ni} \\ \sum_i D_{1i}D_{2i} & D_{2+} & \cdots & \sum_i D_{2i}D_{Ni} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_i D_{1i}D_{Ni} & \sum_i D_{2i}D_{Ni} & \cdots & D_{N+} \end{pmatrix}$$

- The  $i^{th}$  diagonal element of  $\mathbf{D}^2$  equals the number of agent  $i$ 's links or her degree.
- The  $\{i, j\}^{th}$  element of  $\mathbf{D}^2$  gives the number of links agent  $i$  has in common with agent  $j$  (i.e., the number of “friends in common”).



## Paths (continued)

- graph theory: the  $\{i, j\}^{th}$  element of  $\mathbf{D}^2$  gives the number of **paths** of length two from agent  $i$  to agent  $j$ .
- if  $i$  and  $j$  share the common friend  $k$ , then a length two path from  $i$  to  $j$  is given by  $i \rightarrow k \rightarrow j$ .

## Paths (continued)

$$\mathbf{D}^3 = \begin{pmatrix} \sum_{i,j} D_{1i} D_{ij} D_{j1} & \cdots & \sum_{i,j} D_{1i} D_{ij} D_{jN} \\ \vdots & \ddots & \vdots \\ \sum_{i,j} D_{1i} D_{ij} D_{jN} & \cdots & \sum_{i,j} D_{Ni} D_{ij} D_{jN} \end{pmatrix}$$

- $\{i, j\}^{th}$  element gives the number of paths of length 3 from  $i$  to  $j$ .
- If both  $i$  and  $j$  are connected to  $k$  as well as to each other, then the  $\{i, j, k\}$  triad is transitive (i.e., “the friend of my friend is also my friend”).

## Paths (continued)

- The  $i^{th}$  diagonal element  $\mathbf{D}^3$  is a count of the number of transitive triads or **triangles** to which  $i$  belongs (with  $i-j-k$  and  $i-k-j$  counted separately).
  - If  $\{i, j, k\}$  is a closed triad it is counted twice each in the  $i^{th}$ ,  $j^{th}$  and  $k^{th}$  diagonal elements of  $\mathbf{D}^3$ .
  - $\text{Tr}(\mathbf{D}^3) / 6$  equals the number of *unique* triangles in the network.

## K-Length Paths

- The  $\{i, j\}^{th}$  element of  $\mathbf{D}^K$  gives the number of paths of length  $K$  from agent  $i$  to agent  $j$ .
- Let  $D_{ij}^{(K)}$  denote the  $\{i, j\}^{th}$  element of  $\mathbf{D}^K$ .
- $\mathbf{D}^0 = I_N$ , the only zero length walks in the network are from each agent to herself.
- Under the maintained hypothesis,  $D_{ij}^{(K)}$  equals the number of  $K$ -length paths from  $i$  to  $j$ . The number of  $K + 1$  length paths from  $i$  to  $j$  then equals

$$\sum_{k=1}^N D_{ik}^{(K)} D_{kj},$$

which equals the  $\{i, j\}^{th}$  element of  $\mathbf{D}^{K+1}$ . The claim follows by induction.

## Distance

- The **distance** between agents  $i$  and  $j$  equals the minimum length path connecting them.
- If there is no path connecting  $i$  to  $j$ , then the distance between them is infinite.
- We can use powers of the adjacency matrix to calculate these distances:

$$M_{ij} = \min_k \left\{ k : D_{ij}^{(k)} > 0 \right\}$$

- If the network consists of a single, giant, connected component, we can compute average path length as

$$\overline{M} = \binom{N}{2}^{-1} \sum_{i=1}^N \sum_{j < i} M_{ij}. \quad (3)$$

## Small World Problem

Frequency of minimum path lengths in the Nyakatoke network

	1	2	3	4	5
Count	490	2666	3298	557	10
Frequency	0.0698	0.3797	0.4697	0.0793	0.0014

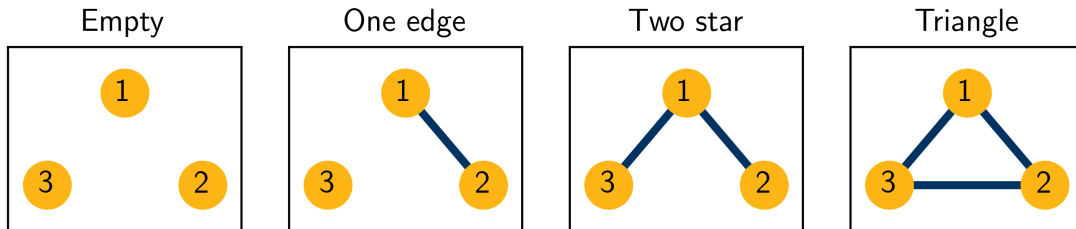
**Source:** de Weerd (2004) and author's calculations.

