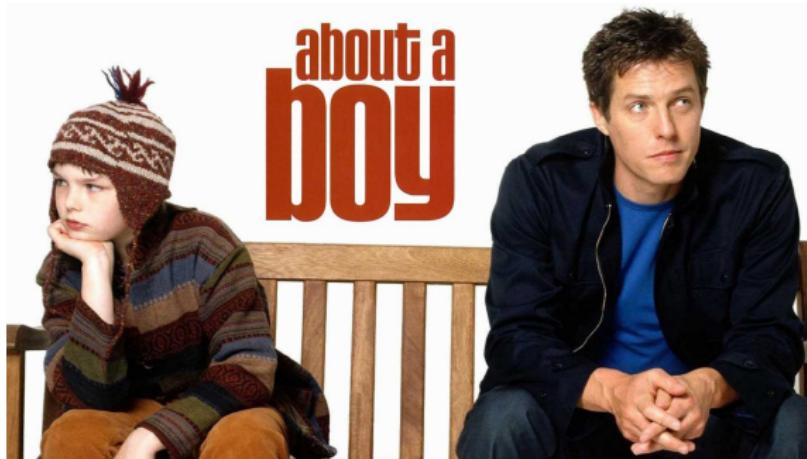


Social Network Analysis in Comparative Politics

Chao-Yo Cheng

Tsinghua University

November 26, 2019



"No man is an island entire of itself; every man is a piece of the continent, a part of the main."

— John Donne, *Devotions upon Emergent Occasions* (1624)

Motivation

- ▶ Why network?
- ▶ What is a network?
- ▶ How do we study a network?

Why network?

- ▶ What explains different varieties of social, political, and economic outcomes?

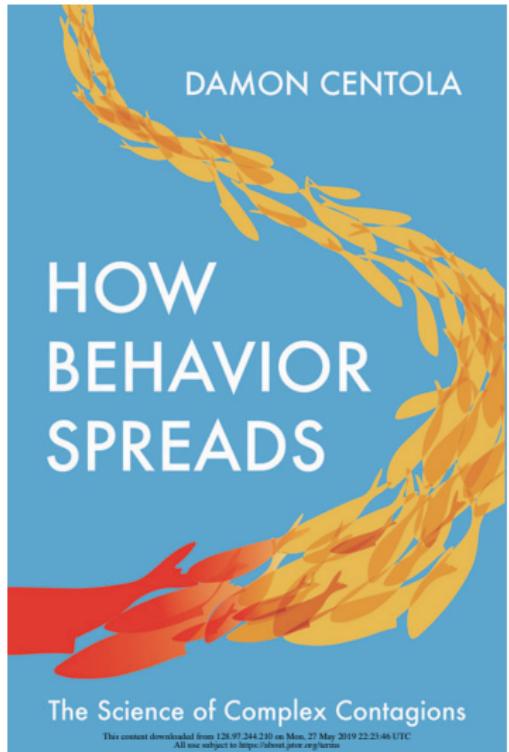
<!--

Why network?

- ▶ What explains different varieties of social, political, and economic outcomes?
<!-->
- ▶ Conventional approaches focus on the traits of individual observations.
- ▶ The outcomes can be a function of individual's ties with others as well as many features of the networks that individuals belong to.
<!--

Why network?

