INTRODUCTION TO SOCIAL NETWORK ANALYSIS AND NETWORK SCIENCE METHODS

Workshop, ICHPS 2023, Scottsdale, AZ

James O'Malley, Ph.D.

Department of Biomedical Data Science

The Dartmouth Institute for Health Policy and Clinical Practice

Geisel School of Medicine at Dartmouth

Email: James.OMalley@Dartmouth.edu

Pre-workshop and during-workshop: Supplemental Materials Used in Workshop

- Github site with supplemental material, R scripts and data used in the illustrative examples presented in workshop:

  https://github.com/kiwijomalley/ICHPS-Social-Network-Analysis-Workshop-2023

- GRANDPA algorithm
  - Use to generate random networks for analysis that approximates a base network
  - Very useful if data is confidential
  - Bobak CA, Zhao Y, Levy JJ, and O'Malley AJ. (2022). GRANDPA: GeneRAtive Network sampling using Degree and Property Augmentation applied to the analysis of partially confidential healthcare networks. doi.org/10.48550/arXiv.2211.15000
  - Presented in a conference poster by Carly Bobak (for Yifan Zhao) on January 9, 2023

Outline: The next 2 hours

I. Descriptive analyses involving networks (35min)
- Definition and examples of networks (5 min)
- Notation and descriptive features of networks (20 min)
- Statistical models involving comparative analysis of descriptive features of multiple networks (10 mins)

II. Statistical analysis of relational data ("sociology") (50 min)
- What factors affect the formation/dissolution of relationships or structure of society?
- Examples, similarity of characteristics (homophily), reinforcement of relationships, test for presence of transitivity?

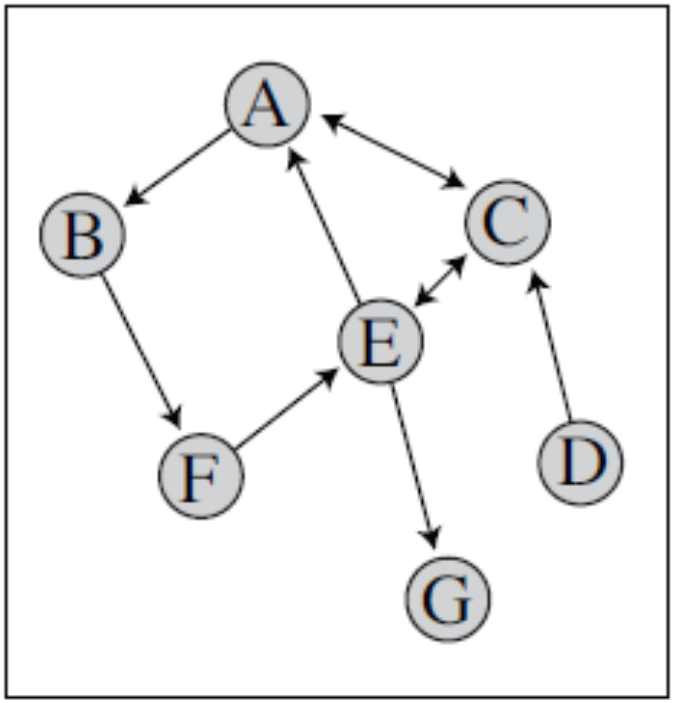
III. Statistical analyses of social influence or peer effects ("medicine") (20 min)
- Do individuals influence one another? Example, diffusion of treatment tendencies or innovations across physicians (Coleman, 1957, 1966)
- Longitudinal and cross-sectional cases

Questions interspersed (≈ 15 min)

I. Definitions, descriptive measures of networks, and models involving comparative analysis of multiple networks

Definition of a social network

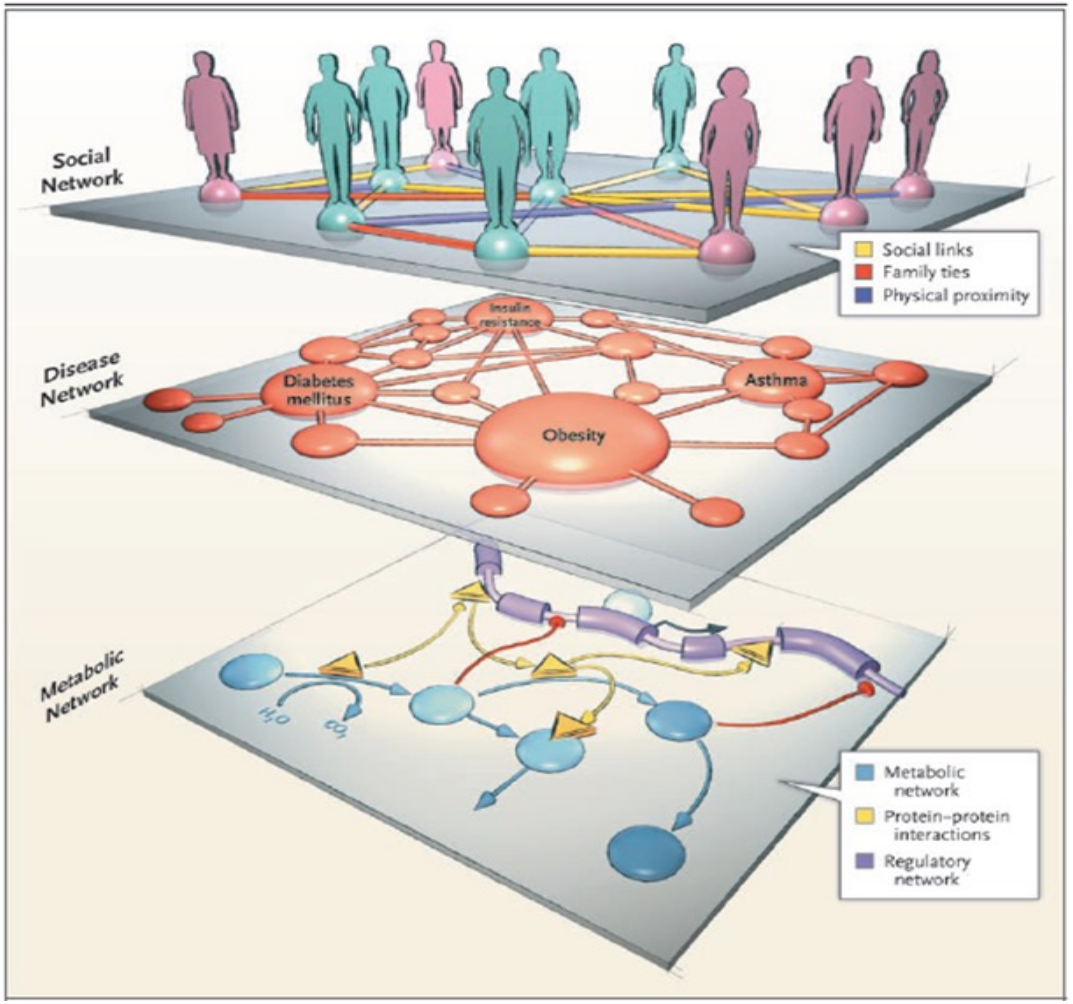
- A social network consists of one or more sets of actors—also known as "units," "nodes," or "vertices"—together with the possibly directed relationships or social ties among them



Components of a social network

- Actors:
  - Individual persons (e.g., patients or clinicians)
  - Organizations (e.g., hospitals)
  - Health and genetic states (e.g., phenotypes and genotypes)
  - Work products (e.g., academic papers)
- Social ties:
  - Communication
  - Influence
  - Trust or affect (e.g., friendship)
  - Affiliations (e.g., co-authors)
- Attributes:
  - Of actors (most typical), relationships, or both

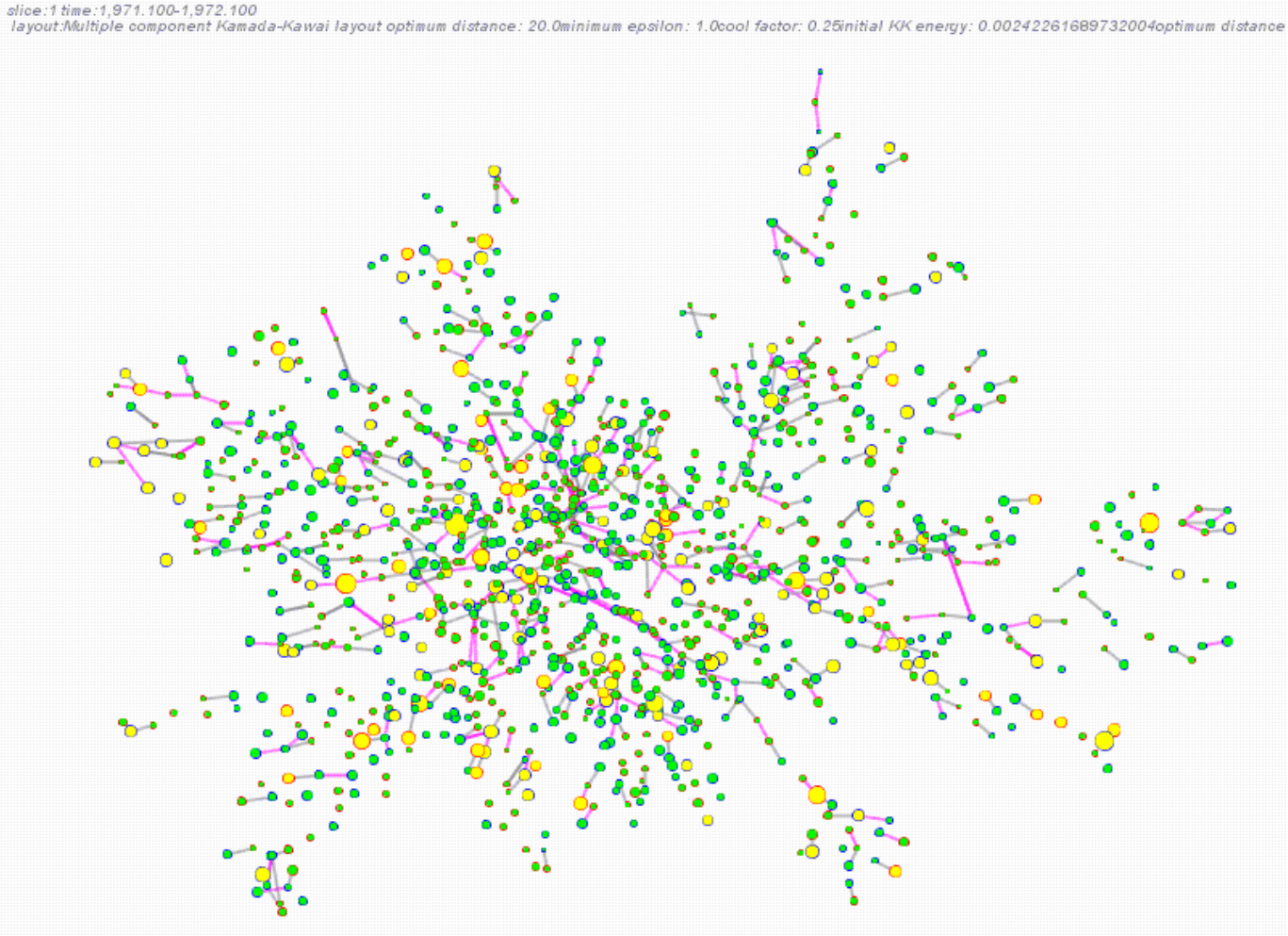
Layers of Networks in Medicine



Barabási, New England Journal of Medicine (2007)

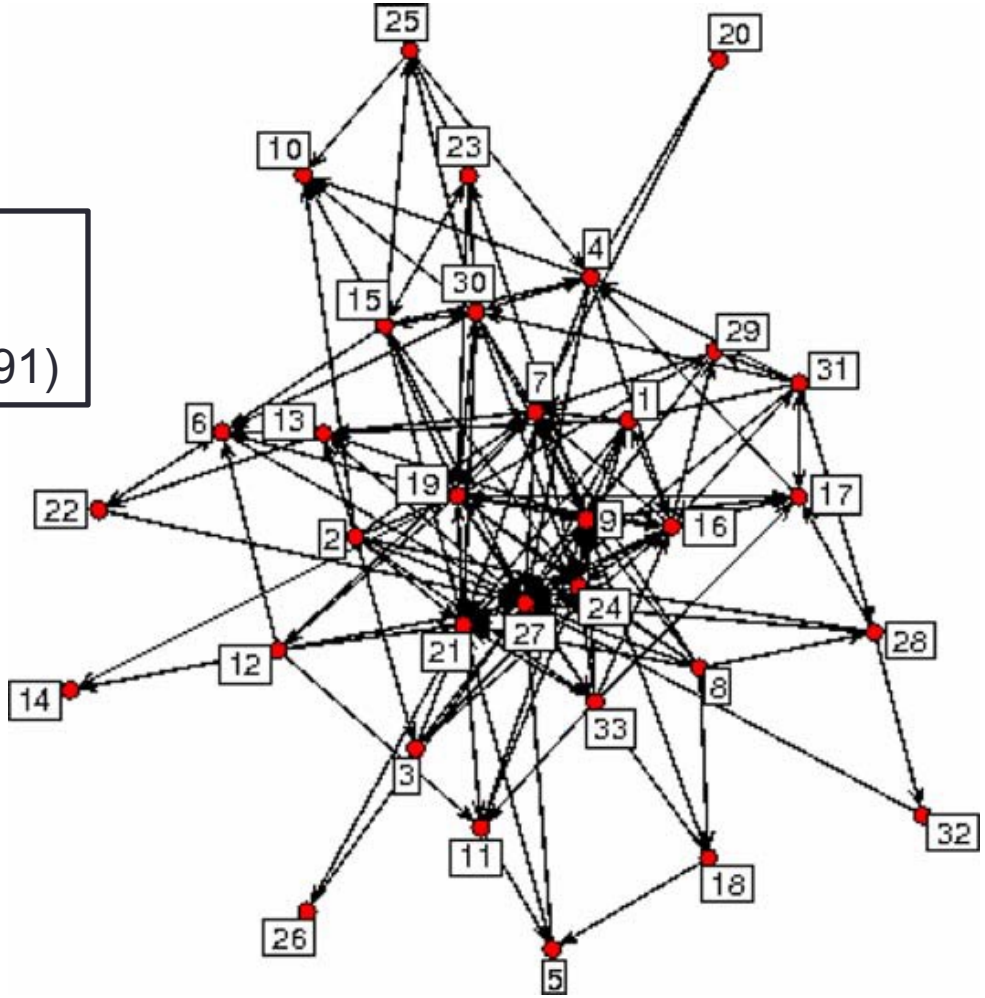
Social network: Spread of Obesity in Framingham Heart Study (Christakis and Fowler, 2007)

https://www.nejm.org/doi/full/10.1056/nejmsa066082

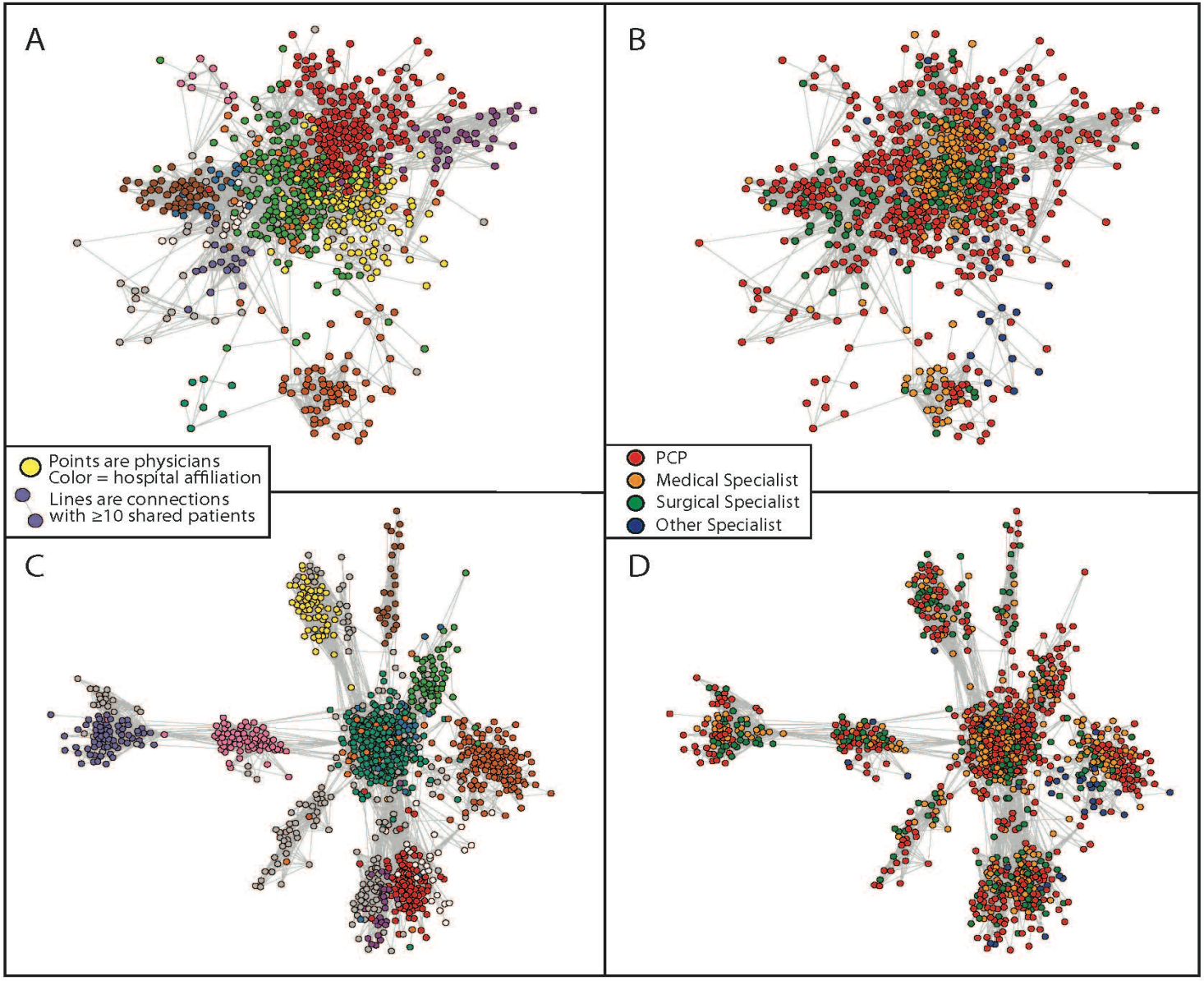


Physician social network within a clinic

Spring embedder algorithm
determines positions of actors
(Fruchterman and Reingold 1991)