- Less than 7 percent of all *pairs* of households are directly connected in Nyakatoke.
- ...but over 40 percent dyads are no more than two degrees apart.
- ..and over 90 percent are separated by three or fewer degrees.

## Small World Problem (continued)

- **diameter:** largest distance between two agents.
- The diameter of the Nyakatoke network is 5.
- Small-world problem: why do we see sparsity and low diameter together (Milgram, 1967)?

## Triad Census



- **Triads**, a set of three unique agents, come in four types (isomorphisms):
  - no connections or **empties**
  - one connection or **one-edges**
  - two connections or **two-stars**
  - three connections or **triangles**
- A complete enumeration of them into their four possible types constitutes a *triad census*.



## Triad Census: Triangles

- Each agent can belong to as many as

$$(N - 1) (N - 2)$$

triangles.

- The counts of these triangles are contained in the  $N$  diagonal elements of  $\mathbf{D}^3$ .
- However each such triangle appears 6 times in these counts: as  $\{i, j, k\}$ ,  $\{i, k, j\}$ ,  $\{j, i, k\}$ ,  $\{j, k, i\}$ ,  $\{k, i, j\}$  and  $\{k, j, i\}$ . Thus

$$T_T = \frac{\text{Tr}(\mathbf{D}^3)}{6} \quad (4)$$

equals the total number of unique triangles in the network.

## Triad Census: Two-Stars

- Each dyad can share of up to  $N - 2$  links in common.
- These counts are contained in the lower (or upper) off-diagonal elements of  $\mathbf{D}^2$ .
- Each triad appears three times in these counts: as  $\{i, j, k\}$ ,  $\{i, k, j\}$  and  $\{j, k, i\}$ . If it is a
  - two star, then only one of  $D_{ji}D_{ki}$ ,  $D_{ij}D_{kj}$ , or  $D_{ik}D_{jk}$  quantities will equal one,
  - triangle, then all three will equal one.

## Two-Stars (continued)

- $\Rightarrow \text{vech}(\mathbf{D}^2)'_{\iota}$  gives the network count of *three times* the number triangles *plus* the number of two-stars, hence

$$T_{TS} = \text{vech}(\mathbf{D}^2)'_{\iota} - \frac{\text{Tr}(\mathbf{D}^3)}{2} \quad (5)$$

equals the number of two-star triads in the network

## Triad Census: One-Edges & Empties

- If *all* triads are empty or have only one edge, then there will be  $(N - 2) \text{vech}(\mathbf{D})_\iota$  one edge triads.
- If some triads are two-stars or triangles this count will be incorrect.
- Subtracting twice the number of two stars and three times the number of triangles gives the correct answer:

$$T_{OE} = (N - 2) \text{vech}(\mathbf{D})'_\iota \quad (6) \\ - 2 \text{vech}(\mathbf{D}^2)'_\iota + \frac{\text{Tr}(\mathbf{D}^3)}{2}$$

- The number of empty triads,  $T_E$ , equals  $\binom{N}{3}$  minus the total number of other triad types.

## Triad Census: Nyakatoke Network

	<b>empty</b>	<b>one-edge</b>	<b>two-star</b>	<b>triangle</b>
<b>Count</b>	221,189	48,245	4,070	315
<b>Proportion</b>	0.8078	0.1762	0.0149	0.0012
<b>Random</b>	0.8049	0.1812	0.0136	0.0003