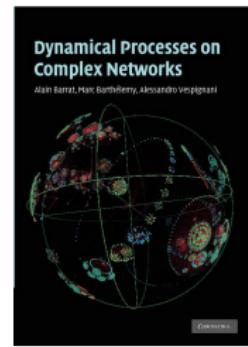
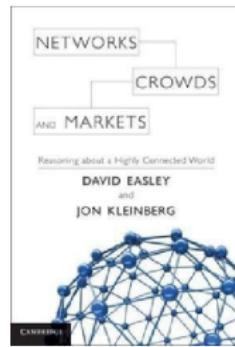
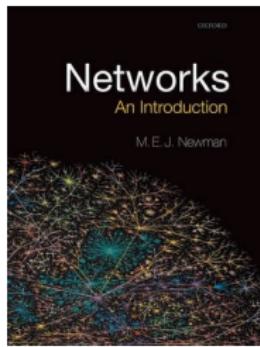
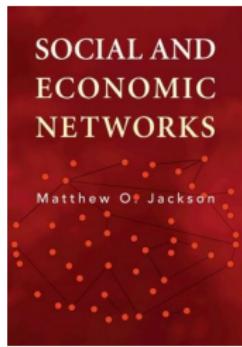
- ▶ What explains different varieties of social, political, and economic outcomes?
<!-->
- ▶ Conventional approaches focus on the traits of individual observations.
- ▶ The outcomes can be a function of individual's ties with others as well as many features of the networks that individuals belong to.
<!-->



Network science as a field

Network science is an interdisciplinary enterprise.

- ▶ Theory: Math, physics, and computer science.
- ▶ Application:
 - Engineering (e.g., power grid and signal exchanges)
 - Natural and life sciences (e.g., disease transmission; yeast Interactome among proteins)
 - Humanities and social sciences (e.g., market transactions and financial exchanges; friendships; mutual trust; ideas and texts)



What is a network?

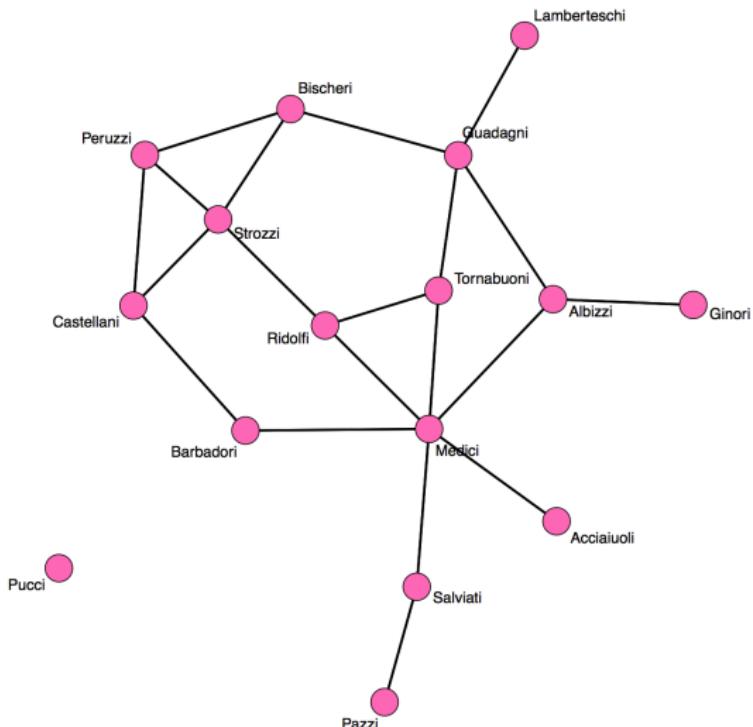
- ▶ Network: Objects + connections.
- ▶ Social network: People, ideas, countries, groups, etc. + relationships.
- ▶ Complex social systems: Many different parts with strong interactions; collective behavior is surprising, hard to predict, namely "emergent."

What is a network?

- ▶ A **network (graph)** is composed of a set of objects, called **nodes** (or **vertices**), with some or all pairs of these objects connected by **ties (edges)**.
- ▶ Two nodes are **neighbors** if they are connected.

