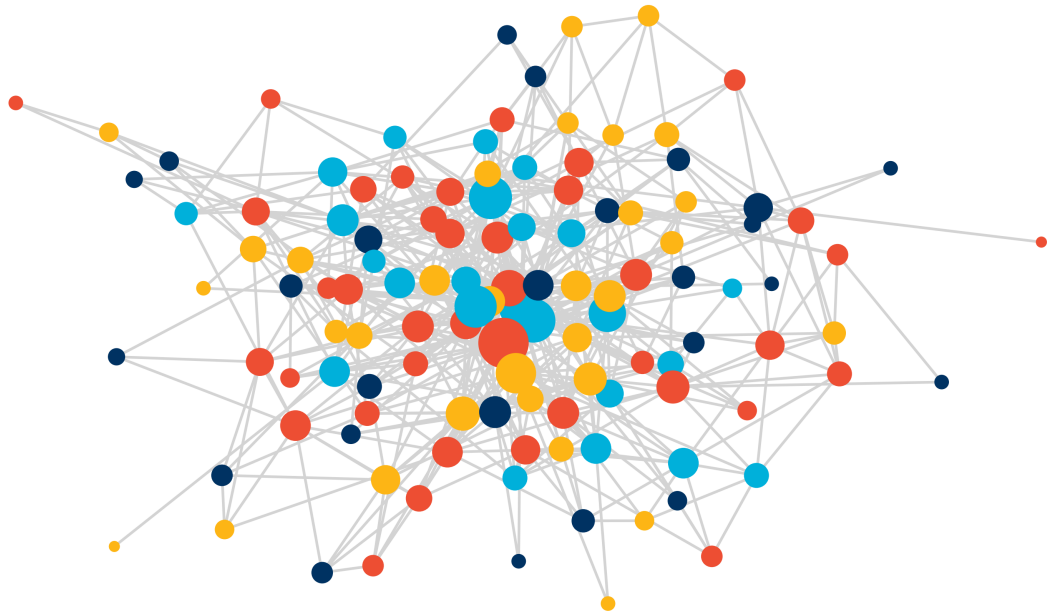
Describing Social Networks

Econometric Methods for Networks,

SMU, May 29th & June 1st, 2017

Bryan S. Graham

University of California - Berkeley

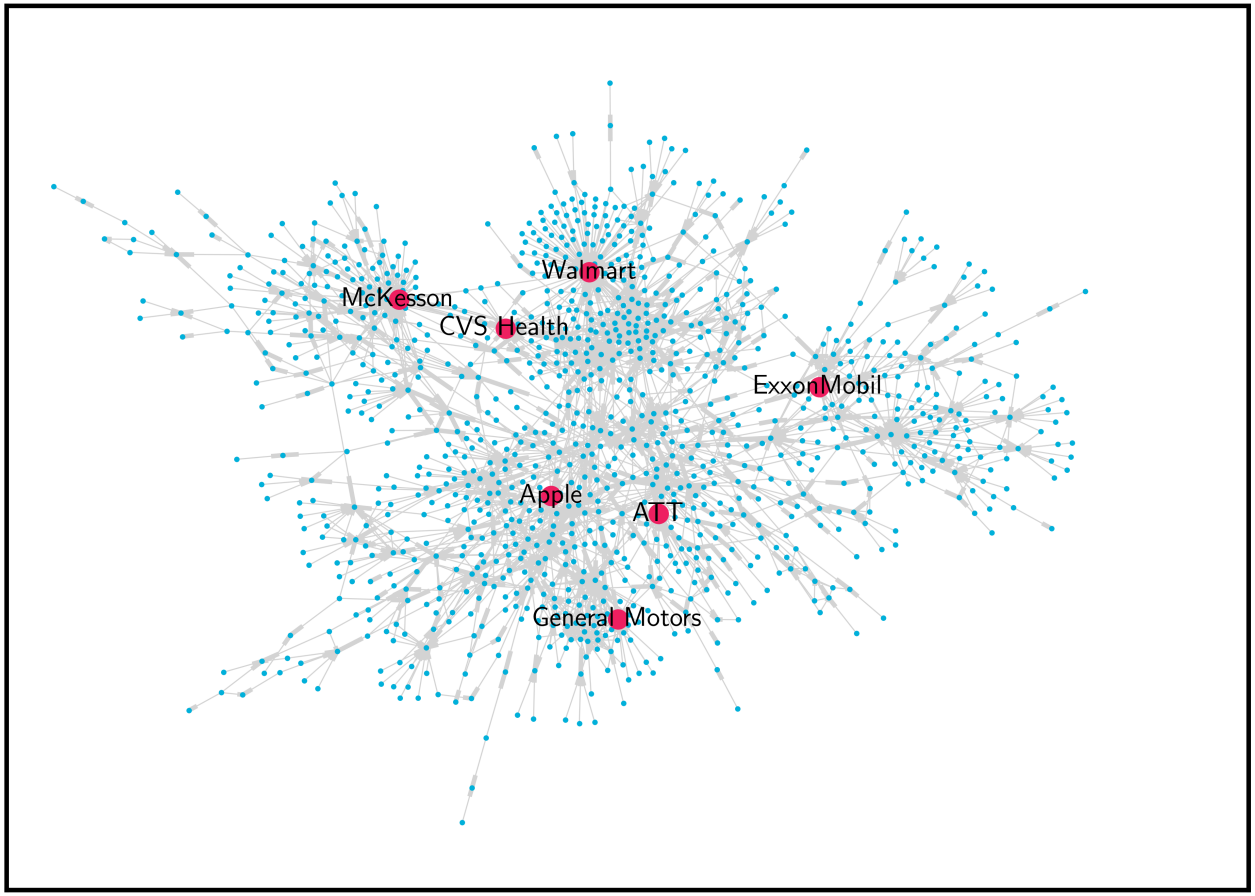