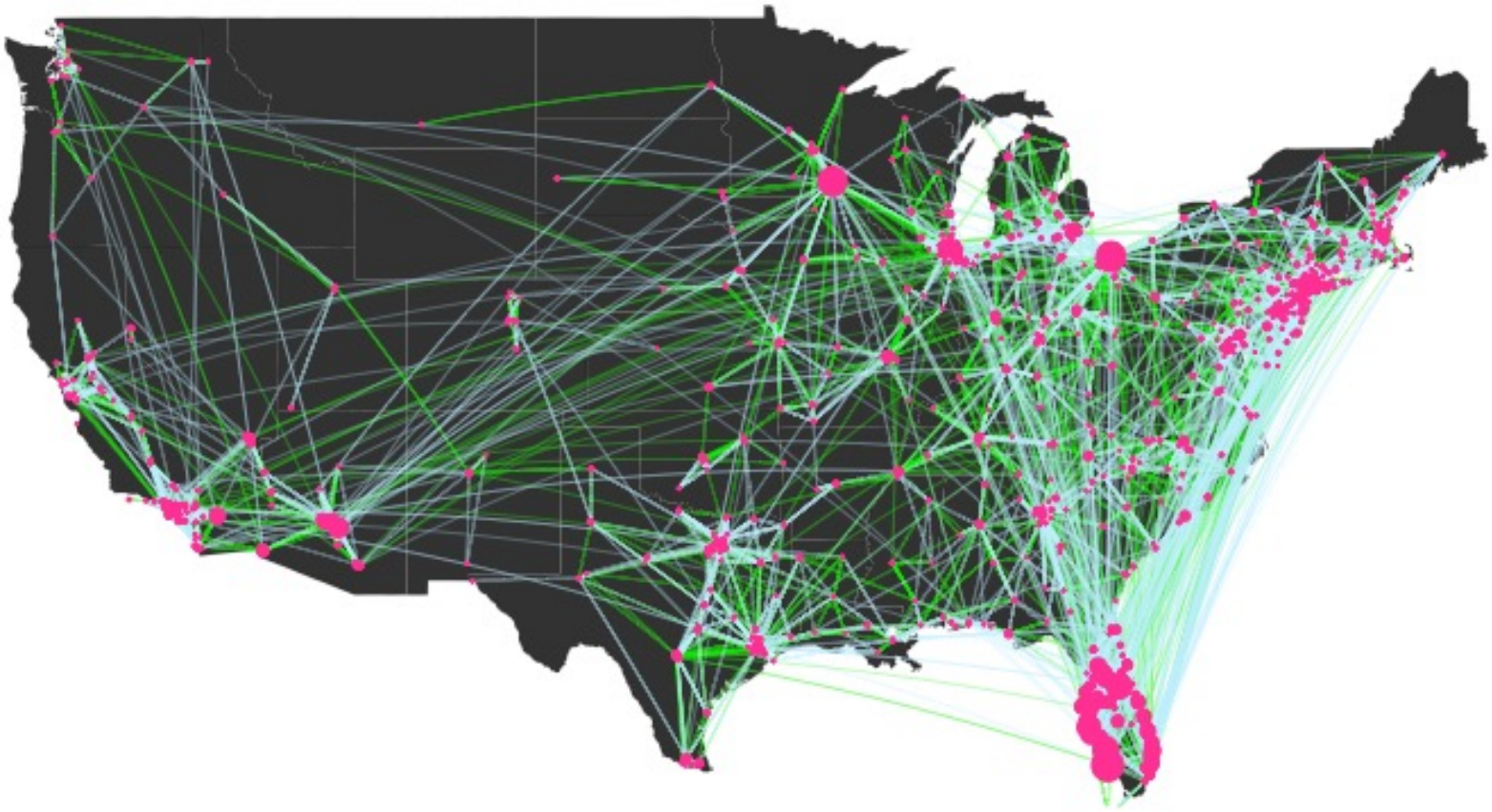


Network of physicians in a Boston hospital (Keating et al, 2007; O'Malley and Marsden, 2008)



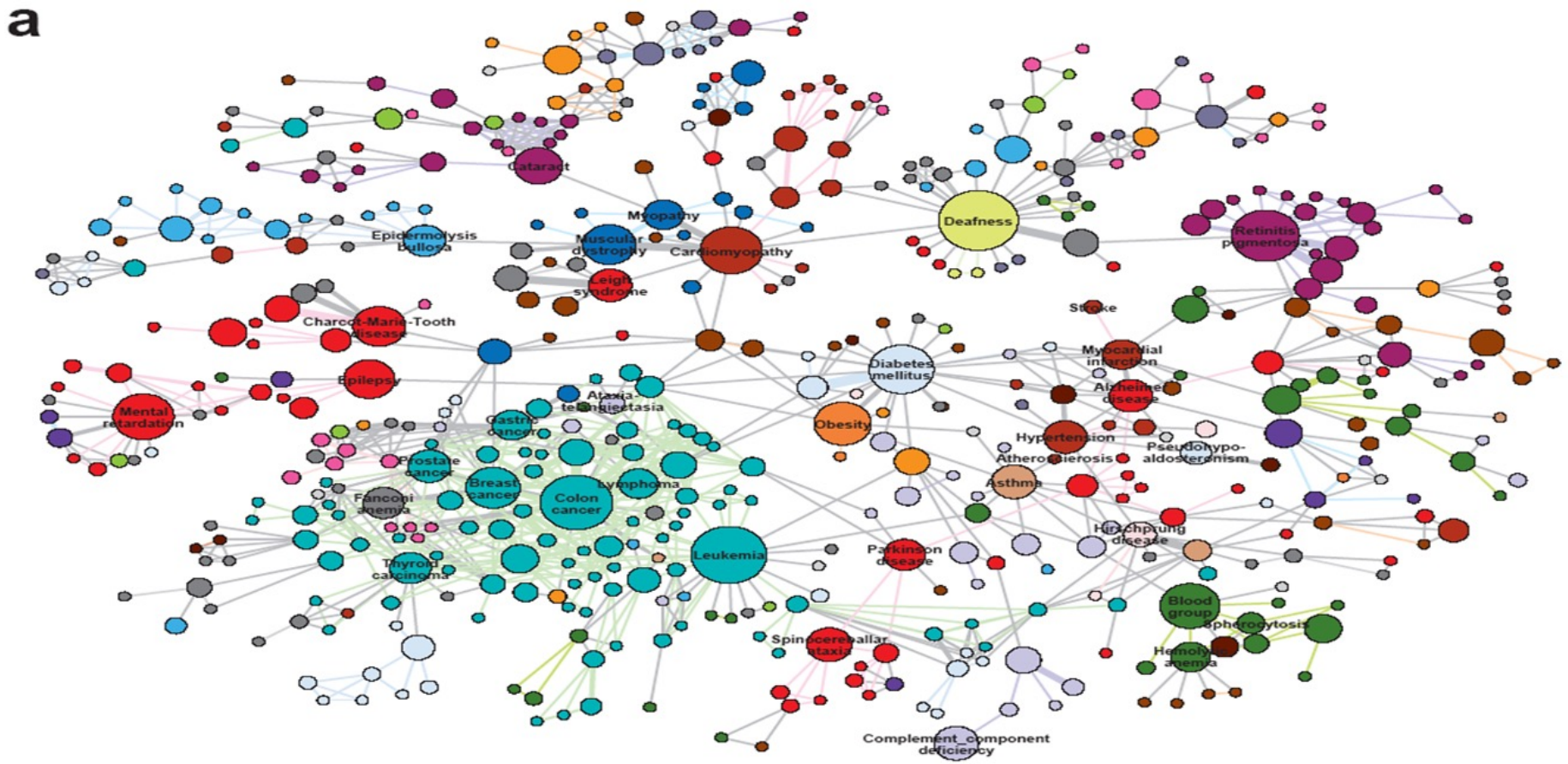
Networks of Physicians in 2 Health Referral Regions (Landon et al 2012, JAMA)

Aggregated US Physician Network (Moen et al, 2018)



Edges reflect number of shared patients between physicians at different hospitals (the 10% of hospitals with the most connections are presented)

Human Disease Network (edges reflect shared genes)



Adapted from: Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

Other networks

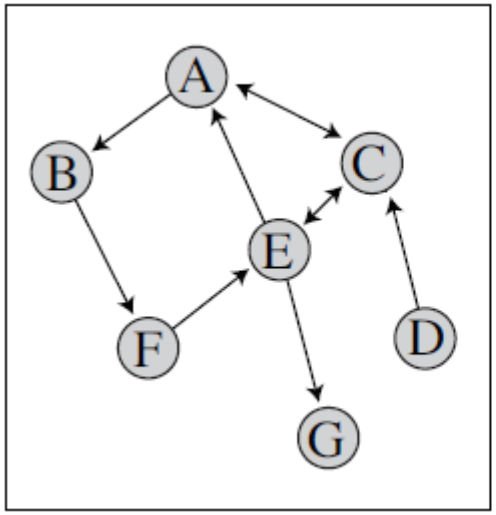
- Networks of congress based on sponsorship of bills
- Honduran villages
- World trade
- School classroom networks
- Company networks in Japan
- Other biological networks involving omics data
- ...

Types of Network Studies

- Sociocentric
  - Measure relationship between all pairs of individuals in the network
  - Richest form of network attainment
  - Often costly to obtain
- Bipartite
  - Nodes are of two distinct types
  - Measure relationships between nodes of different types
  - Can project to a one-mode network from a two-mode network
- Egocentric
  - Focal subjects' self-report their relationship to their peers and the relationships between their peers
  - Multiple networks; one for each study subject ("ego")
  - Networks are limited: often small, obtained using a survey, only a few things can be said about them, subject to miss-reporting by ego

Representing Relational Data

Digraph



Adjacency Matrix

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	0	0	0	0	1	0
C	1	0	0	0	1	0	0
D	0	0	1	0	0	0	0
E	1	0	1	0	0	0	1
F	0	0	0	0	1	0	0
G	0	0	0	0	0	0	0

- Letters = actors; numbers = relationship status
- In a binary (1, 0) network of *N* actors:
  - Dyads (pairs of actors) have $2^2 = 4$ possible states
  - $N(N-1)/2$ dyads
  - $4^{N(N-1)/2}$ possible networks
  - Some analyses of adjacency matrices related to models of lattices and the Ising (Pott's) model in statistical physics
- Can generalize to edges representing strength ("weighted network")

Entering Relational Data

1. Adjacency matrix
   - Rows and columns correspond to actors
2. Edge-list includes only edges with non-null values
   - Two-columns (2 ID fields) in a binary network and three-columns (2 ID fields, weight) in a weighted network
   - Excluded edges have 0 weight
   - Multiple edge-list files if have multiple measurements of relationship
   - Extensions to multiplex or multilayer networks
3. Incidence matrix
   - One row per actor and one column per edge

- Separate file for actor attribute information
  - Actor specific data forms a regular rectangular data set
  - Link to actor and edge data using actor IDs

Example Edgelists

Binary

npi1	npi2
1	7
1	12
1	24
1	25
1	36
1	38
1	39
1	42
1	49
1	51
…	
2	29
2	40
2	54
2	90
2	121
2	125
…	

Weighted

npi1	npi2	beneficiaries
1	7	2
1	12	2
1	24	4
1	25	2
1	36	8
1	38	2
1	39	1
1	42	2
1	49	1
1	51	8
…		
2	29	2
2	40	1
2	54	2
2	90	2
2	121	5
2	125	2
…		

Link to actor information

IDs for which relationship ("tie") exists; edge value = 1

Weight for value of tie

Network notation and properties

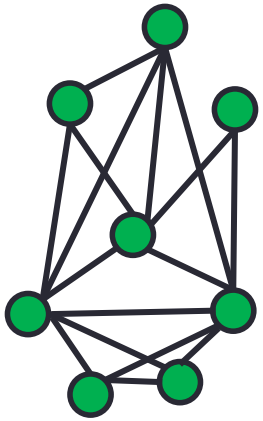
- A_{ij} denotes the relationship from actor *i* to actor *j*
- The array of relationships forms an adjacency matrix, denoted A
  - By convention $\mathrm{diag}(A) = 0$
  - In an undirected network, $A = A^T$
- In a weighted network, generalize edge from tie-existence (1 = yes, 0 = no) to allow different levels of strength

Network Features

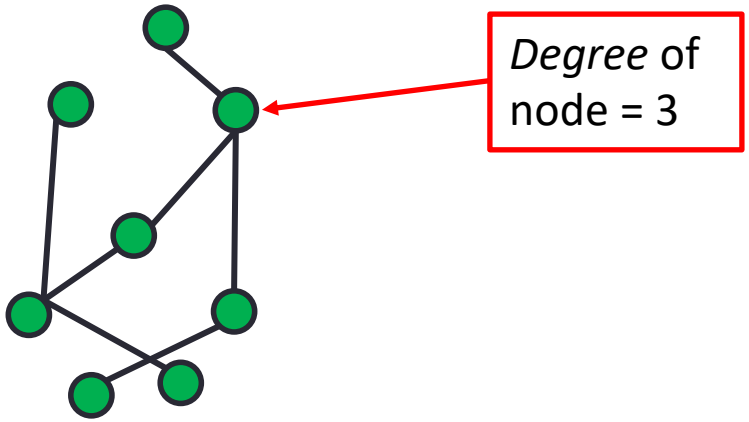
- Assume binary directed network (unless otherwise stated)
  - Many results simplify to undirected networks or generalize to weighted networks
- Size:
  - Number of nodes: *N*
  - Number of edges: $L = \sum_{i \neq j} A_{ij}$
- Density: Ratio of edges to possible edges
  - $\mu = L/((N(N-1))$
- Degree of actor *i*
  - Out-degree: $d_i^{out} = \sum_{j \neq i} A_{ij}$
  - In-degree: $d_i^{in} = \sum_{j \neq i} A_{ji}$

Density and degree

- Average-degree = $(N - 1) \times$ Density
- In a weighted network, the sum of an actor's edge-weights is the actor's Strength ("weighted degree")



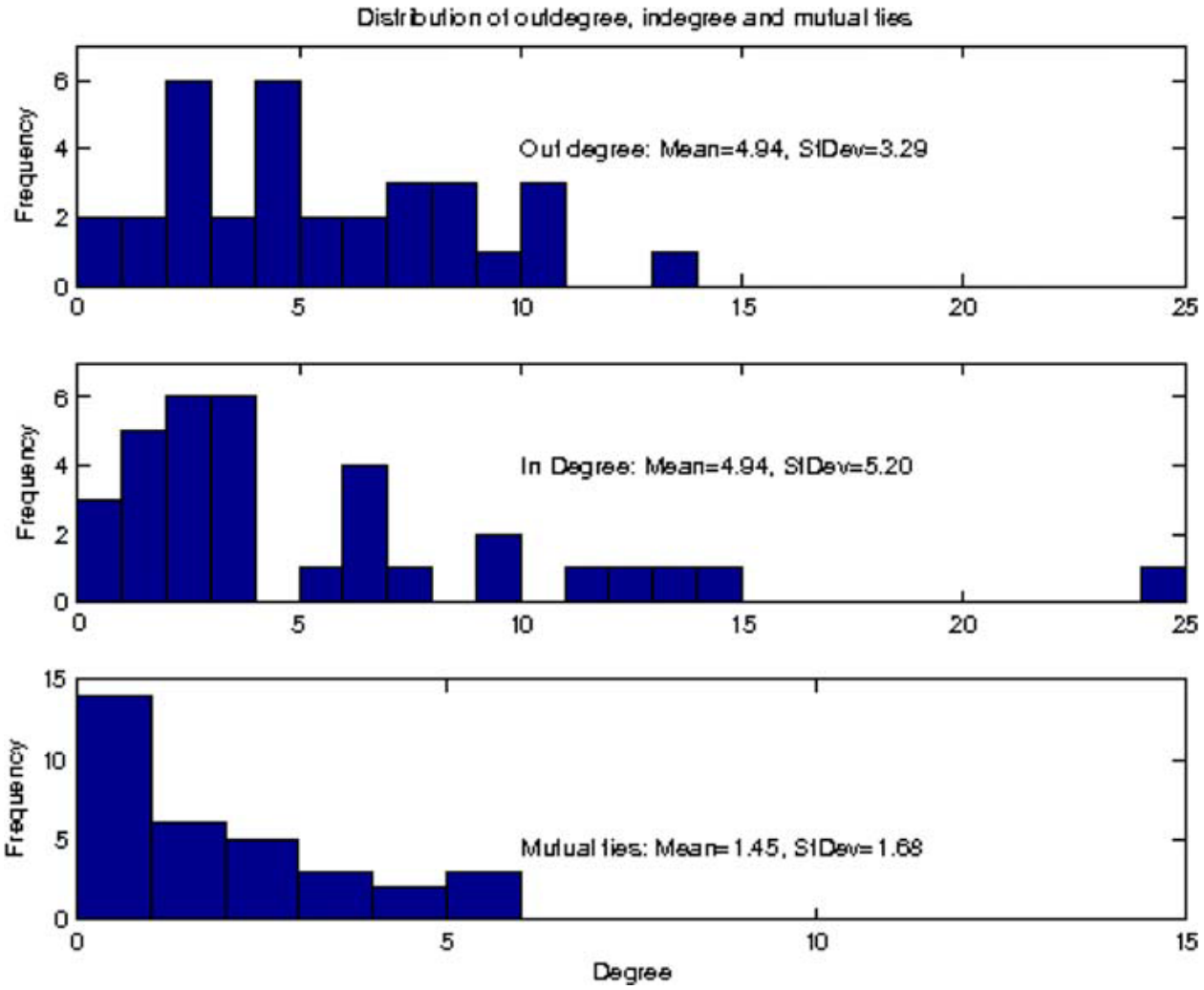
Density = 16/28 = 0.57



Density = 7/28 = 0.25

Degree distributions for N = 33 physicians in a medical practice

Edge denotes occurrence of important discussions about women's health issues (Keating et al, 2007; O'Malley and Marsden, 2008)