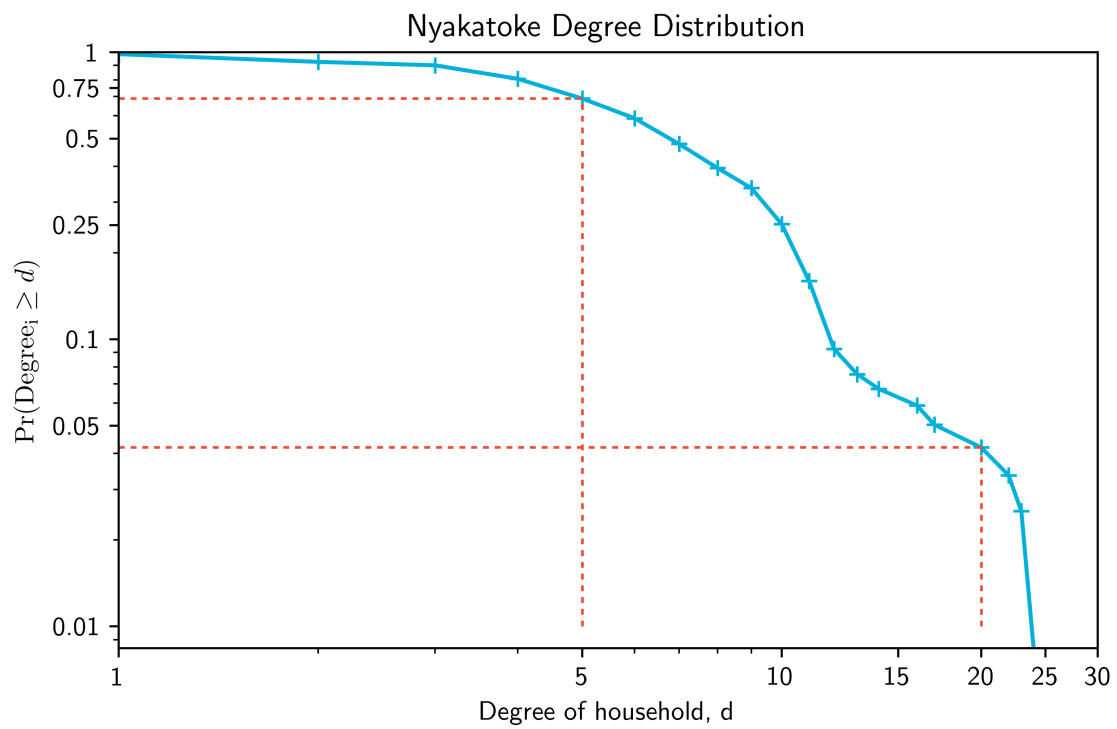
## Transitivity

- The **Transitivity Index**, sometimes called the clustering coefficient, is

$$\begin{aligned}\text{TI} &= \frac{3T_T}{T_{TS} + 3T_T} \\ &= \frac{1}{2} \frac{\text{Tr}(\mathbf{D}^3)}{\text{vech}(\mathbf{D}^2)'}\end{aligned}$$

- In random graphs TI should be close to network density.
- For the Nyakatoke network  $\text{TI} = 0.1884$  and  $P_N = 0.0698$ .
- Network transitivity *may*
  - facilitate risk sharing and other activities which require monitoring (cf., Jackson et al., 2012).

## Nyakatoke Degree Distribution



## Degree Distribution Redux

- Average degree equals  $\lambda_N = \left( \frac{2T_{OE} + 4T_{TS} + 6T_T}{N(N-2)} \right)$ .

- Degree variance equals

$$S_N^2 = \frac{2}{N} (T_{TS} + 3T_T) - \lambda_N [1 - \lambda_N].$$

- Knowledge of mean degree, degree variance and the number of triangles is equivalent to knowledge of the triad census.
- The degree distribution constrains other features of the network.
  - models of network formation should allow for arbitrary degree distributions.



## Power Laws

- Barabási and Albert (1999) assert that the degree distributions of many networks, at least over some range, follow discrete Pareto or ‘power law’ distributions (cf., Yule, 1929):

$$F(d_+) = 1 - \left(\frac{\underline{d}_+}{d_+}\right)^\alpha$$

for  $d_+ = \underline{d}_+, \dots, N$  and  $F(d_+) = \Pr(D_{i+} \leq d_+)$ .

- Here  $\underline{d}_+ > 0$  is some threshold degree level below which the power law distribution may not apply.

- Taking logs yields the linear relationship

$$\ln(1 - F(D_{i+})) = \alpha \ln(\underline{d}_+) + \alpha \ln D_{i+}.$$

The coefficient,  $\alpha$ , may be estimated by OLS (cf., Clauset, Shalizi and Newman, 2009).

## Wrapping Up

- Network data, as encapsulated by adjacency matrices are complex
  - rich combinatoric structure
  - strong dependencies across different statistics of  $\mathbf{D}$
- Researchers have motivated the various statistics reviewed here both formally and heuristically.
- ....methods of (frequentist) inference associated with the statistics reviewed here are still under development.
  - how do we compute a asymptotic standard error for average degree?