Note: node sizes are proportional to household degree

- Wealth $<$ 150,000 TSh
- 150,000 TSh $\leq$ Wealth $<$ 300,000 TSh
- 300,000 TSh $\leq$ Wealth $<$ 600,000 TSh
- Wealth $\geq$ 600,000 TSh

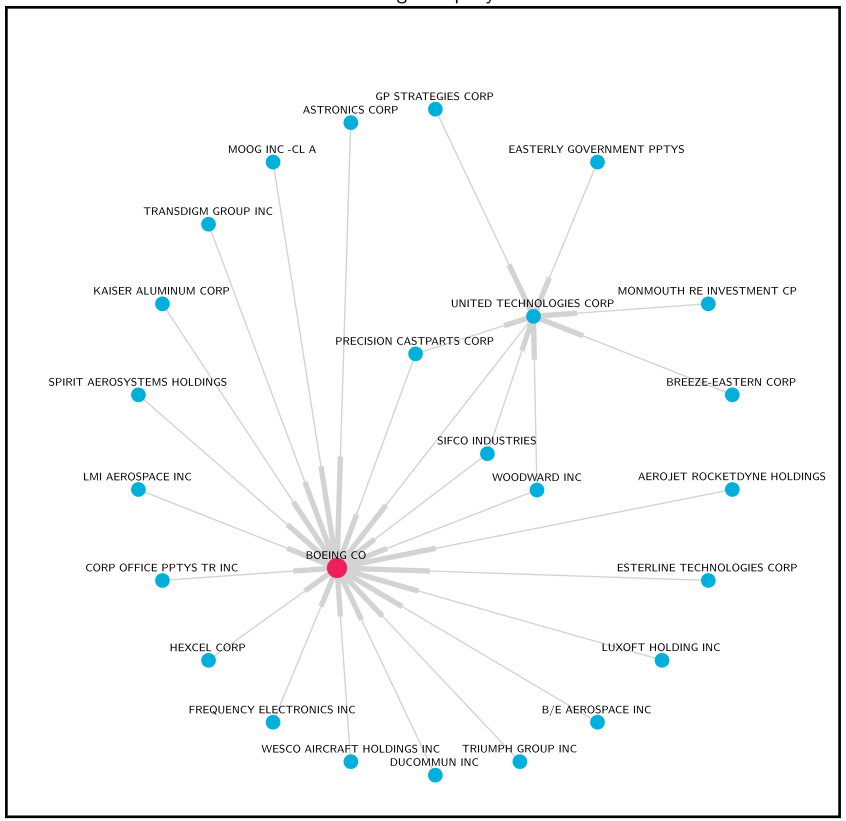
(N = 119, n = 7,021)