General properties of density and degree

- Out-degree measures expansiveness
- In-degree measures popularity
- Average out-degree = average in-degree
- Corr(out-degree, in-degree)
  - Within-actor degree correlation
  - Are expansive actors popular?
- Degree is a measure of an actor's importance in the network
  - Concept generalized later as centrality

Power law distribution

- In network science there is extensive interest in whether the degree distribution of a network follows a power law
- Under a power law, the probability actor *i* has degree *k*:

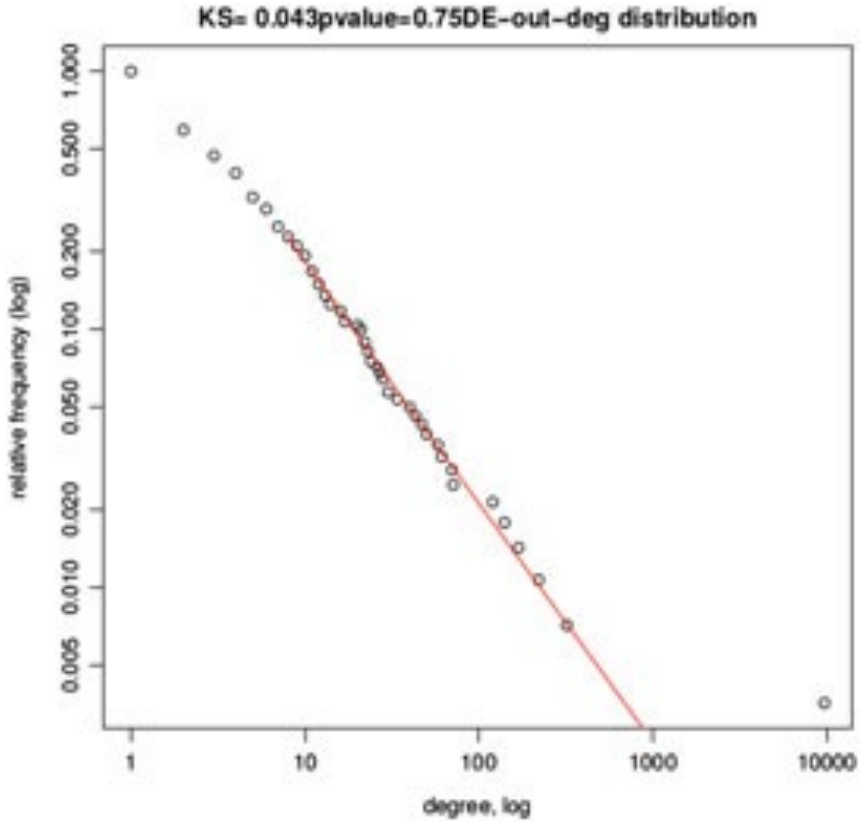
$$\Pr(\text{Degree}_i = k) \propto k^{-\lambda}$$

for a parameter λ, typically $\lambda > 2$

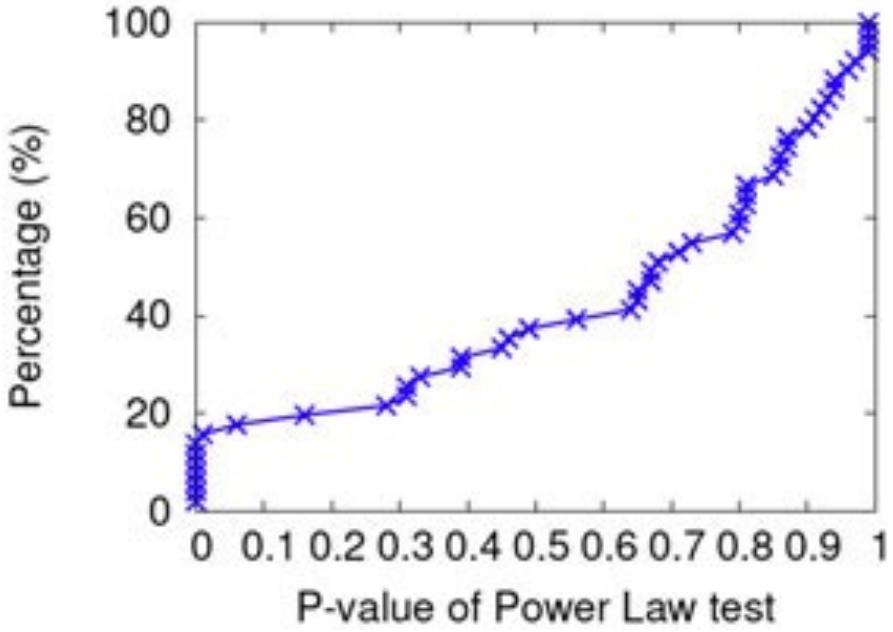
- Power-law distributions/networks are also known as "scale-free distributions/networks" (Barabasi and Albert 1999, Newman 2010)
  - Probability an actor forms a new tie is proportional to their degree

Power law distribution tests

1. Regress log-frequency of actor degree in network on log(degree) and test for lack-of-fit
2. Compare distribution of p-values for many networks to uniform distribution using Kolmogorov-Smirnov test

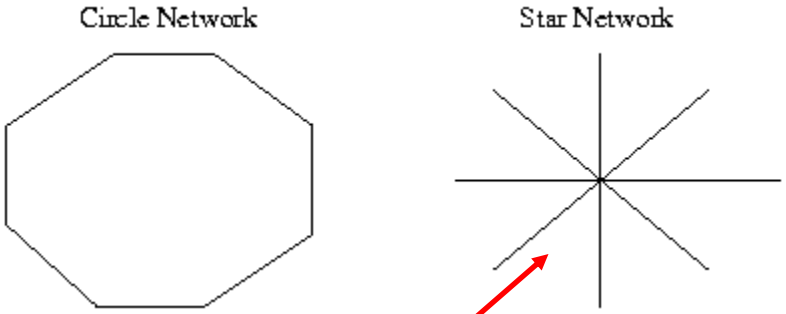


Figures from An et al (2018)



Centralization

- Heterogeneity of actors' network prominence:



- var(degree) = 0; $= \frac{\left(N-3+2N^{-1}\right)^2}{N-1} + \left(2N^{-1}-1\right)^2$
- Standardizing by maximum possible degree, d_{max} $(= N - 1$ in binary networks), might make comparisons between networks more meaningful
- Centralization index:
  - $C_D = \frac{\sum_i \left(\max_i\{d_i\} - d_{\{i\}}\right)}{(N-1)(N-2)}$, where $\max_i\{d_i\}$ is the maximum observed degree

Paths and Components

- Two actors are connected if there is a path linking them
  - Relevant to network meta-analysis
- Two actors are in separate components if there is no path linking them
- Number of components = number of groups of actors with no path between them
- A network is *connected* if it only contains a single component (i.e., a path between all pairs of actors exists)

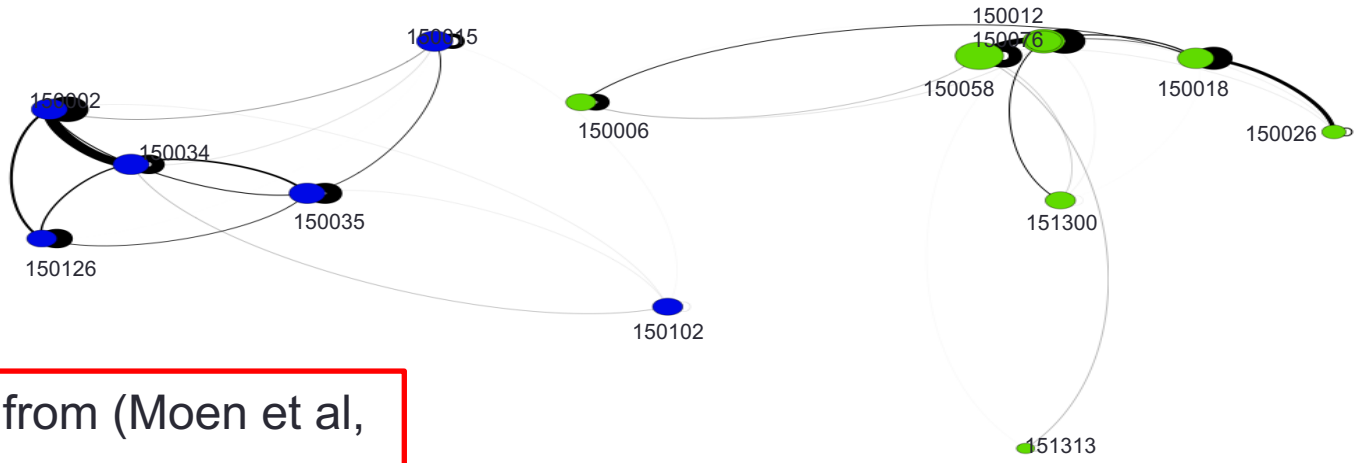


Figure from (Moen et al, 2016)

Distance

- The length of a path is the number of edges on it
- Shortest path between two actors = geodesic distance
  - Infinite if no path exists between actors
  - Equals 1 in the complete graph
- Size, degree, and density measures are well-defined irrespective of whether the network is connected but many other network measures are ill-defined if the network has multiple components
  - Distance-based measures ill-defined if multiple components
  - Motivates analysis of largest connected component

Path and distance calculations can utilize matrix algebra

In a binary network, A^k tells you about paths of length *k*

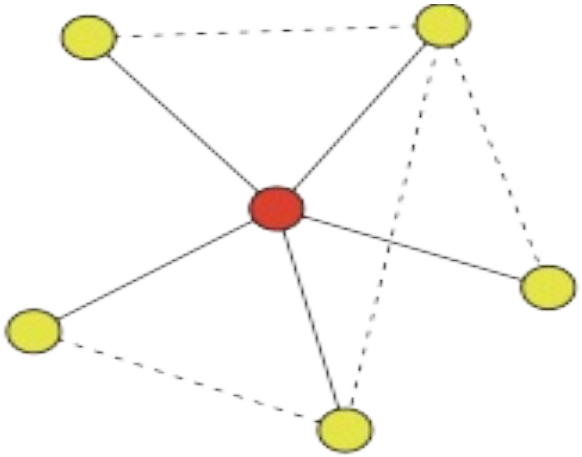
- $[A^2]_{ii}$ = number of cycles of length 2 = number of reciprocated ties (or degree in an undirected network)
- $[A^2]_{ij}$ = number of two-paths from actor *i* to actor *j*
- $\sum_{j=1}^{N} I([A^2]_{ij} > 0)$ = number of second-degree actors to actor *i* (does not count first-degree actors)
- $[A^k]_{ii}$ = number of k-cycles back to actor *i*
- $[A^k]_{ij}$ = number of k-paths from *i to j*

Geodesic distance is the shortest distance (path length)

- Geodesic distance from *i* to *j* $= \min\{k: [A^k]_{ij} > 0\}$

Clustering coefficient

- Extent to which the network neighbors of a given node are directly connected to one another



- Clustering coefficient for a given node = probability that any two randomly chosen network neighbors of the focal individual (e.g., in red) are directly connected (e.g., = 0.4)

Small world property

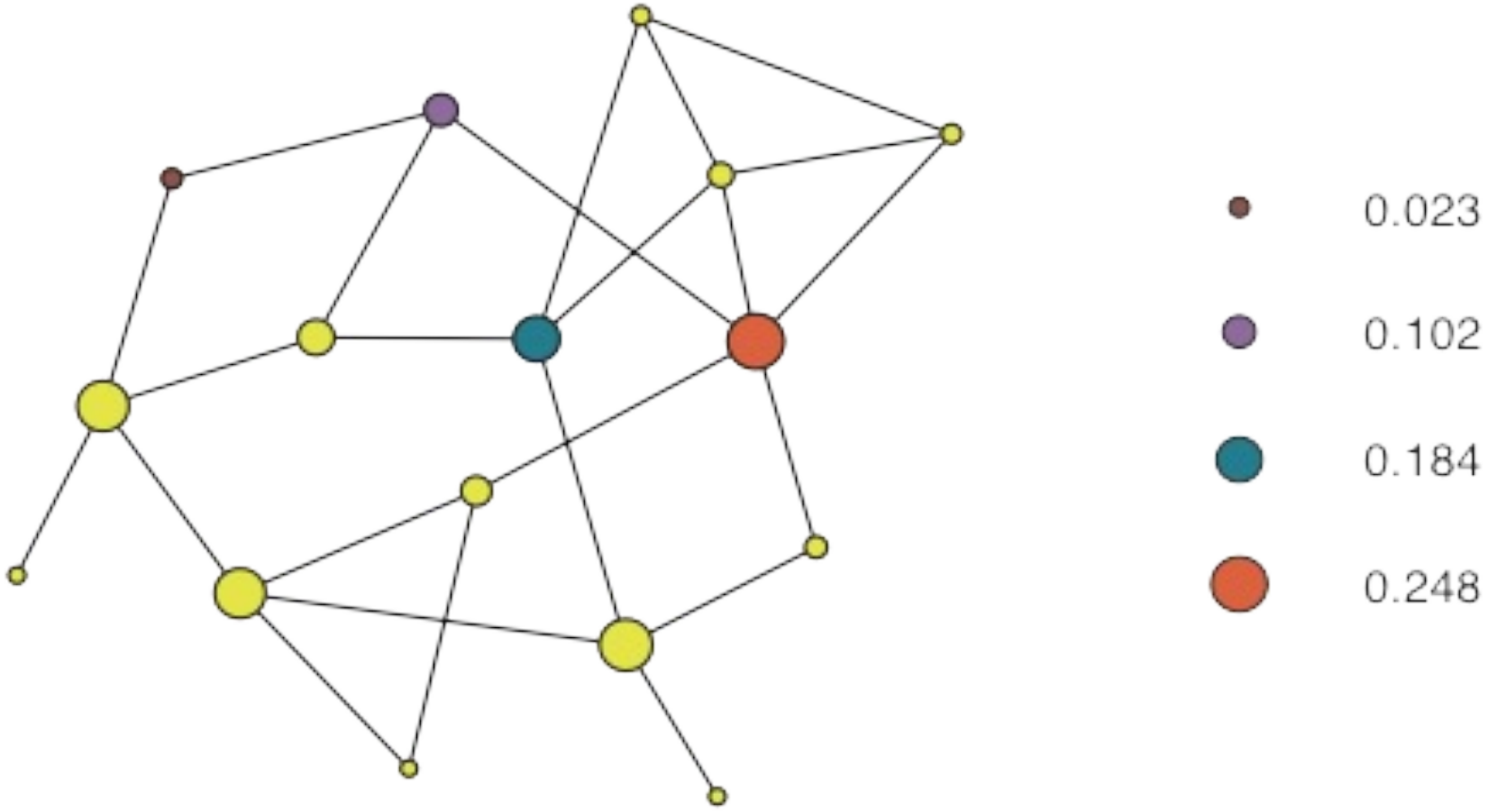
- Another concept that features in network science
- Six-degrees of separation (Watts and Strogatz 1998)
- Holds when there is greater than expected local connectivity and average path length is smaller than expected compared to the completely random (Erdos-Renyi) network
  - Erdos-Renyi (ER) network: State of each edge is an independent and identically distributed Bernoulli random variable
- Small world property → network has a higher (average) clustering coefficient than would be expected by chance (i.e., under the ER network)

Four common measures of centrality

1. Degree (defined earlier)
2. Closeness centrality
   - The inverse of the mean of the geodesic distances from a given actor to all others
3. Betweenness centrality
   - The proportion of times an actor is in an intermediary position on the geodesic paths linking pairs of other actors
4. Eigenvector centrality
   - Construct: Connections to central actors indicate greater prominence than similar-strength connections to peripheral actors
   - Google's Page Rank in same family of measures