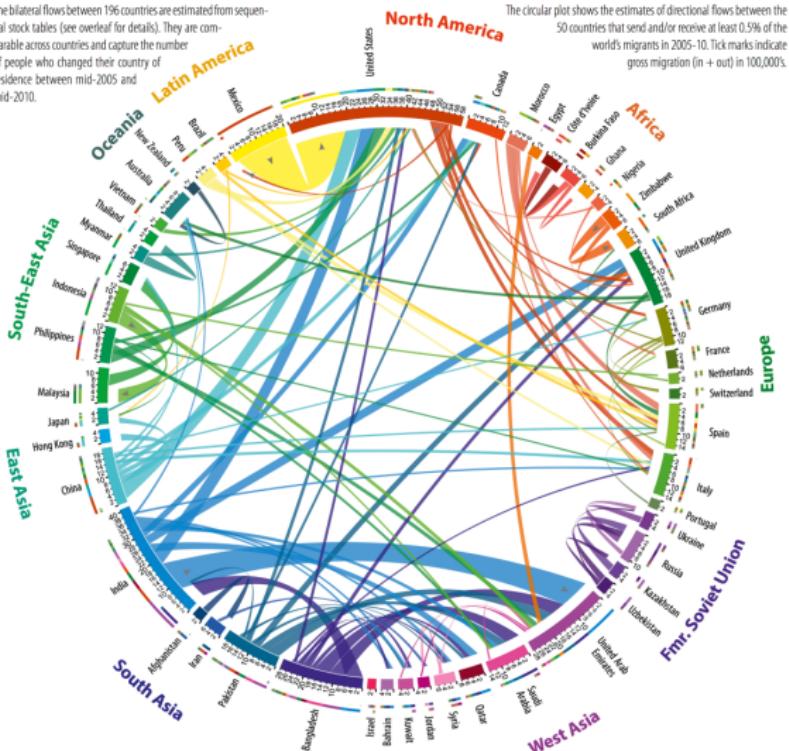
$$\begin{aligned} \text{vertices} &= \{A, B, C, D, E\} \\ \text{edges} &= (\{A, B\}, \{A, C\}, \{B, C\}, \{C, E\}) \end{aligned} \tag{1}$$