United States Inter-Firm Buyer-Seller Network, 2015

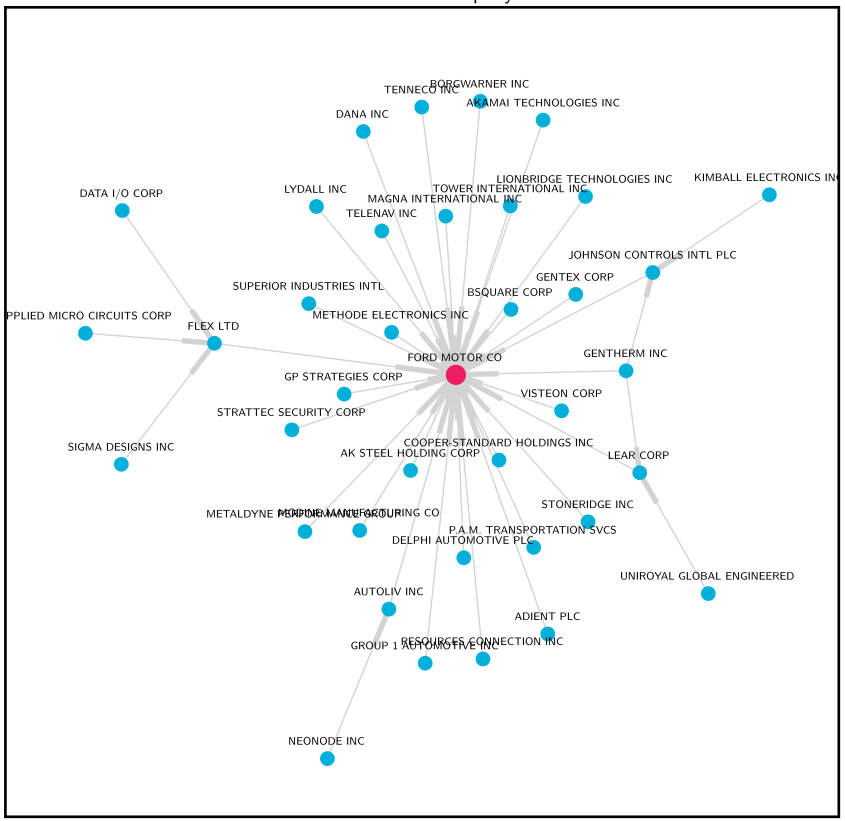


Source: Compustat - Capital IQ and author's calculations.
Raw data available at https://wrds-web.wharton.upenn.edu/wrds/ (Accessed January 2017)

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 ("key players") 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)

Applications... (continued)

- 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.

Outline of Course

- Lecture 1 (5/29/17, PM): Describing social networks
  - introduction to network data
  - definition and computation of basic summary network statistics
- Lecture 2 (5/30/17, AM): Centrality, spillovers and shocks
  - centrality, PageRank, social multiplier
  - networks and aggregate volatility

Outline of Course (continued)

- Lecture 3 (5/30/17, PM): Nonparametrics
  - graphons, graph limits
  - nonparametric estimation of link probabilities
- Lecture 4 (5/31/17, AM): Inference
  - network moments
  - importance sampling from networks w/ fixed degree

Outline of Course (continued)

- Lecture 5 (5/31/17, PM): Link formation
  - dyadic models of link formation
  - strategic and dynamic models
- Lecture 6 (6/1/17, AM): Peer effects
  - network structure & peer effects

Computation

- Some computational illustrations in class.