Average over actors → network-level centrality
Variance over actors → network heterogeneity or centralization

Illustration of betweenness centrality



Bigger node has greater betweenness centrality

Insight into Eigenvector Centrality

- Recall A denotes the adjacency matrix of network or weighted variant
  - A_{ji} = relationship from *j* to *i*
- Inbound eigenvector centrality: Actor *i* is central if he/she is named by central actors
- Because actor *i*'s centrality simultaneously defines the centrality of actors' that name her/him, this suggests that we should solve the system of equations:

$$c_i = A_{1i}c_1 + A_{2i}c_2 + \cdots + A_{Ni}c_N$$

for the centrality c_i of each actor $i = 1, \ldots, N$

- Generalizes to a family of measures: Eigenvector, Prestige, Katz, Bonachich, and PageRank centrality

Reciprocity ("Mutuality")

- The Dyad census is a count of the number of (Mutual, Asymmetric, Unconnected/Null) dyads
- Mutual or reciprocated dyad indicator: $M_{ij} = A_{ij}A_{ji} = M_{ji}$
  - Number of mutual dyads: $M = \sum_{i<j} M_{ij}$
- One index of reciprocity expresses ρ as the deviation in the conditional probability of a tie from that expected under dyadic independence:

$$\Pr\left(A_{ij} = 1 \middle| A_{ji} = 1\right) = \Pr\left(A_{ij} = 1\right) + \rho \Pr\left(A_{ij} = 0\right)$$

- If $\rho = 0$ then independence holds
- If $\rho = 1$ then $\Pr\left(A_{ij} = 1 \middle| A_{ji} = 1\right) = 1$ and the presence of $j \to i$ guarantees the presence of $i \to j$
- Anti-reciprocity, $\rho < 0$, may occur

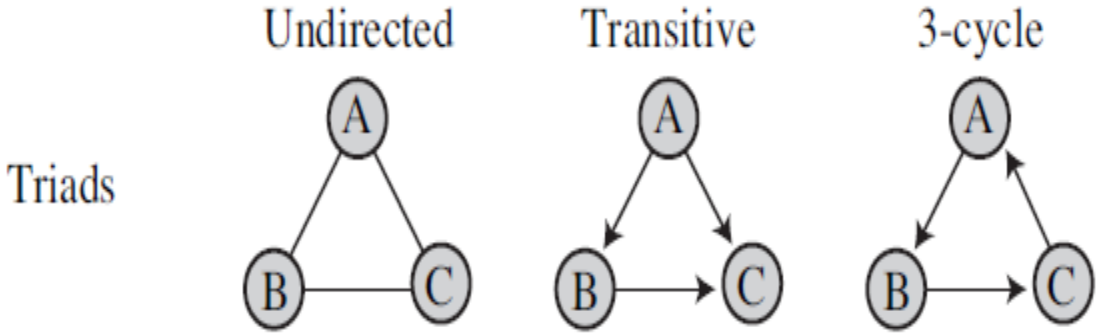
Example mutuality measure

- Quantifies extent that M exceeds the number expected by chance in the absence of reciprocity
  - Depends on degree distribution of the network (and thus density)
- If all actors have degree *d* in a network of *N* actors $\Pr(A_{ij} = 1) = d/(N-1)$
- Then solve for ρ:

$$\hat{\rho} = \frac{2(N-1)M - Nd^2}{Nd(N-1-d)}$$

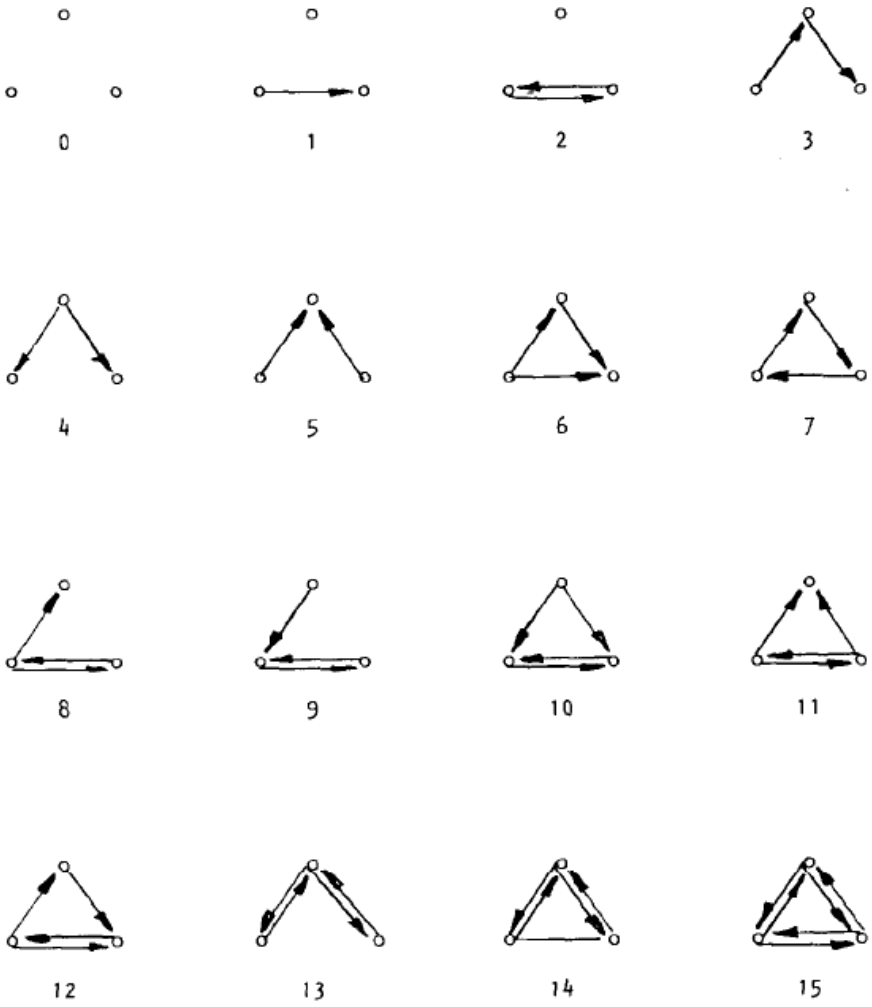
- See Wasserman and Faust (1994) for non-model-based statistical tests of reciprocity

Transitivity ("A friend of a friend is a friend")



- Sociologists → Triads are an important building block of society
- Triadic clustering is a special form of clustering
- Undirected network: the count of triangles is the basis of measures of transitivity
- Directed network:
  - $4^3 = 64$ states of a triad
  - 16 triad groups that are non-isomorphic, embodying several sociological constructs (including transitivity)
  - The transitive triad (shown above for actor A) is often of primary interest

The 16 non-isomorphic triads (Frank, 1986)



Classical (method-of-moments) tests for transitivity (Wasserman and Faust 1994) in undirected network:

- Condition on the dyad census (M = mutual, A = asymmetric, U = unconnected) and degree-distribution.
- Using randomly drawn networks with no triadic dependence, generate null distribution of triad counts
- Subtract expected count from observed and divide by standard error

Assortativity or Preferential Attachment

- The degree to which "likes link to likes"
  - Also known as homophily
- Degree assortativity:
  - Predilection of high degree nodes to link to other high degree nodes
  - Popular actors forming ties with each other
  - Expansive actors forming ties with each other
  - Across-degree (In-out, Out-in) degree assortativity
- Distinct from reciprocity or within-actor degree (popularity-expansiveness) correlation
- Attribute assortativity:
  - E.g., Do high spending physicians link to each other more often than expected by chance?

Properties of Assortativity

- A_{ij} denotes the status of the edge from physician *i* to *j*
- x_i denotes an attribute and d_i the degree of actor *i*
  - d_i could be in-degree or out-degree (or strength)
- $2m = \sum_{i=1}^{N} d_i = \sum_{i,j} A_{ij}$ is the total number of ties (or strength) in the network
- Assortativity with respect to x:

$$r = \frac{\sum_{ij}\left(A_{ij} - d_i d_j / 2m\right)x_i x_j}{\sum_{ij}\left(d_i I(i=j) - d_i d_j / 2m\right)x_i x_j}$$

- Denominator proportional to the maximum covariance of $\boldsymbol{x}$ conditional on A
  - Realized if $x_i = x_j$ whenever $A_{ij} = 1$
- Modularity optimization (Newman, 2006)
  - Finds x_i (categorical) for each actor by seeking to maximize modularity equation; resulting labels referred to as communities

Other Descriptive Network Measures

- Core-periphery
  - Core-ness measure for each node evaluates extent to which node is in the core
  - Gini coefficient of core-ness measures for the network
- K-coreness
  - Extent to which there are groups of nodes each connected to k other nodes within the group
- Gravity
  - Strength of tie inversely related to distance (macro-economic trade)
- Structural balance
  - Relevant to signed networks (e.g., edges valued as 1, 0, and -1)
- …

Many more measures developed in Social network or Network Science paradigms!

Computing Materials Used in Workshop

- Github site with supplemental material, R scripts and data used in the illustrative examples presented in workshop:

https://github.com/kiwijomalley/ICHPS-Social-Network-Analysis-Workshop-2023

- Supplemental slides: ICHPS2023_ShortCourseCode.pdf
- R scripts (include data sets): WorkshopICHPS2023.R, WorkshopICHPS2023grandpa.R, WorkshopICHPS2023models.R

- GRANDPA algorithm
  - Use to generate random networks for analysis that emulate a base network
  - Very useful if network data is confidential
  - Bobak CA, Zhao Y, Levy JJ, and O'Malley AJ. (2022). GRANDPA: GeneRAtive Network sampling using Degree and Property Augmentation applied to the analysis of partially confidential healthcare networks. doi.org/10.48550/arXiv.2211.15000
  - Presented in a conference poster by Carly Bobak (for Yifan Zhao) on January 9, 2023

Using igraph to graph a network in R

- Let edgelist be a N by 2 matrix in R with the network represented as an edgelist
- nodes <- unique(c(edgelist[,1],edgelist[,2]))

Make graph object from edgelist:

- gnet <- graph_from_data_frame(d=edgelist, vertices=nodes, directed=TRUE)

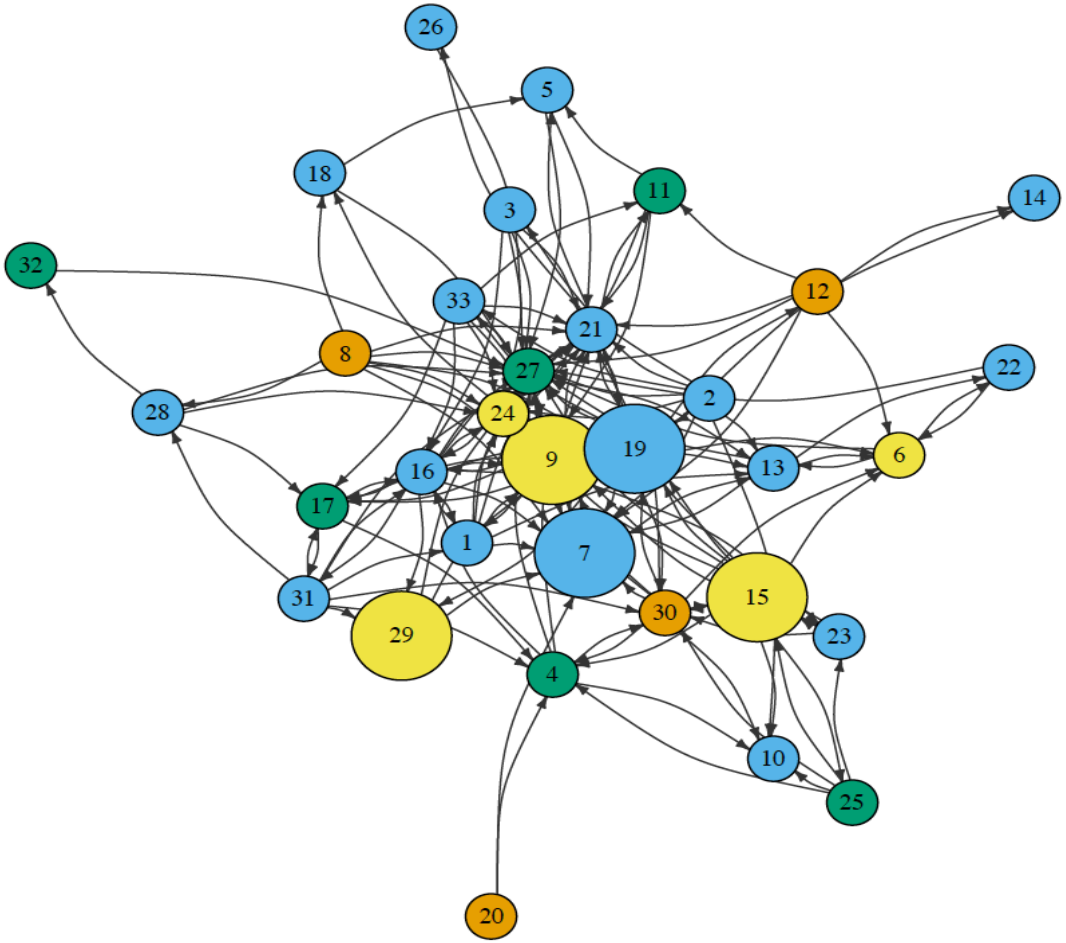
Plot graph

- print(gnet, e=TRUE, v=TRUE)

Plotting network in igraph

- V(gnet)$color <- covdata$practice
- V(gnet)$size <- 10*(covdata$whexpert+1)
- par(mar=c(0,0,0,0))
- plot(gnet,
- vertex.color = V(gnet)$color, # Color of nodes
- vertex.size = V(gnet)$size, # Size of nodes
- vertex.label.color = "black", # change color of labels
- vertex.label.cex = .75, # change size of labels to 75% of original size
- edge.curved=.25, # add a 25% curve to the edges
- edge.color="grey20", # change edge color to grey
- edge.arrow.size=0.3)

Visualization of network of physicians' professional relationships within a medical practice



Computing summary measures of networks in R

- reldir <- scan("ICHPS_PhysPractBin.txt")
- nr <- sqrt(length(reldir)) #Number of physicians
- reldir <- matrix(reldir,ncol=nr,nrow=nr,byrow=T)

- Degree distributions
  - idegree=degree(reldir,cmode="indegree")
  - odegree=degree(reldir,cmode="outdegree")
  - central=centralization(reldir,degree)

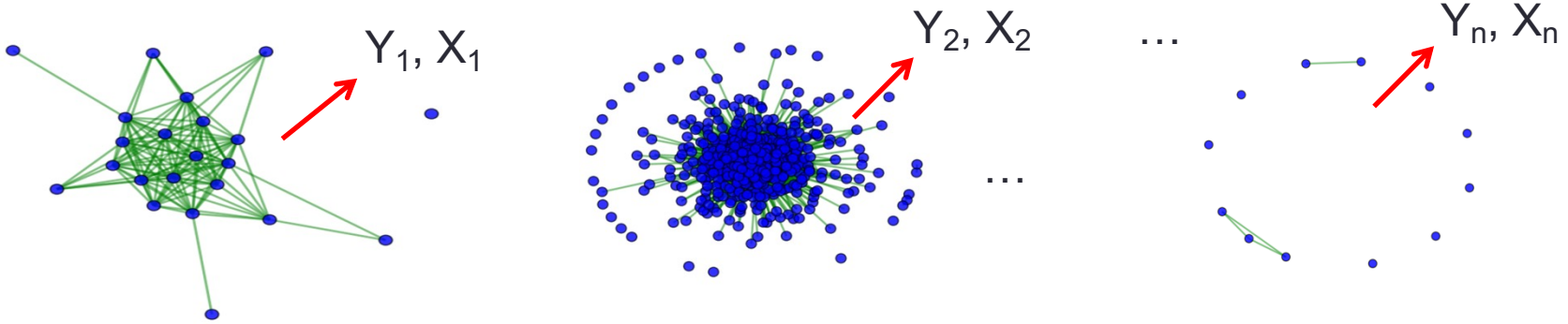
- Centrality measures
  - closecent=closeness(reldir,gmode="digraph")
  - bcent=betweenness(reldir,gmode="digraph")
  - eigcent=evcent(reldir,gmode="digraph",use.eigen=FALSE)
  - powcent=bonpow(reldir,gmode="digraph")

Statistical models involving comparative analysis of multiple networks

- Do social network characteristics correlate with other variables of interest?
- Often involves use of regression and hierarchical regression models

Networks are observational units

- For each network, generate summary measures that are used as either the outcome(s) or predictor variable(s) in a subsequent analysis involving other variables
- Applies when it is reasonable to think of the units on which networks are evaluated as distinct
- “Standard” statistical analysis might have 1 obs per network; need to account for clustering if units observed multiple times



Application: US physician network based on Medicare claims

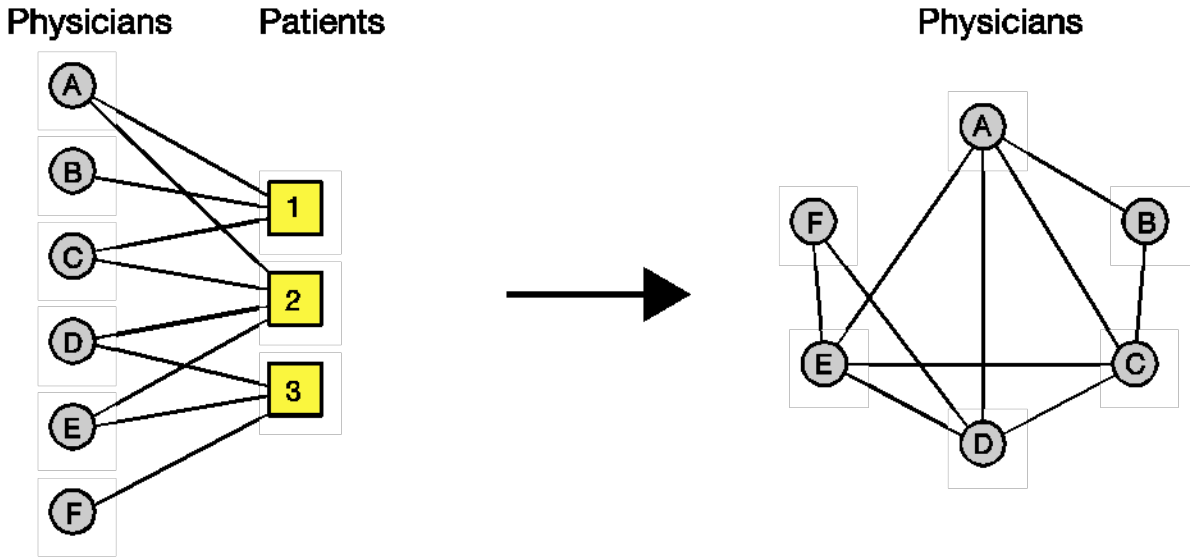
- Question: Does variation in network features across health organizations or regions correlate with outcomes of interest:
  - Health care utilization, intensity, and cost of care?