Inter-marriages between big families in Renaissance Florentine

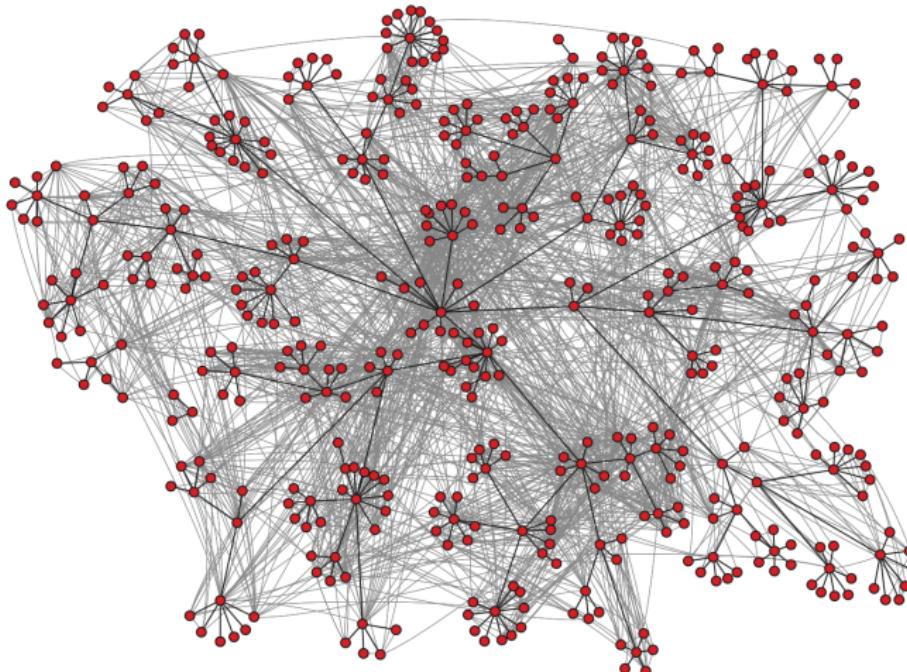


Global migration network

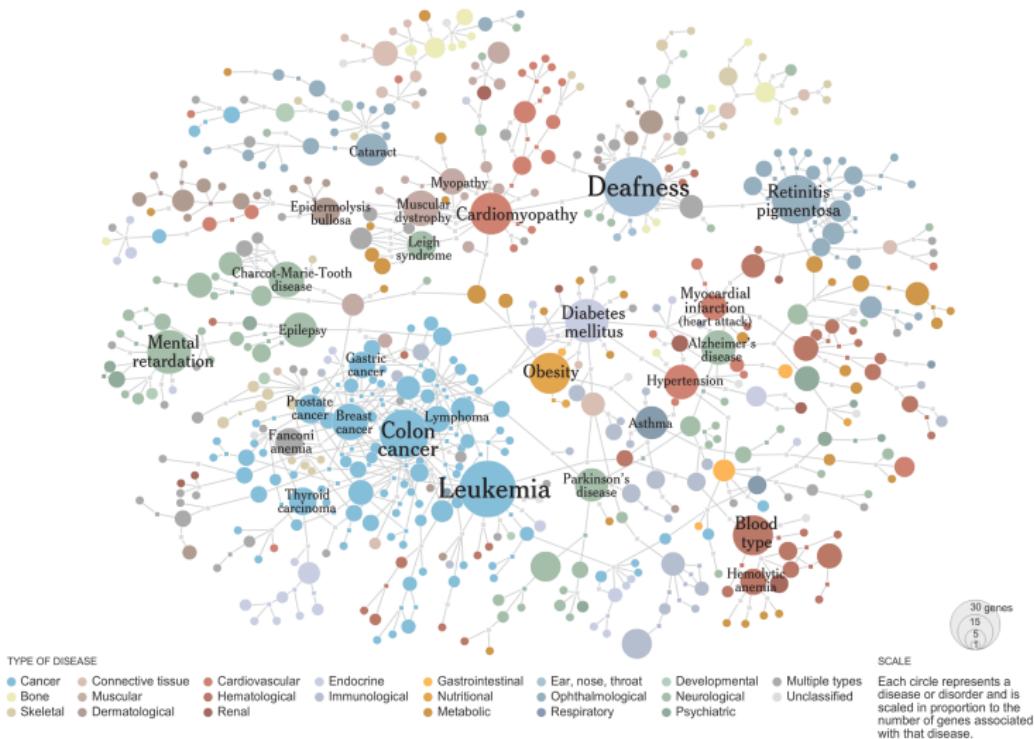
The bilateral flows between 196 countries are estimated from sequential stock tables (see overleaf for details). They are comparable across countries and capture the number of people who changed their country of residence between mid-2005 and mid-2010.



Email communication network in HP labs



Network of diseases with common associated genes

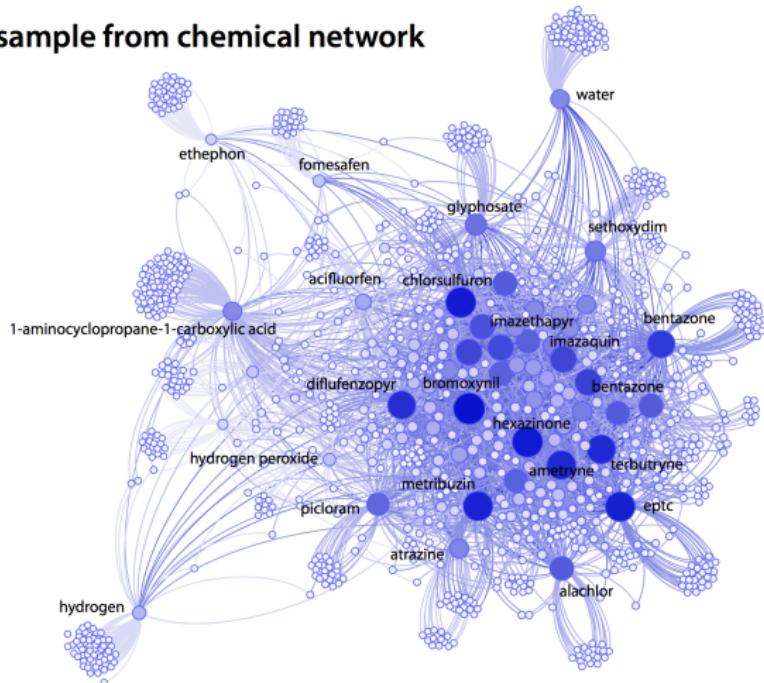


Sources: Marc Vidal; Albert-Laszlo Barabasi; Michael Cusick;
Proceedings of the National Academy of Sciences