- Most code is available on the course GitHub repository (`https://github.com/bryangraham/short_courses`).

- If you want to follow along (recommended, but not required) use the *Anaconda* distribution of Python v 2.7.12 `https://www.continuum.io/downloads`.

Computation (continued)

- *Anaconda* 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

QuantEcon

http://quant-econ.net/

Basic Terms & Notation

- An **undirected graph** $G(\mathcal{N}, \mathcal{E})$ consists of a set of **nodes** $\mathcal{N} = \{1, \ldots, N\}$ and a list of unordered pairs of nodes called **edges** $\mathcal{E} = \{\{i, j\}, \{k, l\}, \ldots\}$ for $i, j, k, l \in \mathcal{N}$.

- A graph is conveniently represented by its **adjacency matrix** $\mathbf{D} = \left[D_{ij}\right]$ where

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

- No self-ties & unordered edges $\Rightarrow \mathbf{D}$ is a symmetric binary matrix with a diagonal of so-called structural zeros.

Basic Terms & Notation (continued)

- vertex: node, agent or player.
- edges: links, friendships, connections or ties.
- We will extend our framework to accommodate directed ties subsequently (e.g., as needed for buyer-supplier networks).

Basic Terms & Notation (continued)

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

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

Basic Terms & Notation (continued)

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

- 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).

Basic Terms & Notation (continued)

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

- 3 out of 10 possible ties are present in the network.

Agents, Dyads, Triads and Tetrads

- A network consists of
  - N agents
  - $\binom{N}{2} = \frac{1}{2}N(N-1) = O\left(N^2\right)$ pairs of agents or **dyads.**
  - $\binom{N}{3} = \frac{1}{6}N(N-1)(N-2) = O\left(N^3\right)$ triples of agents or **triads**.
  - $\binom{N}{4} = \frac{1}{24}N(N-1)(N-2)(N-3) = O\left(N^4\right)$ quadruples of agents or **tetrads**.

Agents, Dyads and Triads (continued)

In summarizing a network adjacency matrix it is convenient to conceptualize statistics as measures of

1. agent-,
2. dyad-,
3. triad- or
4. 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 $\mathbf{D}_+ = \left(D_{1+}, \ldots, D_{N+}\right)'$.

- The **degree distribution** gives the frequency of each possible agent-level degree count $\{0, 1, \ldots, N\}$ 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 (more on this below).
- 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:

$$\widehat{\rho}_N = \widehat{P}(\ \text{•—•}\ ) = \binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} D_{ij}. \qquad (2)$$

- Note that $\widehat{\lambda}_N = (N-1)\,\widehat{P}(\ \text{•—•}\ )$ coincides with average degree.

Dyad-level Statistics: Density (continued)

- 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 $\mathbf{D}^3$.
  - Tr $\left(\mathbf{D}^3\right)/6$ 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.

K-Length Paths (continued)

- 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.

- Agents separated by a finite distance are *connected*, otherwise they are *unconnected*.

Distance (continued)

- 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

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

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 Weerdt (2004) and author's calculations.

Small World Problem (continued)

- Less than 7 percent of all *pairs* of households are directly connected in Nyakatoke.