- Example: The network of physicians in hospitals or health referral regions across the US
  - Conjecture: physician influence and diffusion attitudes/practice-patterns operate through professional relationships

Bipartite Physician Hospital Networks

- Ideal: Directly observe ties between physicians
- In lieu of direct measurement use patients to infer professional relationships between providers
  - "Surrogate" relationship: Overlap of patients ("shared-patients")
- Assign physicians a "hospital" according to where the greatest proportion of patients they care for are hospitalized
- Then form a network of physicians for each hospital

Bipartite projection



$$\mathbf{B} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline A & 1 & 1 & 0 \\ B & 1 & 0 & 0 \\ C & 1 & 1 & 0 \\ D & 0 & 1 & 1 \\ E & 0 & 1 & 1 \\ F & 0 & 0 & 1 \end{array}$$

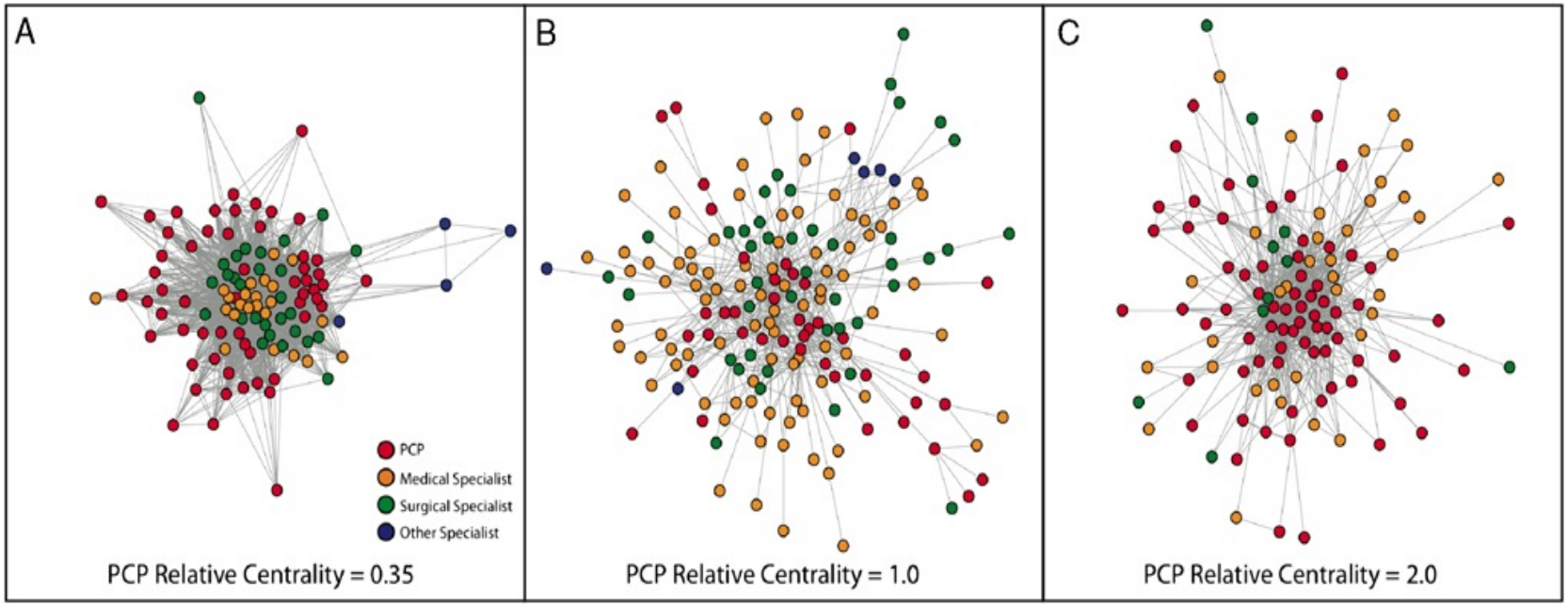
$$\mathbf{A} = \mathbf{B}\mathbf{B}^T = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 2 & 1 & 2 & 1 & 1 & 0 \\ B & 1 & 1 & 1 & 0 & 0 & 0 \\ C & 2 & 1 & 2 & 1 & 1 & 0 \\ D & 1 & 0 & 1 & 2 & 2 & 1 \\ E & 1 & 0 & 1 & 2 & 2 & 1 \\ F & 0 & 0 & 0 & 1 & 1 & 1 \end{array}$$

Comparative outcome analysis with respect to network features

- Compute network statistics for each network and the actors in them
  - Network-level summary measures
  - Actor within-network summary measures
- Model the outcome of interest (cost, intensity of care) in terms of network statistics and other predictors
  - Barnett et al (2012) analyzed 526 hospitals regressing case-mix adjusted cost of care and other outcomes on 3 summary features of the physician network within each hospital
  - An et al (2017) regressed state health litigation and health outcome variables for multiple years on an extensive range of network summary measures for each US state
  - Moen et al (2018) analyzed 4,000 hospitals relating the positions of physicians in hospital networks, hospital network features, and of hospital position in the US national network to appropriateness (guideline consistency) of implantable cardiac defibrillator outcomes

Example: Hospital Network Analysis (Barnett et al 2012)

- Predictor of interest: relative centrality of PCPs



- In panel A, PCP relative (betweenness) centrality to that of specialists is 0.35, so PCPs are about a third as central as other physicians in this network
  - Reflects the tight group of medical and surgical specialists at center of network
- In panels B and C, PCPs move more toward the center; PCP relative centrality increases

$$C_{\text{Rel.Bet}} = \frac{\bar{C}_{\text{Bet}}(\text{PCP})}{\bar{C}_{\text{Bet}}(\text{Non}-\text{PCP})}$$

General statistical modeling approach in the three papers

1. Barnett et al (2012): Network-level cross-sectional

$$Y_i = \beta_0 + \beta_1 relbetcent_i + \beta_2^T X_i + \varepsilon_i$$

where *i* = hospital

2. An et al (2017): Network-level longitudinal:

$$Y_{it} = \alpha_i + \lambda_t + \beta_1^T netvar_{it} + \beta_2^T netvar_{it} t + \varepsilon_{it}$$

where *i* = state and *t* = year

3. Moen et al (2018): Network and within-network levels

$$\text{logit}(E[\text{InGuide}_{ijk}|\theta_i, \delta_{ij}]) = \beta_0 + \beta_1 \text{Covariates}_{ijk} + \beta_2 \text{ProvPos}_{ij} + \beta_3 \text{ReferralHospPos}_i + \beta_4 \text{ReferralHospStructure}_i + \beta_5 \text{SurgeryHospPos}_i + \beta_6 \text{SurgeryHospStructure}_\text{i} + \theta_i + \delta_{ij}$$

$$\text{where } \theta_i \sim \text{Normal}(0, \sigma^2) \text{ and } \delta_{ij} \sim \text{Normal}(0, \tau^2)$$

where *i* = HRR, j = provider, and k = patient while ref and surg denote the referral and the surgical hospital of the patient

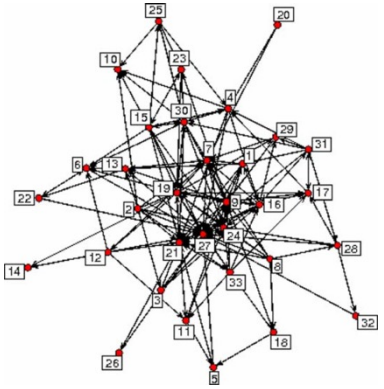
General bipartite or affiliation networks

- Network of physicians deduced from “shared patients” is an affiliation network
- Other affiliation networks
  - Network of scientists based on shared publications
  - Network of phenotypes based on shared genotypes (genes linked to multiple phenotypes)
  - Network of genotypes based on shared phenotypes or diseases
  - Network of diseases based on common comorbidities
- A wide-ranging topic!

2: Statistical analysis of relational data

- Observed network is the outcome
- Typically only observe network once (cross-sectional data)
  - Multiple observations (e.g., longitudinal data) becoming more common
- Network structure: Is global network structure explained by local configurations or sub-networks?
  - Closed dyads: reciprocity
  - Closed triads: transitivity, 3-cycles, …
- Social selection: Are individuals with similar characteristics more likely to form ties (homophily, assortative mixing)?
  - Do (latent) communities underlie the network?
- Non-standard and challenging statistical analyses required!

Model multivariate outcome: A ←

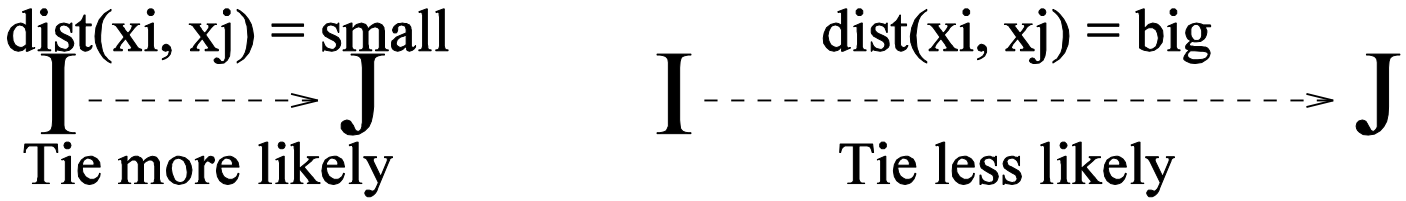


Why identify factors affecting relationship status?

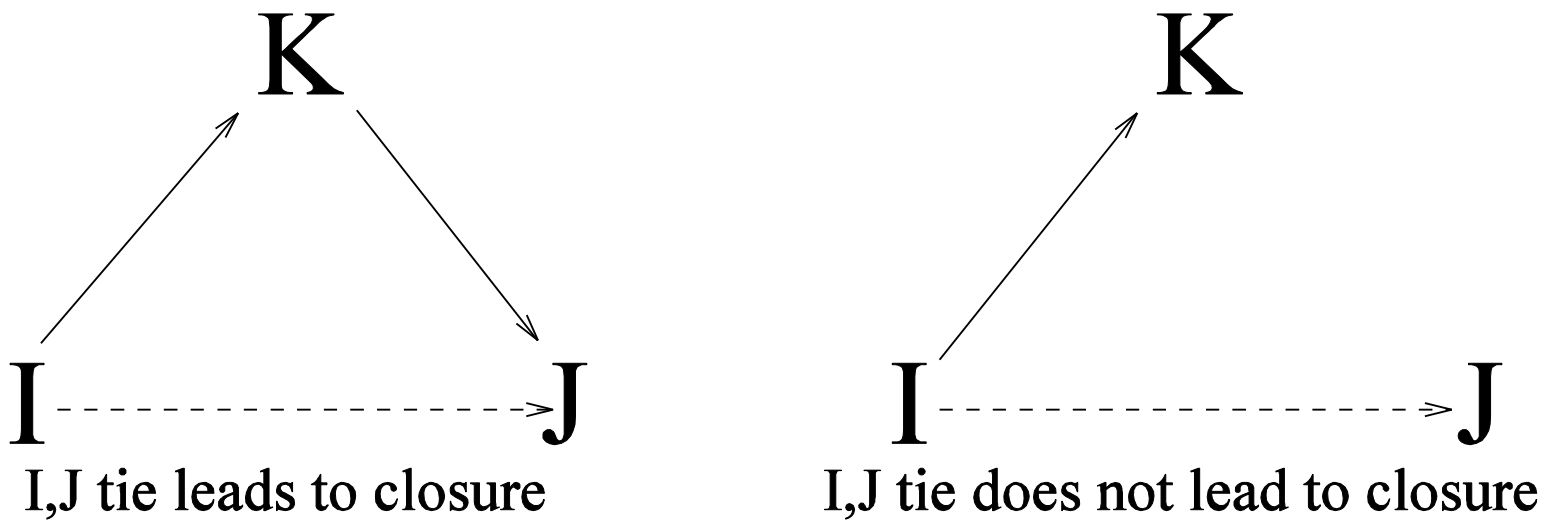
- Understand and replicate successful organizations
  - Replicate the best Accountable Care Organization (ACO)
- Provides recipe for developing network-based interventions
  - Interventions that change the network or an individual's position in the network
  - Determine factors that are most strongly associated with network status and use this to catalyze desired change in the network
  - Intervene on individuals' relationships with peers

Some Common Sociological Hypotheses

Homophily



Triadic Closure (Transitivity)



Model for random graph (Erdos-Renyi, 1959)

- Bernoulli or completely-random graph model:

$$A_{ij} \sim \text{Bernoulli}(p) \text{ for all } i, j$$

Key step!

implies that the model for the network is:

$$\Pr(A = a|\mu) = \prod_{i,j}^{N} p^{a_{ij}}(1-p)^{1-a_{ij}}$$

$$= \frac{\exp(\mu L)}{(1+\exp(\mu))^{N(N-1)}}$$

$$\propto \exp(\mu L) \text{ as a function of } A \text{ where}$$

$\exp(\mu) = \frac{p}{1-p}$, or equivalently $\mu = \log\left(\frac{p}{1-p}\right) = \text{logit}(p)$, and

$L = t_1(a) = \sum_{i \neq j} a_{ij}$ is the lone network statistic

- Use maximum likelihood to estimate parameter μ knowing that the resulting value yields the most likely network as a whole

Generalizing Erdos-Renyi Model

- Generalize model to allow tie-probabilities for an edge to depend on observed predictors, X_{ij}
- Further, allow each actor to have a different baseline propensity to form and to receive edges
- For example,

$$\text{logit}(\Pr(A_{ij} = 1|X_{ij})) = \mu + X_{ij}^T\beta + \alpha_i + \gamma_j$$
$$= \mu_{ij} + \alpha_i + \gamma_j$$

- $\mu_{ij} = \mu + X_{ij}^T\beta$ is the edge-specific systematic part of the linear predictor. Ignoring the denominator of the probabilities
  - $\Pr(A_{ij} = 1|X_{ij}) \propto \exp(\mu_{ij} + \alpha_i + \gamma_j)$
  - $\Pr(A_{ij} = 0|X_{ij}) \propto 1$
- α_i, γ_j could be fixed or random effects
  - Edge-independence or conditional edge-independence retained!
  - May use logistic regression (or mixed-effect) logistic regression to estimate model parameters

Dyadic Independence models

- Allowing reciprocity (statistical dependence) between ties from the same dyad
  (Dropping explicit appearance on X_{ij} in the following)
- Under edge-independence, $\text{logit}\left(\Pr(A_{ij}=1)\right) = \mu_{ij} + \alpha_i + \gamma_j$ implies

$$\Pr(A_{ij}=1) \propto \exp(\mu_{ij} + \alpha_i + \gamma_j) \text{ and thus that}$$
$$\Pr(A_{ij}=1, A_{ji}=1) \propto \exp(\mu_{ij} + \alpha_i + \gamma_j + \mu_{ji} + \alpha_j + \gamma_i)$$

- Therefore, a model that allows dyadic dependence ("reciprocity") is the multinomial dyad-independent model:

$$\Pr(A_{ij}=0, A_{ji}=0) \propto 1$$
$$\Pr(A_{ij}=1, A_{ji}=0) \propto \exp(\mu_{ij} + \alpha_i + \gamma_j)$$
$$\Pr(A_{ij}=0, A_{ji}=1) \propto \exp(\mu_{ji} + \alpha_j + \gamma_i)$$
$$\Pr(A_{ij}=1, A_{ji}=1) \propto \exp(\mu_{ij} + \alpha_i + \gamma_j + \mu_{ji} + \alpha_j + \gamma_i + \rho_{ij})$$

  where the proportionality constant is the sum of the above four probabilities

- Within-dyad independence occurs if $\rho_{ij} = \rho_{ji} = 0$
  - The extent of within-dyad dependence is governed by ρ_{ij}
- Heterogeneity in μ_{ij} and ρ_{ij} may be explained by covariates

P_1 Model of Network

- A special case is the P_1 model (Holland Leinhardt 1981) obtained by setting $\mu_{ij} = \mu$ and $\rho_{ij} = \rho$
- Then multiply the dyadic probabilities (on the previous slide) together to obtain the implied model for the network:

$$\Pr(A = a) \propto \exp\left(\mu t_1(a) + \sum_i \alpha_i t_{2i}(a) + \sum_j \gamma_j t_{3j}(a) + \rho t_4(a)\right)$$

where $t_1(a) = \sum_{i \neq j} a_{ij}$ is the number of ties, $t_{2i}(a)$ is the out-degree for actor *i*, $t_{3j}(a)$, is the in-degree for actor *j*, and $t_4(a)$ is the number of mutual dyads

Four (sets of) network statistics characterize network!