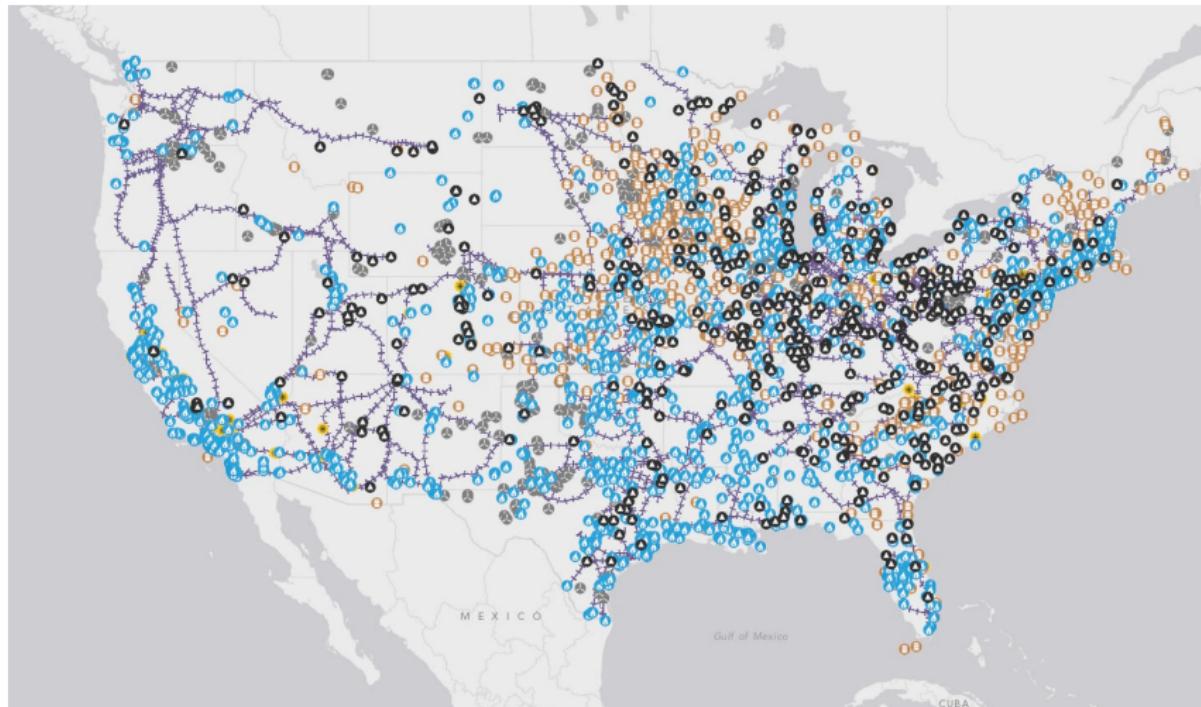
The New York Times

Co-mentioned chemicals in articles and patents

MEDLINE random sample from chemical network



Network of energy supply in the US



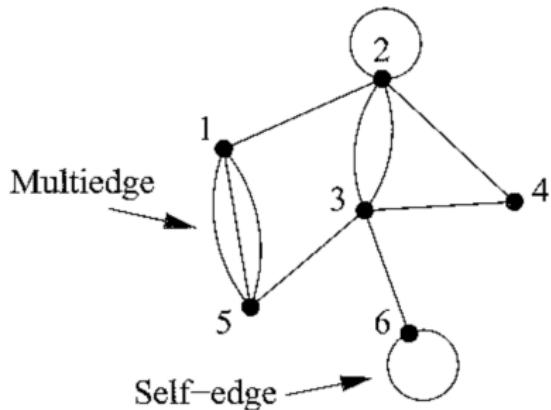
Basic types of networks (graphs)