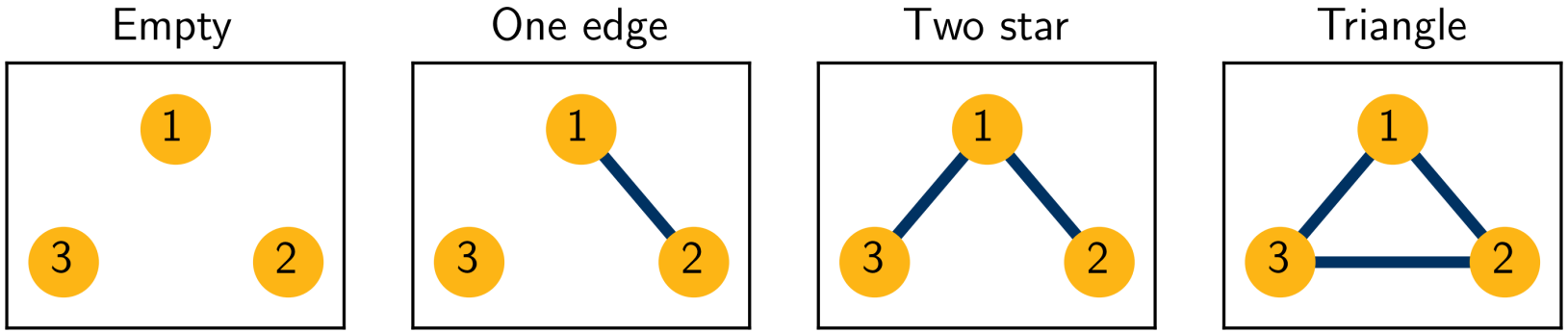
- ...but over 40 percent dyads are no more than two degrees apart.

- ..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



Triad Census (continued)

- **Triads**, sets 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

$$T_T = \frac{\mathrm{Tr}\left(\mathbf{D}^3\right)}{6} \tag{4}$$

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)

- We have that $\text{vech}\left(\mathbf{D}^2\right)'\iota$ gives the network count of *three times* the number triangles *plus* the number of two-stars.

- Therefore

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

  equals the number of two-star triads in the network.

Triad Census: One-Edges

- 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:

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

Triad Census: Empties

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

Triad Census: Nyakatoke Network

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

Transitivity

- The **Transitivity Index** (a.k.a. clustering coefficient) is

$$\text{TI} \quad = \quad \frac{3T_T}{T_{TS} + 3T_T}$$

- In random graphs TI should be close to network density (next slide).

- For the Nyakatoke network $\text{TI} = 0.1884$ and $\widehat{\rho}_N = 0.0698$.

- Network transitivity *may* facilitate risk sharing and other activities which require monitoring (cf., Jackson et al., 2012).

Transitivity (continued)

- Let $\rho_N = \Pr\left(D_{ij} = 1\right)$ with all edges forming independently.
- Probability that a randomly drawn triad takes a triangle configuration is ρ_N^3.
- Probability that a randomly drawn triad takes a two star configuration is

$$3 \times \rho_N^2 \left(1 - \rho_N\right).$$

- Note: 3 is the number of two-star isomorphisms (i.e., $\left|\text{iso}\left(\text{  }\right)\right|$) and $\rho_N^2 \left(1 - \rho_N\right)$ the probability that a triad is configured according to any one of them.

Transitivity (continued)

- In a random graph the transitivity index should therefore approximately equal

$$\begin{aligned} \text{TI} &\approx \frac{3\binom{N}{3}\rho_N^3}{3\binom{N}{3}\rho_N^2(1-\rho_N)+3\binom{N}{3}\rho_N^3} \\ &= \rho_N. \end{aligned}$$

- In practice TI often substantially exceeds network density (i.e., $\text{TI} \gg \rho_N$).