A closer look at the network statistics

$$\Pr(A = a) \propto \exp\left(\mu t_1(a) + \sum_i \alpha_i t_{2i}(a) + \sum_j \gamma_j t_{3j}(a) + \rho t_4(a)\right)$$

where:

- $t_1(a) = \sum_{i \neq j}^{N} a_{ij}$ is the total number of edges in the network
- $t_{2i}(a) = \sum_{j \neq i}^{N} a_{ij}$ is the out-degree for actor *i*,
- $t_{3j}(a) = \sum_{i \neq j}^{N} a_{ij}$, is the in-degree for actor *j*,
- $t_4(a) = \sum_{i<j}^{N} a_{ij} a_{ji}$ is the total number of mutual dyads in the network
- More powerful than the dyad census in terms of testing for mutuality and estimating the magnitude of mutuality (reciprocity)

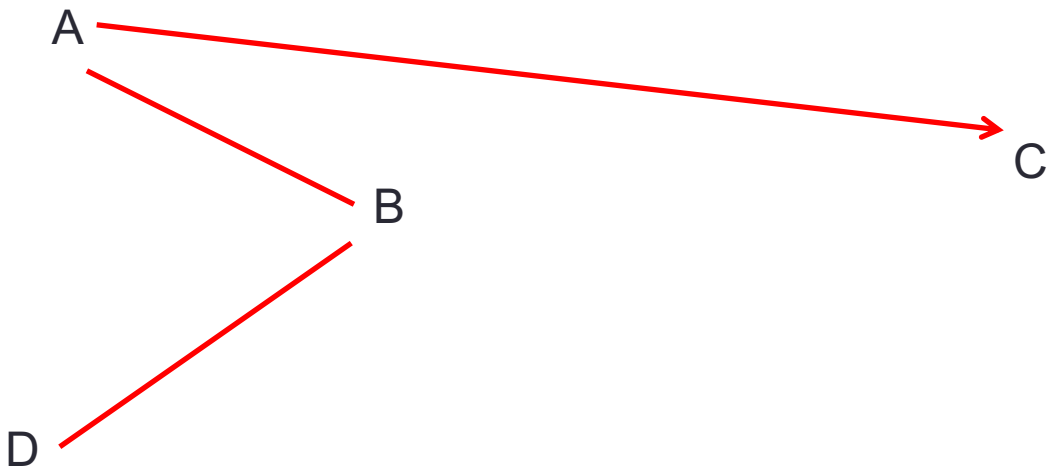
Dyadic Independent Model Extensions

- Stochastic block model: Allow density and reciprocity to vary between actors with different attributes (e.g., gender, age, ethnicity) (Fineberg and Wasserman 1981, Tanner and Wong 1987, Karrer and Newman 2010)
  - $\mu_{ij} = \mu_0 + \mu_1^T x_{ij}$; x_{ij} may include directional covariates
  - $\rho_{ij} = \rho_0 + \rho_1^T z_{ij}$; $z_{ij} = z_{ji}$ (covariates are symmetric)
- Model block as a latent variable (i.e., unobserved attribute)
  - Alternative to modularity-based community detection (Hoff 2008)
- Multiple-membership stochastic block models (Airoldi et al 2008; J Mach Learn Res)

Beyond dyadic independence

- Dyadic independence allows model for the network to be generated from the model for the dyad by multiplying the state probabilities for each dyad
- Dependence between dyads arises whenever the state of one dyad depends on the state of another dyad beyond actor-specific effects
- Triadic dependence: an edge is more (or less) likely to form if its actors have a common third actor
- Higher (k-)order dependence: the probability distribution of the state of the ties in the network depends simultaneously on the states of groups of ties involving $k-1$ other actors

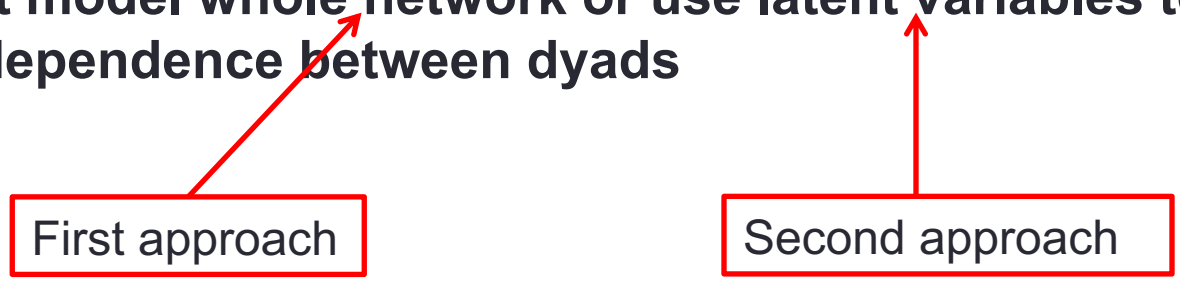
Illustration of structural constraint due to higher-order dependence



- Status of A-B-C triads restricts the possible states of the A-B-D triad!
  - Reduces from 8 to 4 possible states of the triad in a binary network
  - The triads are not structurally distinct nor separable

Quintessential Challenge of Sociocentric Data

- Triad = subgraph of three individuals
- Triads are not separable units
  - Observed state of one triad restricts state of triads with which it shares two actors
- ⟹ Cannot multiply dyad probabilities to generate corresponding model (or likelihood function) for the network
- **… Unit of observation becomes the whole network**
  - **Must model whole network or use latent variables to account for dependence between dyads**



Exponential random graph models (ERGMs)

- Key contributions: Handcock, Hunter, Butts, Goodreau, and Morris (2003); Robins, Pattison, Kalish, and Lusher (2007); Snijders, Pattison, Robins, and Handcock (2006)
- Let $t(a)$ denote a vector of network statistics
  - Functions of elements of adjacency matrix (A)
- An exponential random graph (or *p**) model has the form

$$\Pr(A = a; \theta) = K(\theta)^{-1} \exp\left(\sum_k \theta_k t_k(a)\right)$$

- where $K(\theta) = \sum_{a \in R(a)} \exp(\sum_k \theta_k t_k(a))$ and $R(a)$ is the set of all possible realizations for the network
- The dependent variable has $4^{N(N-1)/2}$ possible states
  - Observe one!
  - Does not in general factorize into models for dyads (only does so if dyadic independence holds)

Implied edge probability

- ERGMs imply the probability of a single edge is given by

$$\text{logit}\left(\Pr\left(a_{ij}=1\middle|a_{ij}^{c}\right)\right)=\sum_{k}\theta_{k}\delta_{k}(a_{ij}^{c})$$

  where a_{ij}^{c} denotes all ties other than a_{ij} and

$$\delta_{k}\left(a_{ij}^{c}\right)=t_{k}\left(a_{ij}^{c}\cup\{a_{ij}=1\}\right)-t_{k}\left(a_{ij}^{c}\cup\{a_{ij}=0\}\right)$$

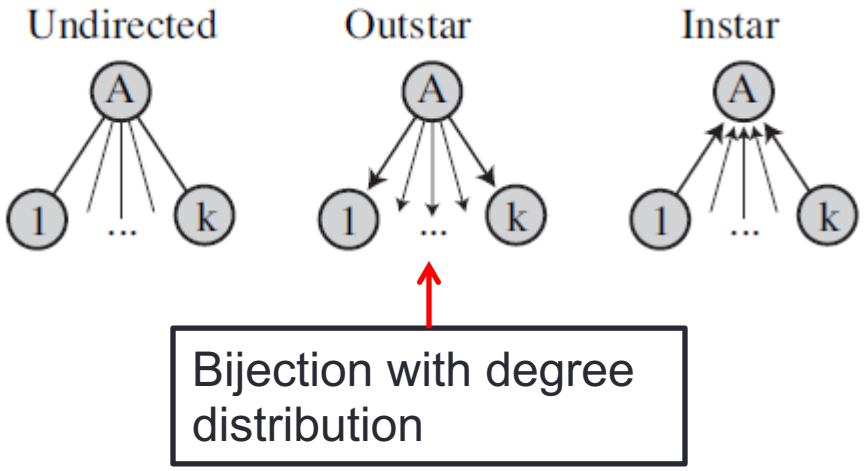
  is the difference in the network statistics when $a_{ij}=1$ to when $a_{ij}=0$

- θ_k is the log-odds of $a_{ij}=1$ to $a_{ij}=0$ if the presence of a_{ij} would lead to a one-unit increase in $t_k(a)$, conditioned on the rest of the network
- Key point: Model for edge probabilities does not in general imply model for the network
  - Does so only under edge- or dyadic-independence
  - From a statistical perspective, specifying network statistics to define model for the network seems back-to-front!

Common network statistics, $t_k(a)$

- Density: $\sum_{i \neq j} a_{ij}$
- k-out stars: $\sum_i \binom{a_{i+}}{k}$
- k-in stars: $\sum_i \binom{a_{+i}}{k}$
- Reciprocity: $\sum_{i<j} a_{ij}\, a_{ji}$
- Transitive triad: $\sum_{i<j<k} a_{ij}\, a_{ik} a_{jk}$
- 3-cycle: $\sum_{i<j<k} a_{ij}\, a_{jk} a_{ki}$

k-stars



- Nodematch or homophily covariate: $\sum_{i<j} a_{ij} I(x_i = x_j)$
- Preferential out-degree attachment: $\sum_{i<j} a_{ij}\, f(a_{i+} a_{j+})$
- For more extensive list see (e.g.,) Snijders et al (2007, SIENA manual)
- In R: ?ergm-terms

Estimation of ERGMs

- Markov Chain Monte Carlo (MCMC) methods allow inferences to be based on the true likelihood function
  - MLE via MCMC integration (Geyer and Thompson 1992)
  - Fully Bayesian estimation (Caimo arXiv:1703.05144v2 [stat.CO])
- Statnet R package (Handcock et al. 2003) can fit models on networks of modest size (in the 1000s of nodes)
- Obtaining convergence can be difficult because the likelihood surface often has a highly irregular shape
  - Traps at local maxima, failure to converge, or convergence to inappropriate "degenerate" solutions (Handcock 2003)
  - Degenerate solution: Individual simulated draws of the network under the model or estimated model are nothing like the observed network

Handling Degeneracy

- ERGMs with k-star, triadic or other higher-order terms are plagued by degeneracy
- If don't care about interpreting the higher-order effects but want to control for them, use a more general higher-order term whose parameters are constrained such that the effects of successive higher-order statistics balance each other
- For example, to account for triadic dependence, Snijders et al (2006) suggests using a measure that combines all *k*-triangles statistics into a single statistic

$$Alt_{\lambda}^{T}(a) = 3t_1(a) + \sum_{k=2}^{N-2} (-1)^k \frac{t_k(a)}{\lambda^{k-1}}$$

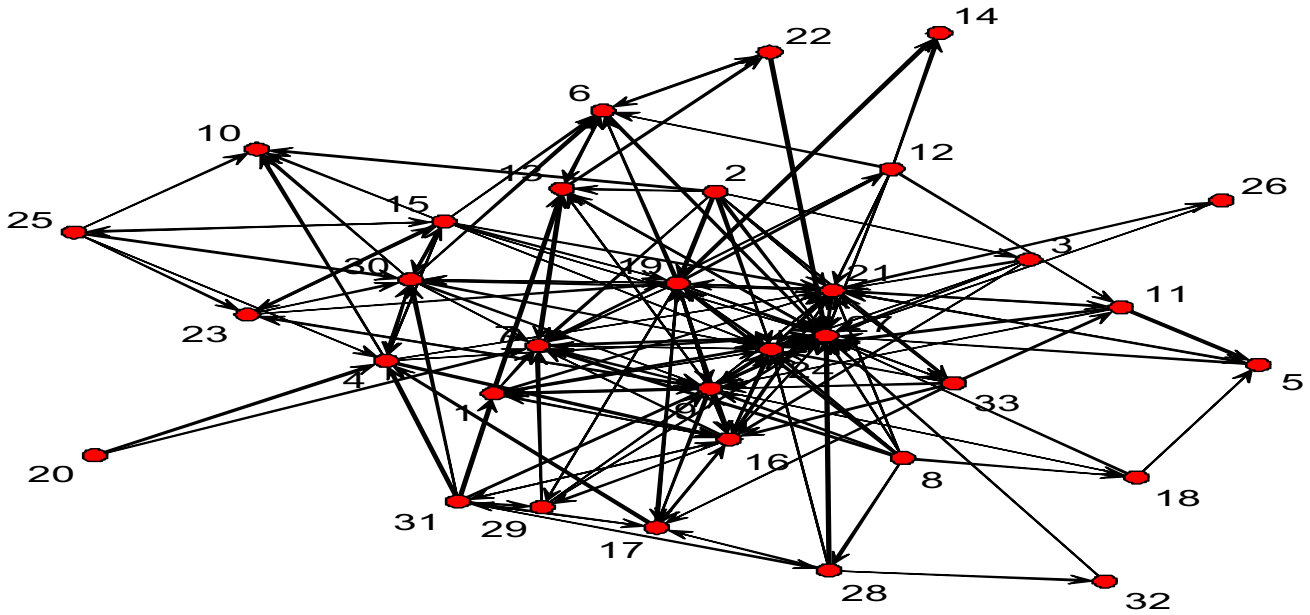
- where $t_k(a)$ is the number of sets of *k* individual triangles sharing a common base ("*k*-triangles")
  - $t_1(a) = t(a)$ is the count of the total number of triangles
  - $t_k(a) = -t_{k-1}(a)/\lambda$ for $k \geq 3$

Small physician practice example (Keating at al, 2007)

- R script: WorkshopICHPS2023models.R
- Goal: We are interested in whether presence of tie between physicians depends on the receiving physician having expertise in women's health, the proportion of female patients, and the number of clinical sessions they hold per week
- Enter network as an adjacency matrix (okay as network is small)
- Estimate ERGM with terms for network density, reciprocity, receiver covariates, and various similiarity (homophily) variables

Example plot and analysis of network (O'Malley and Marsden, 2008)

- pnet <- network(physnetwork, directed=TRUE, matrixtype="adjacency",
  vertex.attr=nodecov,
  vertex.attrnames = c("male", "whexpert",
  "pctwom", "numsess", "practice", "bcma", "bima",
  "bpp", "wnhlth", "numcat", "pctcat"))
- plot(pnet,mode = "fruchtermanreingold", displaylabels=T)



Simplest model

- model1a <- ergm(pnet~edges)
- Evaluating log-likelihood at the estimate.
- Formula: pnet ~ edges

- Iterations: 5 out of 20

- Monte Carlo MLE Results:
- Estimate Std. Error MCMC % p-value
- edges -1.70084 0.08517 0 <1e-04 ***
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- Null Deviance: 1463.9 on 1056 degrees of freedom
- Residual Deviance: 908.6 on 1055 degrees of freedom
-
- AIC: 910.6 BIC: 915.5 (Smaller is better.)

What is this the log-odds of?

Model of interest

- Formula: pnet ~ edges + mutual + nodeocov("whexpert") + nodeocov("pctwom") +
- nodeocov("numsess") + nodematch("male", diff = F) + nodematch("bcma",
- diff = F) + nodematch("bima", diff = F) + nodematch("bpp",
- diff = F) + nodematch("wnhlth", diff = F)

Ego (originator) covariate

Homophily covariate

- Monte Carlo MLE Results:
- Estimate Std. Error MCMC % p-value
- edges -4.560879 0.510615 0 < 1e-04 ***
- mutual 0.851292 0.293985 0 0.003862 **
- nodeocov.whexpert -0.391256 0.279232 0 0.161455
- nodeocov.pctwom -0.001583 0.004530 0 0.726830
- nodeocov.numsess 0.159686 0.039360 0 < 1e-04 ***
- nodematch.male 0.646502 0.181881 0 0.000396 ***
- nodematch.bcma 0.739658 0.278889 0 0.008119 **
- nodematch.bima 0.277871 0.199195 0 0.163321
- nodematch.bpp 1.396742 0.255284 0 < 1e-04 ***
- nodematch.wnhlth -0.046937 0.221694 0 0.832366
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- AIC: 838.3 BIC: 887.9 (Smaller is better.)