- ▶ Simple graph
- ▶ Weighted graph
- ▶ Directed graph
- ▶ Signed graph

Basic types of networks: Simple graph

A network that has neither self-edges nor multiedges. That is, in addition to the definition above,

$$A_{ii} = 0 \quad \forall \quad i. \quad (2)$$

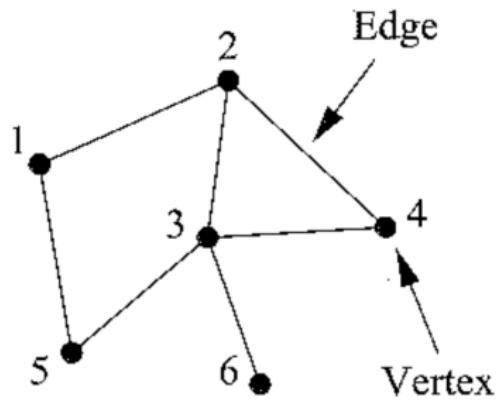


Math representation of networks

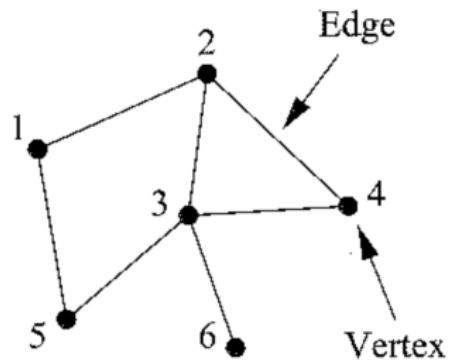
The **adjacency matrix** or **sociomatrix \mathbf{A}** of a graph (network) can be represented as an $n \times n$ matrix \mathbf{A} such that

$$A_{ij} = \begin{cases} 1 & \text{if there exists an edge between } i \text{ and } j \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

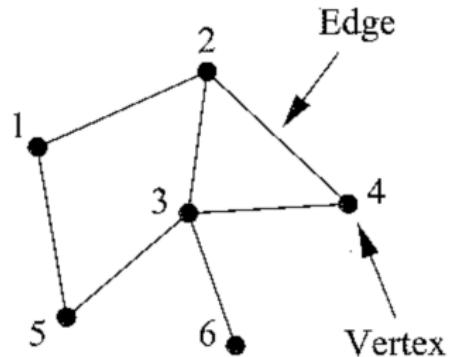
where n refers to the number of **nodes**.



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (4)$$



$$A = (\{1, 2\}, \{1, 5\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{3, 5\}, \{3, 6\}). \quad (5)$$



$$A = \begin{cases} 1 : \{2, 5\} \\ 2 : \{1, 3, 4\} \\ 3 : \{2, 4, 5, 6\} \\ 4 : \{2, 3\} \\ 5 : \{1, 3\} \\ 6 : \{3\} \end{cases} \quad (6)$$

Basic types of networks: Weighted graph

The adjacency matrix or sociomatrix \mathbf{A} of a weighted network is an $n \times n$ matrix with elements

$$A_{ij} = \begin{cases} r \in \mathbb{R} & \text{if there exists an edge between } i \text{ and } j \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

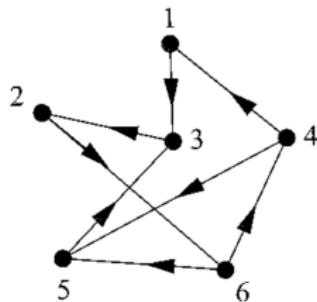
where n refers to the number of **nodes**.

Basic types of networks: Directed graph

A directed network is a network in which edges have a direction, pointing from one vertex to another. The adjacency matrix \mathbf{A} of a *directed* graph is the matrix with elements

$$A_{ij} = \begin{cases} 1 & \text{if there exists an edge from } j \text{ and } i \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

where n refers to the number of **nodes**.



$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad (9)$$

in which columns and rows refer to the starting and ending nodes respectively.