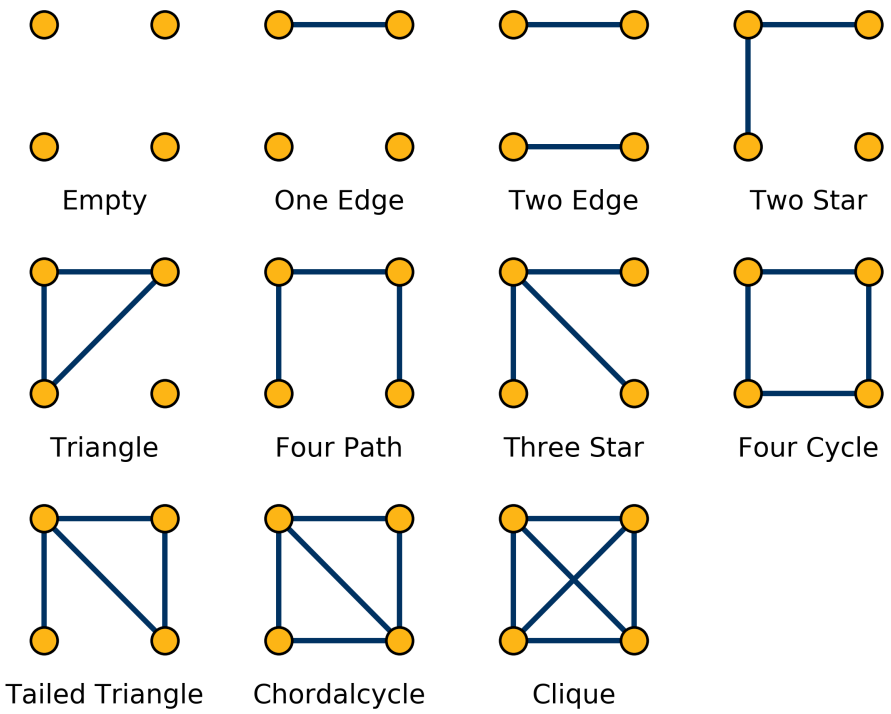
Transitivity (continued)

- Links are more clustered than would be expected under the homogenous random graph null.

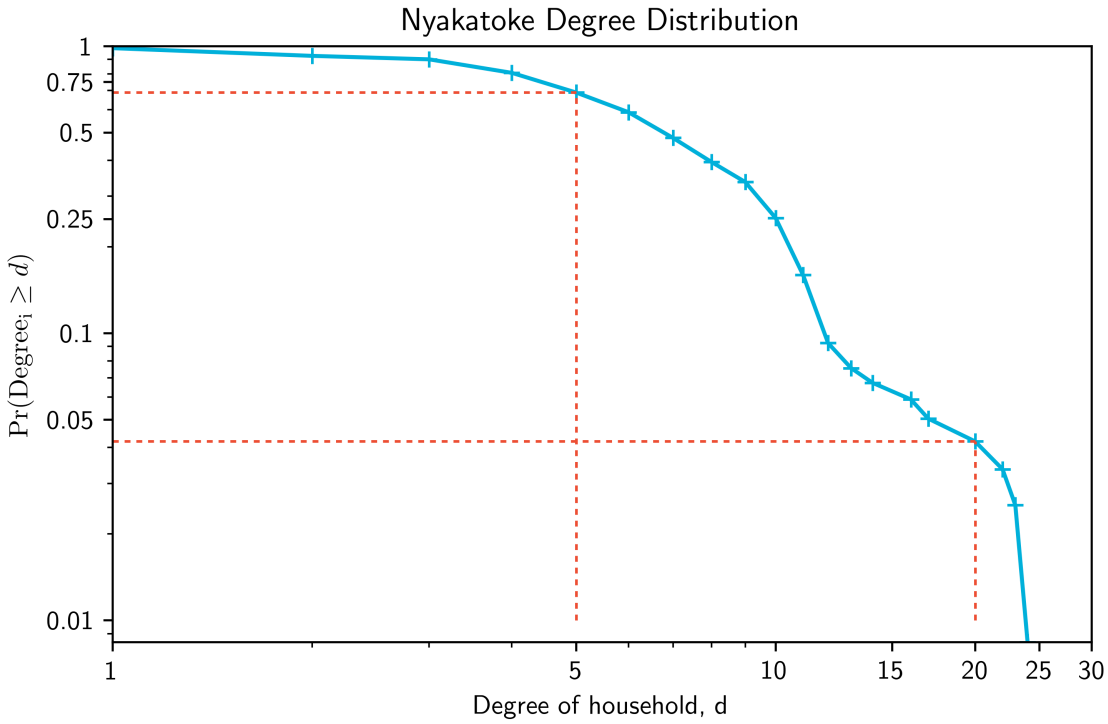
- i and j are more likely to be connected if they share a friend k in common (structural or homophily?).

- Granovetter (1973, AJS): two star is a "forbidden triad".