Re-estimate with separate node-match (homophily) coefficients

- model1f <- ergm(pnet~edges + mutual + nodeocov("whexpert") +nodeocov("pctwom") + nodeocov("numsess") + nodematch("male",diff=T) + nodematch("practice",diff=T))

Allows nodematch.male coefficient to vary by gender

	Estimate	Std. Error	MCMC %	p-value	
edges	-1.134295	0.417598	0	0.00671	**
mutual	0.615252	0.316615	0	0.05226	.
nodeocov.whexpert	-0.509118	0.293552	0	0.08315	.
nodeocov.pctwom	-0.022326	0.005604	0	< 1e-04	***
nodeocov.numsess	-0.007532	0.045438	0	0.86837	
nodematch.male.0	1.534152	0.247805	0	< 1e-04	***
nodematch.male.1	-1.112795	0.420071	0	0.00819	**
nodematch.practice.1	1.320874	0.849377	0	0.12022	
nodematch.practice.2	0.466084	0.215630	0	0.03088	*
nodematch.practice.3	2.131637	0.423777	0	< 1e-04	***
nodematch.practice.4	1.981935	0.489350	0	< 1e-04	***

Pros and Cons of ERGMs

- Model implied through specification of sufficient statistics of network; reverse of specifying model first and then seeking sufficient statistics
  - Pro: Allows tests for specific dependencies
  - Con: If network predictors include network configurations involving 3 or more actors, solutions susceptible to degeneracy (Handcock et al, 2003; Goldenberg et al, 2009)
  - Con: Computational barriers when *N* is ``large''
  - Con: Not generative
- Actor oriented (choice) models (Snijders, 2005, 2006) in SIENA are an alternative family of models that is less vulnerable to degeneracy
  - But feasible to estimate only on smaller networks

Conditional Dyadic Independence Models

- Instead of using network statistics to account for dyadic dependence, condition on latent variables
- Motivation: A random (or latent) effect extension of the P_1 model is the P_2 model (van Duijn et al, 2004):

$$\Pr(A_{ij} = a_{ij}, A_{ji} = a_{ji}) \propto \exp(\mu_{ij}a_{ij} + \mu_{ji}a_{ji} + \rho_{ij}a_{ij}a_{ji})$$

where

$$\mu_{ij} = \mu + \alpha_i + \gamma_j + \beta_1^T x_{1ij}$$
$$\alpha_i = a_i + \beta_2^T x_{2i}$$
$$\gamma_i = b_i + \beta_3^T x_{3i}$$
$$\rho_{ij} = \rho + \beta_4^T x_{4ij}$$

and x_{1ij}, x_{2i}, x_{3i}, and x_{4ij} are covariates impacting density, propensity to extend ties, propensity to receive ties, and propensity for mutual ties

- The extent of within-dyad dependence is governed by $\rho_{ij} = \rho_{ji}$
- $\begin{pmatrix} a_i \\ b_i \end{pmatrix} \sim N(0, \Sigma)$ is a bivariate latent variable capturing unexplained variation in and correlation of actors' propensities to "send" and "receive" ties

Latent Space Models

- The P_2 model is a conditionally dyadic independent models that provides a base for developing models that use latent variables to account for between dyad dependence
- These dyadic dependent models include:
  - Latent class models (Nowicki and Snijders, 2001; Airoldi et al 2008, 2010; Choi et al 2010)
  - Latent space and latent factor models (Hoff et al 2002, 2005, 2008; Paul, O'Malley et al 2014)
- Can estimate as mixed-effects models via two-step (iterative) maximum likelihood or as Bayesian models

Latent space models cont.

- Adapt P_2 model by using additional latent variables to account for dependence between dyads
- Peter Hoff and colleagues have led the development of such models (Hoff 2002, 2005, 2008)
- Add $\xi(z_i, z_j)$ to μ_{ij} where

$$\xi(z_i, z_j) = \begin{cases} \lambda_{z_i,z_j} \text{ where } z_i, z_j \in \{1, \ldots, K\} \text{ and } \lambda_{z_i,z_j} = \lambda_{z_j,z_i} \\ -|z_i - z_j|^c \text{ where } c > 0 \text{ and } z_i, z_j \text{ have } K \text{ elements} \\ z_i^T \boldsymbol{U} z_j \text{ where } z_i \sim N(0, \Sigma_z) \text{ and } U \text{ is a } K\text{-dimensional diagonal matrix} \end{cases}$$

- These correspond to z_i being a categorical latent variable (latent class or attribute), a position in continuous latent space, and a translated position, respectively.

Latent space models: interpretation

- Case 1 (latent class model) captures latent homophily by allowing individuals with the same value of the latent variable to have a greater probability of sharing an edge
- Cases 2 and 3 portray similarity as a distance measure → capture structural equivalence. The dimension of *Z* is arbitrary but generally is low
  - Case 2 model accounts for transitivity by requiring that the latent distances between actors obey the triangle inequality
  - Case 3 model accounts for both structural equivalence and latent homophily (bilinear model when $U = I$)
- Hoff (2015) developed and contributed the amen package in R to estimate all three cases
- Some latent space models can be estimated using the (older) latentnet package in R
  - Enter network into R as for using StatNet to estimate ERGMs

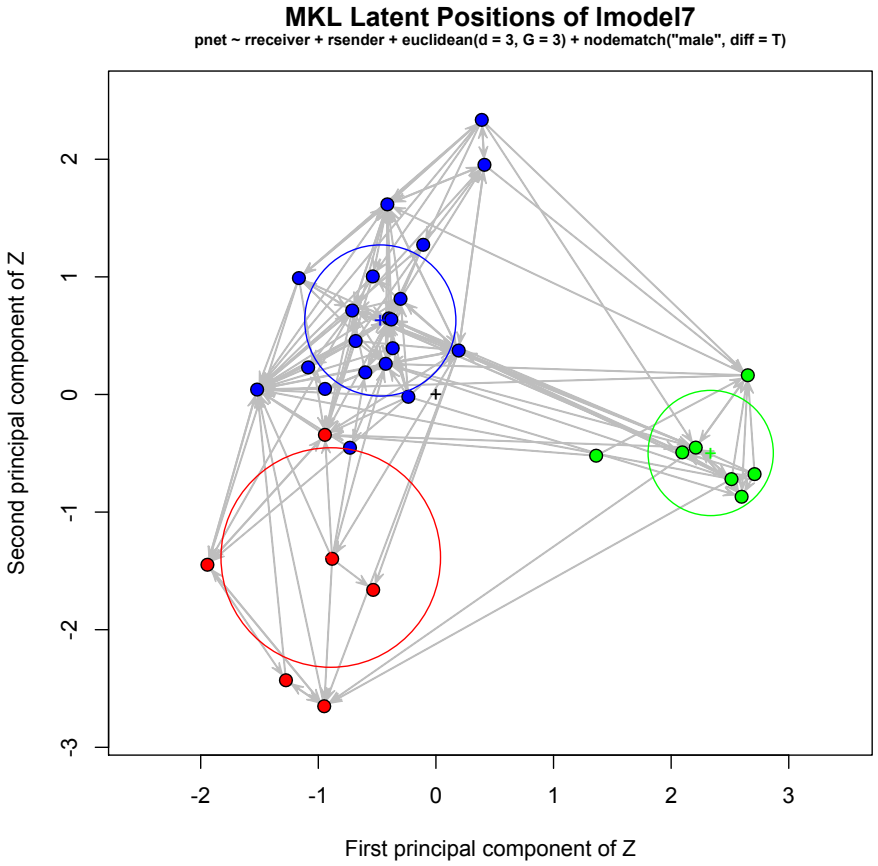
Estimating Latent-space models in R (see Github R scripts for more detail)

- ergmm(formula, response = NULL, family = "Bernoulli", fam.par = NULL, control = control.ergmm(), user.start = list(), prior = ergmm.prior(), tofit = c("mcmc", "mkl", "mkl.mbc", "procrustes", "klswitch"), Z.ref = NULL, Z.K.ref = NULL, seed = NULL, verbose = FALSE)
  - "family" command allows different distributions in the exponential family (as for generalized linear models)
  - "prior" command gives some control over prior distributions for Bayesian analysis
  - "tofit" controls which estimation methods are used
- ?ergmm to get help and then terms.ergmm to get list of terms that are supported
  - Note that there are no mutual, triadic or higher-order network statistics allowed in this model!

Euclidean distance with $d = 3$ dimensions and 3 clusters (or groups)

```
lmodel7 <- ergmm(pnet ~
rreceiver+rsender+euclidean(d=3,G=3)+nodematch("male",diff=T))
```

An application is to cluster actors into groups!



3. Statistical analyses of social influence or peer effects

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL ARTICLE

The Spread of Obesity in a Large Social Network Over 32 Years

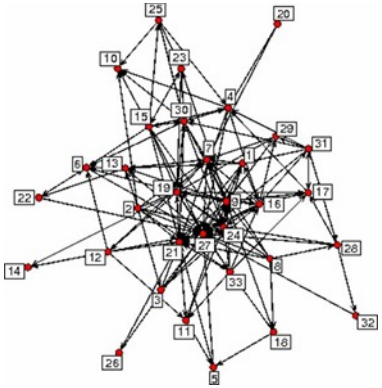
Nicholas A. Christakis, M.D., Ph.D., M.P.H., and James H. Fowler, Ph.D.

Peer Effects

- One person(s) influence on another
- Also referred to as "social influence, induction or contagion"
- History: Modeling diffusion of innovations
  - Coleman et al (1957): Social network/social structure related to diffusion of information among physicians
  - Coleman et al (1966): Diffusion of tetracycline, a new medical drug, among doctors (more interpersonal connections = faster adoption)

Peer and Spillover Effects

- Endogenous peer effects or contagion et al: Does the behavior of my peers affect my own behavior?
- Exogenous peer effects or spillover effects: Does the treatment received by my peers affect my outcome (above and beyond my treatment)?
  - Violates stable unit treatment value assumption (SUTVA)
- Network defines predictors: Overlapping groups of individuals yield the predictor(s) of individuals' outcomes!



* $Y, X = Y_{peer}, X_{peer}$: Predictors

Why do peer effects matter in Medicine?

1. Justify behavioral interventions reliant on influence between persons
   - Peer-support and group interventions
   - May be enhanced by targeted ("seeded") interventions
2. Evaluate full effect of an intervention (“collateral effects”)
   - Intervene on the untreated
   - Account for spillover effects (Sobel 2006)

Dyadic Influence: Longitudinal data

- Longitudinal is the easier case!
- Focus on a single dyad for now (individuals *i*, *j*)

Y_{it} = Outcome for individual *i* at time *t*

X_{it} = Other characteristics of individual *i* at *t*

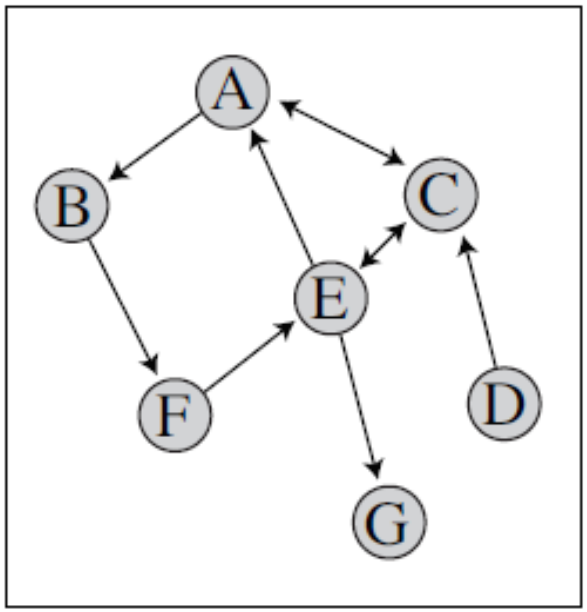
- Example Model:

$$Y_{it} = \beta_1 Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + \beta_3 X_{i(t-1)} + \theta_i + \varepsilon_{ijt}$$

- Longitudinal, lagged version of an Autoregressive Outcome Model (O'Malley and Marsden 2008)
- $\beta_1 > 0$: positive effect of *j*'s outcome on subsequent outcome for *i*
- $\theta_i, \varepsilon_{ijt}$ are latent variables, error terms
- May add $X_{j(t-1)}$ as an additional predictor
- Roles of *i* and *j* reversed in another observation!
  - Reflection problem (Manski, 1993)
  - Lagged predictors allows effect of *i* on *j* to be distinguished from the feedback effect of *j* on *i*
  - Model is unnatural when there are multiple peers (*j*) for a focal individual (*i*)

Beyond Dyads: Network Influence through outbound edges

Digraph



Influence Matrix, W

	A	B	C	D	E	F	G
A	0	1/2	1/2	0	0	0	0
B	0	0	0	0	0	1	0
C	1/2	0	0	0	1/2	0	0
D	0	0	1	0	0	0	0
E	1/3	0	1/3	0	0	0	1/3
F	0	0	0	0	1	0	0
G	1/6	1/6	1/6	1/6	1/6	1/6	0

Row stochastic matrix

Network Influence Models

- Multiple peers: One specification of W_t assumes influence proportional to exclusivity of connections
- Let $\bar{Y}_{-i(t-1)} = [W_{t-1}Y_{t-1}]_i$ denote alters' average trait, where W_{t-1} is the influence matrix and Y_{t-1} is the vector of the trait of interest, at time $t-1$
  - Influence transmits via weighted average of trait in peers
- Simple network influence model:

$$Y_{it} = \beta_1 \bar{Y}_{-i(t-1)} + \beta_2 Y_{i(t-1)} + \beta_3 X_{i(t-1)} + \theta_i + \varepsilon_{it}$$

- Lots of other specifications!
  - For example, don't make W_{t-1} row stochastic
  - Base W_{t-1} on geodesic distances as opposed to the adjacency matrix
- Estimation straight-forward when predictors are lagged
  - Use procedures for hierarchical or mixed-effect models

Examples: Hospital and physician-level diffusion

- Example 1: Hospital peer-to-peer influence of implantable cardiac defibrillator (ICD) utilization over time
  - Peer effects: elementary unit of diffusion
  - Is there evidence of physician-physician influence on ICD use and guideline consistent ICD use?
    - If so, is influence modified by structural position in network?
    - → Justify prioritizing select physicians for interventions
  - O'Malley, Moen, Bynum, Austin, and Skinner (2020)

- Example 2: Spillover effect of another physician's patient having an adverse reaction following a colonoscopy (beyond effect of adverse reactions within own patient cohort)
  - With Keating, Landon, and Onnela (2017)
  - Not discussed today

Peer Effects of Hospital ICD equipped status

- Let y_{it} denote ICD status (1 = equipped, 0 = not-equipped) of hospital *i* at time *t*
- Key predictor is weighted average of the vector y_t over the peer hospitals of hospital *i*
- We used the network strength (number of shared patients) of the edges between the hospitals as weights in W_{t-1}
- Thus, model has the form

$$y_{it} | \, y_{i(t-1)} = j \sim Bernoulli(p_{it})$$

where $logit(p_{it})$

$$= \theta_{ij} + \beta_{1j} \, x_{i(t-1)} + \beta_{2j} [W_{t-1} Y_{t-1}]_i$$

$\theta_{ij} \sim Normal(\beta_{0j}, \tau_j^2)$ is a random effect for hospital and $x_{i(t-1)}$ is a vector of control predictors

Example: Heterogeneous peer-effects of Hospital ICD equipped status with peer geographic proximity adjustment

- Full model interacts $[W_{t-1}Y_{t-1}]_i$ with hospital i's network strength
- Model given by: $logit(p_{it})$

$$= \theta_{ij} + \beta_{1j}\, x_{i(t-1)} + \beta_{2j}[W_{t-1}Y_{t-1}]_i + \beta_{3j}[G_{t-1}Y_{t-1}]_i + \beta_{4j}[W_{t-1}Y_{t-1}]_i d_{i(t-1)}$$

where $d_{i(t-1)}$ is the network weighted degree (i.e., strength) of hospital i at time $t-1$

- G_{t-1} is a weight matrix based on physical (geographic) distances
- Could add additional ego and peer variables for number of implants and referrals to account for propensity of adoption and de-adoption of ICD technology ("Equipped status")

ICD Adoption of Equipped Status

ICD Adoption: 306 hrrs, 3720 hospitals, 12716 observations			
Term	Estimate	z-value	p-value
Lag network strength	-1.593	-2.67	0.008
Lag peer equipped	-0.391	-1.29	0.198
Lag peer equipped*network strength	2.295	2.81	0.005
Lag peer referral	0.268	0.58	0.564
Lag peer implant	-0.146	-0.58	0.561
Lag geographic equipped	22.750	4.67	0.000
Lag geographic referral	-0.607	-4.30	0.000
Lag geographic implant	-0.059	-1.27	0.204
Var(hospital, HRR)	1.15 +/- 1.07, 0.49 +/- 0.70		

A non-capable hospital with strong connections to peer hospitals equipped to implant ICDs is more likely to become equipped to implant ICDs

ICD Continuation of Equipped Status

ICD De-adoption: 305 hrrs, 1410 hospitals, 4418 observations			
Term	Estimate	z-value	p-value
Lag network strength	-1.670	-3.01	0.003
Lag peer equipped	-1.456	-3.27	0.001
Lag peer equipped*network strength	2.216	2.75	0.006
Lag peer referral	-0.055	-0.07	0.943
Lag peer implant	-0.322	-1.14	0.254
Lag geographic equipped	-4.753	-1.15	0.248
Lag geographic referral	-0.042	-0.29	0.773
Lag geographic implant	-0.029	-0.78	0.433
Var(hospital, HRR)	0.84 +/- 0.92, 0.00 +/- 0.00		