	Weighted	Binary
Directed		
Undirected		

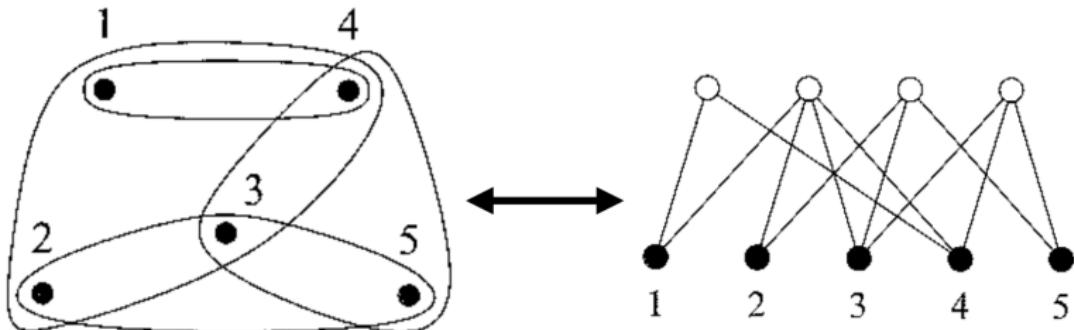
	Weighted	Binary
Directed		
Undirected		SIMPLE

More types of networks

- ▶ Affiliation network (bipartite network and hypergraph);
 - cocitation network (i.e., studies that are cited together in a study)
 - bibliographic coupling network (i.e., studies that use the same references)
- ▶ regular network (i.e., networks in which all nodes have the same degree);
- ▶ planar network (i.e., networks that can be visualized without any edges crossing each other);
- ▶ tree and forest;
- ▶ ... and many more.

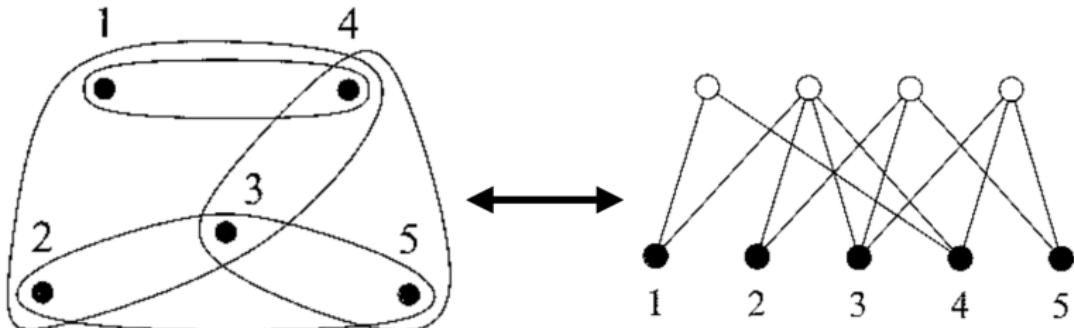
Affiliation network

- ▶ A **bipartite graph** is a graph in which there are two types of node, and edges only run between the two types.
- ▶ A **hypergraph** is a network in which edges can join two or more nodes.



$$B = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad (10)$$

where rows correspond to groups and columns refer to actors.



One-mode projection: From bipartite graph to simple graph.

$$P_{ij} = \sum_{k=1}^g B_{ki} B_{kj} = \sum_{k=1}^g B_{ik}^T B_{kj}, \quad (11)$$

where k refers to affiliated groups.

Two pillars of social network analysis (SNA)

- ▶ Descriptive SNA: Study numerical summary measures of networks.
- ▶ Generative or inferential SNA: Study underlying dynamic process of network formation; hypothesis testing; simulation; Ergm.