Tetrad Isomorphisms



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

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

- Knowledge of mean degree, degree variance and the number of triangles is equivalent to knowledge of the triad census.

Degree Distribution Redux (continued)

- The degree distribution constrains other (local) 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):

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

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

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

Power Laws (continued)

- Taking logs yields the linear relationship

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

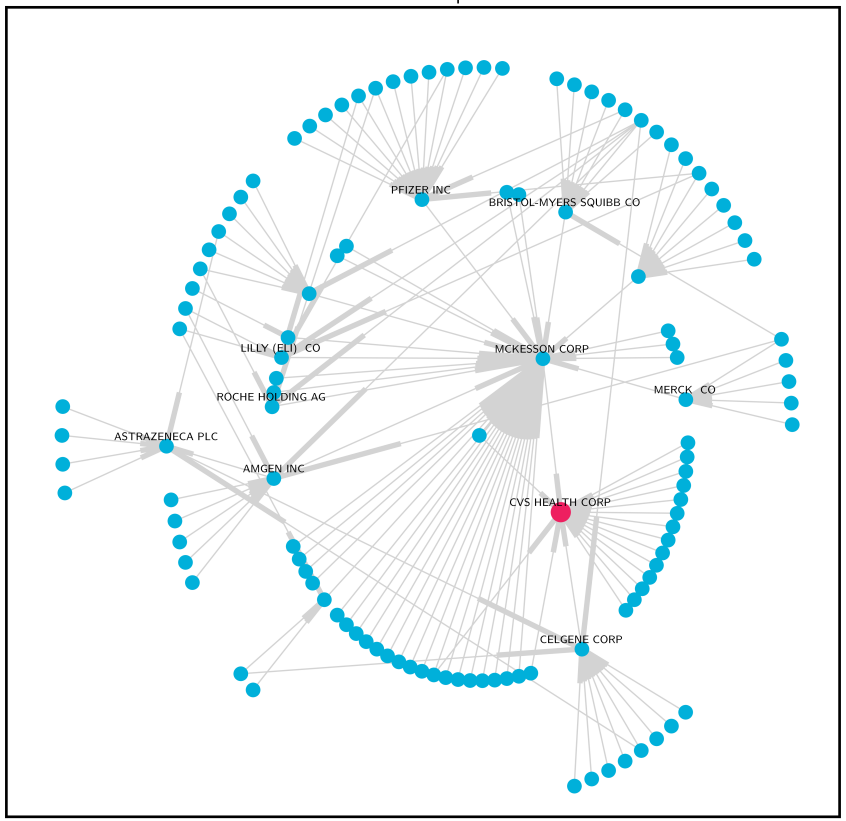
- The coefficient, α, may be estimated by OLS...

- ... but see Clauset, Shalizi and Newman (2009) and "power-law" Python package by Alstott, Bullmore and Plenz (2014) for better methods.