An ICD capable hospital with strong connections to to peer hospitals equipped to implant ICDs is more likely to remain equipped to implant ICDs

Cross-sectional social influence model

- If don't have longitudinal data, it becomes necessary to model the whole network as a system of simultaneous equations

$$Y_i = \beta_1 \bar{Y}_{-i} + \beta_3 X_i + \varepsilon_i$$
$$\equiv Y_i = \beta_1 [WY]_i + \beta_3 X_i + \varepsilon_i$$

- Expressed vector form:

$$Y = \beta_1 WY + X\beta_3 + \varepsilon$$
$$\Rightarrow (I - \beta_1 W)Y = X\beta_3 + \varepsilon$$
$$\Rightarrow Y = (I - \beta_1 W)^{-1} X\beta_3 + (I - \beta_1 W)^{-1}\varepsilon$$

- $E[Y|X] = (I - \beta_1 W)^{-1} X\beta_3$
- $Var(Y|X) = (I - \beta_1 W)^{-1} Var(\varepsilon)(I - \beta_1 W)^{-T}$
- Network analogue to an AR(1) structure in longitudinal or time-series models

Estimation of Network Autocorrelation Models in R

- Likelihood-based methods typically assume normally distributed errors
- Use lnam function
  - Uses numerical approximation to second derivatives in Newton-family optimization routines
- Develop your own estimation routines
  - Use optim function for maximum likelihood based estimation
- Amenable to Bayesian inference

Estimation of linear regression and autoregressive outcome models for cross-sectional network influence of level of use of hormone replacement therapy (hrt)

Data manipulation in R

- on <- as.vector(rep(1,nr))
- x <- as.matrix(cbind(on,regdata[,c("male", "pctwom","numalters")]))
- hrtalt <- wtreldir %*% as.vector(covdata$sumhrt)
- regdata <- data.frame(covdata,hrtalt=hrtalt,noalters=noalters,numalters= numalters)

Estimation in R using reg (don't do) and lnam (good) functions

- reg.adj <- lm(sumhrt~x+hrtalt-1, data=regdata)
- lnam1.adj <- lnam(regdata$sumhrt,x,wtreldir)

Network Autocorrelation Model

- Like network autocorrelated outcome model but assume network effects act on the residuals

$$Y_i = \beta X_i + \varepsilon_i$$
$$\varepsilon_i = \rho \bar{\varepsilon}_{-i} + \delta_i = \rho [W\varepsilon]_i + \delta_i$$

- where $\{\delta_i\}_{1:N}$ are independent random variables
- Expressing the model in vector form:

$$\varepsilon = \rho W \varepsilon + \delta$$
$$\Rightarrow (I - \rho W)\varepsilon = \delta$$
$$\Rightarrow Y = X\beta + (I - \rho W)^{-1}\delta$$

- $E[Y|X] = X\beta$
- $Var(Y|X) = (I - \rho W)^{-1} Var(\delta)(I - \rho W)^{-T}$
- Network analogue to a moving average correlation structure in longitudinal or time-series analysis

- Estimation in R:
- lnam2.adj <- lnam(regdata$sumhrt,x,NULL,wtreldir)

Similarity of Cross-sectional Social Influence models to Spatial Statistics

- Using sna, gdist <- geodist(reldir)$gdist yields geodesic distances
  - You can use this to define a W matrix whose elements reflect the closeness of geodesic distances
  - Use 0 weight if geodesic distance is infinite

- Adjacency or geodesic distance network matrix → Spatial matrix of physical distances
  - Similarity between binary-valued network data and areal spatial data
  - Similarity between geodesic network distance weight matrix and point referenced spatial data (known distances between observation locations)

- Subtle issues arise due to interference between dyads in social network data
  - No longer have “isolated experiments”
  - Stable Unit Treatment Value Assumption (SUTVA) violated!

Social Influence Analysis Challenges

- Longitudinal data helps with identification of causal effects
  - Helps alleviate concerns of reverse causality, simultaneity, …
  - Avoids reliance on strong parametric assumptions
- Statistical analysis challenging if seek causal inferences when dyads not formed at random!

Causality concerns related to social influence of a health behavior

- Homophily (U): “Birds of a feather flock together”
  - Seek individuals with similar habits (e.g., over-eating, smoking) or medical practices then become friends or professional colleagues
  - Tie-dissolution due to diverging traits

- Unmeasured common cause (C)
  - Propaganda about healthy living
  - New fad diet
  - New drug
  - Unknown peers in common
  - Conference attendance or unknown publication of guidelines

How to identify the causal effect of social influence?

- Use a joint modeling approach:
  - Specify parametric distribution (e.g., multivariate normality)
  - Assumptions not conclusively testable (O'Malley and Marsden 2008)
  - Steglich et al. (2010)
- Instrumental variable (IV) methods provide hope but rely on structural assumptions holding
  - Candidate IV: peers' genetic alleles of phenotype of interest
  - Due to concerns over population stratification and possibly pleiotropy, IV needs to be time-varying (O'Malley et al 2014)

Conclusion

- Social network analysis is a very broad field
  - Many disciplines involved
- Comparison of networks of different organizations, analysis of social selection ("homophily"), and analysis of social influence ("peer effects") are structurally different problems!
- Methods for the analysis of social network data often differ from standard methods in statistics
  - Interference of observations!
  - One person's "treatment" may affect another's "outcome"
- Networks are a new and growing topic in statistics; lots to be done
  - A lot that we did not cover today!!!

Thank you for your attention!

References: Descriptive network measures

- Wasserman S. and Faust K. (1994), Social Network Analysis. Cambridge: Cambridge University Press
- Bonacich, P. (1987), Power and Centrality: A Family of Measures," American Journal of Sociology, 92, 1170-1182
- Freeman, L. (1979), Centrality in Social Networks, Conceptual Clarification, Social Networks, 1, 215-239
- Faust, K. (1997), Centrality in Affiliation Networks, Social Networks, 19, 157-191
- O'Malley, A. J. and Marsden, P. V. (2008), The Analysis of Social Networks, Health Services & Outcomes Research Methodology, 8, 222-269

References: Bipartite Networks and Comparison of Multiple Networks

- Borgatti, S. and Everett, M. (1997), Network Analysis of 2-Mode Data," Social Networks, 19, 243-269
- Barnett, M. L., Christakis, N. A., O'Malley, A. J., Onnela, J.-P., Keating, N. L., and Landon, B. E. (2012), Physician Patient-Sharing Networks and the Cost and Intensity of Care in US Hospitals," Medical Care, 50, 152-160
- Landon, B. E., Keating, N. L., Barnett, M. L., Onnela, J. P., Paul, S., O'Malley, A. J., Keegan, T., and Christakis, N. A. (2012), Variation in Patient-Sharing Networks of Physicians Across the United States, Journal of the American Medical Association, 308, 265-273.

References: Bipartite Networks and Comparison of Multiple Networks cont.

- An C, O'Malley AJ, Rockmore DN, Stock CD. Analysis of the U.S. Patient Referral Network. Statistics in Medicine, 2018, 37, (5), 847-866. doi: 10.1002/sim.7565. PMID: 29205445
- Moen EL, Bynum JPW, Austin AM, Chakraborti G, Skinner JS, O'Malley AJ. Assessing variation in implantable cardioverter defibrillator therapy guideline adherence with physician and hospital patient-sharing networks. Medical Care, 2018, 56, (4), 350-357
- Moen EL, Austin AM, Bynum JP, Skinner JS, O'Malley AJ. An analysis of patient-sharing physician networks and implantable cardioverter defibrillator therapy. Health Services and Outcomes Research Methodology, 2016, 16, 132-153
- Landon BE, Keating NL, Onnela J-P, Zaslavsky AM, Christakis NAC, O'Malley AJ. Patient-Sharing Networks of Physicians and Healthcare Utilization and Spending Among Medicare Beneficiaries. JAMA Internal Medicine, 2018, 178 (1), 66-73
- Keating NL, O'Malley AJ, Onnela J-P, Landon BE. Assessing the impact of colonoscopy complications on use of colonoscopy among primary care physicians and other connected physicians: an observational study of older Americans, 2017, BMJ Open, 7, (6), e014239. doi:10.1136/bmjopen-2016-014239
- Moen EL, Bynum JPW, Skinner JS, O'Malley AJ. Physician network position and patient outcomes following implantable cardioverter defibrillator therapy. Health Services Research, 2019, 54 (4), 880-889. PMID: 30937894

References: General Network Analyses and Applications

- Barabasi, A.-L. and Albert, R. (1999), Emergence of Scaling in Random Networks, Science, 286, 509-512
- Newman, M. E. J. (2006), Modularity and community structure in networks, Proceedings of the National Academy of Sciences of the United States of America, 103 (23): 8577–8696
- Keating, N. L., Ayanian, J. Z., Cleary, P. D., and et al (2007), Factors affecting influential discussions among physicians: a social network analysis of a primary care practice, Journal of general internal medicine, 22, 794-798
- Coleman, J., Katz, E., and Menzel, H. (1957), The diffusion of innovations among physicians, Sociometry, 20, 253-270
- Coleman, J., Katz, E., and et al. (1966), Medical Innovation: A Diffusion Study, Bobbs-Merrill
- Pollack, C. E., Soulos, P. R. & Gross, C. P. Physician's peer exposure and the adoption of a new cancer treatment modality. *Cancer* **121,** 2799–2807 (2015).
- Hidalgo, C. A., Blumm, N., Barabasi, A.-L., and Christakis, N. A. (2009), A Dynamic Network Approach for the Study of Human Phenotypes, PLoS Computational Biology, 5, e1000353. doi:10.1371/journal.pcbi.1000353

References: cross-sectional models of relationships (social selection)

- Erdos, P. and Renyi, A. (1959), Random Graphs, Publicationes Mathematicae, 6, 290:297.
- Wang, W. and Wong, G. (1987), Stochastic Blockmodels for Directed Graphs, Journal of the American Statistical Association, 82, 8-19.
- Choi D, Wolfe P. and Airoldi E. (2010) Stochastic blockmodels with growing number of classes. arXiv:1011.4644.
- Fienberg, S. and S. Wasserman. (1981). Categorical data analysis of single sociometric relations. In Sociological Methodology, edited by S. Leinhardt, pp. 156-92. San Francisco: Jossey-Bass.
- Frank, O. and D. Strauss. (1986). Markov graphs. Journal of the American Statistical Association 81: 832-42.
- Goldenberg, A., Zheng, A. X., Fineberg, S. E. and Airoldi, E. M. (2009). A survey of statistical network models. In press: Foundations and Trends in Machine Learning.
- Handcock, M. S., Robins, G. L., Snijders, T. A. B., Moody, J. and Besag, J. (2003). Assessing degeneracy in statistical models of social networks. Journal of the American Statistical Association, 76: 33-50.
- Holland, P. and S. Leinhardt. (1981). An exponential family of probability-distributions for directed-graphs. Journal of the American Statistical Association 76(373): 33-50.

References: cross-sectional models of relationships cont.

- Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. (2010), ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks, http://CRAN.R-project.org/package=ergm. Version 2.2-6. Project home page at http://statnetproject.org
- Handcock, M. S., Robins, G. L., Snijders, T. A. B., Moody, J., and Besag, J. (2003), Assessing degeneracy in statistical models of social networks, Journal of American Statistical Association, 76, 33-50
- Robins, G. L., Snijders, T. A. B.,Wang, P., Handcock, M. S., and Pattison, P. E. (2007), Recent developments in exponential random graph (p∗) models for social networks, Social Networks, 29, 192-215
- Hoff, P. (2005). Bilinear mixed-effects models for dyadic data. Journal of the American Statistical Association 100: 286-95.
- Hoff, P., A. Raftery, and M. Handcock. (2002). Latent space approaches to social network analysis. Journal of the American Statistical Association 97: 1090-98.
- Hoff P. (2008), Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data, in: Advances in Neural Information Processing Systems, volume 20, MIT Press, 657-664
- Paul S, Keating NL, Landon BE, O'Malley AJ. Results from using a new dyadic-dependence model to analyze sociocentric physician networks. Social Science & Medicine, 125, 2015, 51-59

References: dynamic network models of relationships

- Krivitsky P.N. and Handcock M.S. (2010). A Separable Model for Dynamic Networks. arXiv:1011.1937v1[stat.ME].
- Nowicki K and Snijders T.A.B. (2001). Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 96, 1077-1087.
- O'Malley AJ, Christakis NA. (2011). Longitudinal Analysis of Large Social Networks: estimating the Effect of Health Traits on changes in Friendship Ties. Statistics in Medicine, 30, 9, 950-964
- Paul S. and O'Malley A.J. (2013). Hierarchical longitudinal models of relationships in social networks. Journal of the Royal Statistical Society, Series C.
- Pattison, P. and S. Wasserman. (1999). Logit models and logistic regressions for social networks: II. Multivariate relations. British Journal of Mathematical and Statistical Psychology 52 (Pt 2): 169-93.
- Van Duijn, M., T. Snijders, and B. Zijlstra. (2004). A Random Effects Model with Covariates for Directed Graphs. Statistica Neerlandica 58(2): 234-54.

References: dynamic network models of relationships cont.

- Snijders, T. A. B. (2006). "Statistical methods for network dynamics." In S. R. Luchini et al., editors, Proceedings of the XLIII Scientic Meeting, Italian Statistical Society, pages 281-296, Padova: CLEUP.
- Snijders, T.A.B. (2005), Models for longitudinal social network data," in Models and Methods in Social Network Analysis, Cambridge University Press, 215-247.
- Steglich, C.E.G., Snijders, T.A.B. and Pearson, M. (2010). Dynamic Networks and Behavior: Separating Selection from Influence. Sociological Methodology, 40, 329-393.
- Westveld, A. H. and Hoff, P. D. (2011), A Mixed Effect Model for Longitudinal Relational and Network Data, With Applications To International Trade and Conflict," The Annals of Applied Statistics, 5, 843-872.

References: social influence analyses

- Christakis NA, Fowler JH. (2007). The Spread of Obesity in a Large Social Network over 32 Years. New England Journal of Medicine, 357, 370--379
- Christakis NA, Fowler, JH. (2008). Dynamics of Smoking Behavior in a Large Social Network. New England Journal of Medicine, 358, 2249--2258
- Fowler JH, Christakis NA. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. British Medical Journal, 337, doi:10.1136/bmj.a2338.
- Cohen-Cole E, Fletcher JM. (2008). Detecting Implausible Social Network Effects in Acne, Height, and Headaches: Longitudinal Analyses. British Medical Journal, 337, a2533.
- Lyons, R. (2011), The spread of evidence-poor medicine via flawed social network analyses, Statistics, Politics and Policy, 2, 1-26, doi:10.2202/2151-7509.1024
- Shalizi CR, Thomas AC. (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies, Sociological Methods and Research, 40, 211--239.
- O'Malley AJ, Elwert F, Rosenquist JN, Zaslavsky AM, Christakis NA. Estimating peer effects in longitudinal dyadic data using instrumental variables. Biometrics, 2014, 70, 3, 506–515. (Published online: 29 APR 2014, DOI:10.1111/biom.12172).

References: social influence analyses cont.

- Christakis, N. A. and Fowler, J. H. (2013), Social Contagion Theory: Examining Dynamic Social Networks and Human Behavior," Statistics in Medicine, 32, 556-577
- Manski, C. A. (1993), Identfication of endogenous social effects: The Reflection Problem, Review of Economic Studies, 60, 531-542
- Marsden, P. V. and Friedkin, N. E. (1993), Network Studies of Social Influence, Sociological Methods and Research, 22, 127-151
- McPherson, M. L., Smith-Lovin, Cook, and et al (2001), Birds of a Feather: Homophily in Social Networks, Annual Review of Sociology, 27, 415-444
- VanderWeele, T. J. (2011), Sensitivity Analysis for Contagion Effects in Social Networks, Sociological Methods & Research, 40, 240-255
- VanderWeele, T. J., Ogburn, E. L., and Tchetgen Tchetgen, E. J. (2012), Why and When "Flawed" Social Network Analyses still yield Valid Tests of no Contagion, Statistics, Politics, and Policy, Manuscript 1050
- Keating NL, O'Malley AJ, Onnela J-P, Gray SQ, Landon BE. Influence of Peer Physicians on Intensity of End-of-Life Care for Cancer Decedents. Medical Care, 57, 6, 468-474. PMID: 31033059

References: network-influence related models

- Land KC, Deane G. (1992). On the Large-Sample Estimation of Regression Models with Spatial or Network Effect Terms: A Two-Stage Least-Squares Approach. Sociological Methodology, (ed: Peter V. Marsden), Oxford, UK: Basil Blackwell, Ltd, 221-248
- Anselin L. (1988). Spatial Econometrics: Methods and Models, Kluwer Academic Publishers: Dordrecht, The Netherlands
- Sargan JD. (1958). The Estimation of Econometric Relationships Using Instrumental Variables, Econometrica, 26, 393-415
- Haining, R.P. (1978). The moving average model for spatial interaction. Transactions of the Institute of British Geographers, 3, 202-225
- Kelejian, H.H., Robinson, D.P., 1993. A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. Papers in Regional Science 72, 297-312
- O'Malley AJ, Moen EL, Bynum JPW, Austin AM, Skinner JS. Modeling Peer Effect Modification by Network Position: The Diffusion of Implantable Cardioverter Defibrillators in the US Hospital Network. In Press: Statistics in Medicine