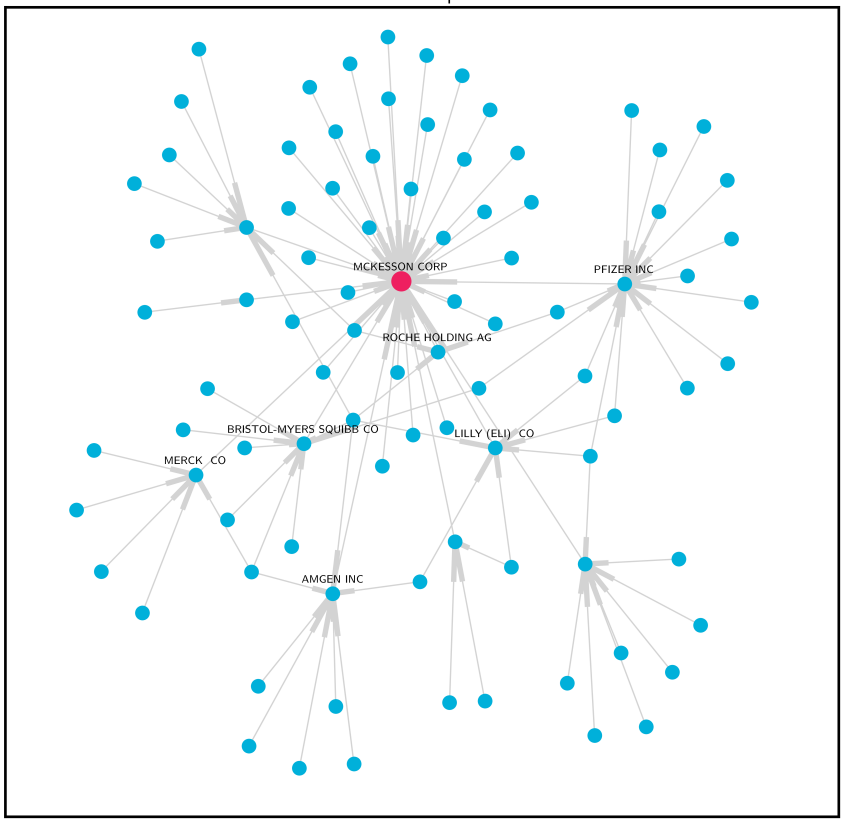
Directed Networks

- In some settings ties are naturally directed:
  - Buyer-Supplier networks
  - International trade flows
  - Financial networks

CVS Health Corporation



McKesson Corporation



Directed Networks (Buyer-Supplier Networks)

- If a firm supplies inputs to another firm, then there exists a *directed edge*  from the supplier to buyer.

- The supplying firm (left node) is called the *tail* of the edge, while the buying firm (right node) is its *head.*

- $G(\mathcal{V}, \mathcal{E})$, a directed network or *digraph* is defined on
  - $N = |\mathcal{V}|$ vertices or firms and
  - $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of all directed links (supplier-buyer relationships) among them.

Paths in Directed Networks

- Paths in directed networks have directionality (think of one way roads).

- It may be possible to travel from i to j via a series of directed paths, but not in the reverse direction.

- If a path runs from i to j, but not from j to i, we say i and j are *weakly connected*.

- If a path runs in both directions, the two agents are *strongly connected*.

Paths in Directed Networks (continued)

- In directed networks $D_{ij} = 1$ if i directs a link to j.
- If j also directs a tie to i, then $D_{ji} = 1$ and we say the link is *reciprocated* ().
- Adjacency matrix for a directed network need not be symmetric.
- The ij^{th} entry of $\mathbf{D}^K$ still gives the number of K length paths *from* i to j.

Reciprocity Index

- The frequency of the *asymmetric* dyad configuration in G equals

$$\widehat{P}\left(\text{  }\right) = \frac{2}{N\left(N-1\right)} \sum_{i=1}^{N} \sum_{j<i} \left[D_{ij} \left(1 - D_{ji}\right) + \left(1 - D_{ij}\right) D_{ji} \right].$$

- The frequency of the *mutual* configuration in G equals

$$\widehat{P}\left(\text{  }\right) = \frac{2}{N\left(N-1\right)} \sum_{i=1}^{N} \sum_{j<i} D_{ij} D_{ji}.$$

Reciprocity Index (continued)

- A standard measure of reciprocity (e.g., Newman, 2010) is given by

$$\hat{R}_N = \frac{2\hat{P}\left(\text{  }\right)}{2\hat{P}\left(\text{}\right) + \hat{P}\left(\text{  }\right)}. \tag{7}$$

Reciprocity Index (continued)

- If edges form *completely at random* with probability ρ_N, then

$$\hat{R}_N \approx \frac{2\rho_N^2}{2\rho_N^2 + (1-\rho_N)\,\rho_N} = \rho_N.$$

- In practice $\hat{R}_N$ is generally far from ρ_N.

- For example, reciprocity is
  - common in social networks (i.e., $\hat{R}_N \gg \hat{\rho}_N$),
  - rare in supply-chains ($i.e., \hat{R}_N \ll \hat{\rho}_N$).

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 (cf., Bickel, Chen and Levina, 2011).