Descriptive SNA

- ▶ Network connectivity;
 - degree
 - density
 - path and geodesic distance
- ▶ node centrality: Measures of node importance in a network;
- ▶ ... and many more (e.g., cosine similarity, cliques, triads, and assortive mixing)

Descriptive SNA: Degree

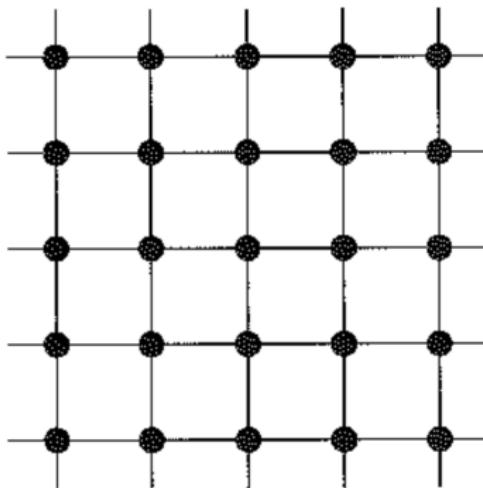
Given an adjacency matrix \mathbf{A} , the degree of a vertex is the number of edges connected to it.

$$k_i = \sum_j^n A_{ij}, \quad (12)$$

where n refers to the number of nodes in the graph.

Descriptive SNA: Degree

A **regular** graph is one in which all nodes have the same degree.



Descriptive SNA: Degree

The total number of edges in graph **A** will be

$$2m = \sum_i^n k_i \quad \Rightarrow \quad m = \frac{1}{2} \sum_i^n k_i, \quad (13)$$

where k_i is the degree of node i in the graph.

Descriptive SNA: Degree

How do we derive the mean degree of graph **A**?

$$c = \frac{1}{n} \sum_i^n k_i = \frac{2m}{n}. \quad (14)$$

Descriptive SNA: Density

- ▶ The density of a graph measures the *connectance* of a graph.
- ▶ Mathematically, it is defined as the number of edges out of the max possible number of edges of a graph.

Descriptive SNA: Density

The density of a graph is defined as the share of the number of edges out of the max possible number of edges of a graph.

$$\rho = \frac{m}{\binom{n}{2}} = \frac{m}{\frac{1}{2}n(n-1)} = \frac{2m}{n(n-1)} = \frac{c}{n-1}, \quad (15)$$

Recall that $c = \frac{1}{n} \sum_i^n k_i = \frac{2m}{n}$, where c is the mean degree.

Descriptive SNA: Density

$$\rho = \frac{m}{\binom{n}{2}} = \frac{m}{\frac{1}{2}n(n-1)} = \frac{2m}{n(n-1)} = \frac{c}{n-1}. \quad (16)$$

- ▶ A network for which the density tends to a constant as the number of nodes tends to infinity is **dense**.
- ▶ A network for which the density tends to 0 as the number of nodes tends to infinity is **sparse**.

Descriptive SNA: Path, walk, and distance

- ▶ A **path** is a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge.
- ▶ A **simple path** is a sequence of nodes with the property that each consecutive pair in the sequence is connected by an edge and each node in the sequence appears only once.

Descriptive SNA: Path, walk, and distance

- ▶ In statistics, a path is also a **walk**.
- ▶ A **walk** of length k between nodes i and j means that the path between i and j contains k edges where $i \neq j$.

Descriptive SNA: Path, walk, and distance

- ▶ Existence of *walks* between nodes tell us about network connectivity.
- ▶ Existence of *walks of minimal length* tells us about **geodesics**, or geodesic distance, between two nodes in the graph.

Descriptive SNA: Path, walk, and distance

The walks of all lengths between a pair of nodes can be counted using matrix multiplication. Define $W = A^k$, where k means we multiply the adjacency matrix of graph \mathbf{A} for k times, then

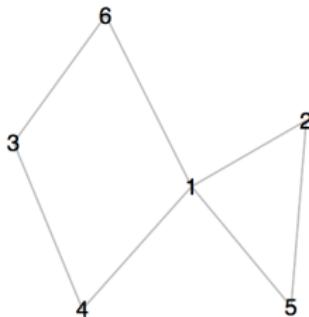
$$W_{ij} = \text{the number of walks of length } k \text{ between } i \text{ and } j. \quad (17)$$

Descriptive SNA: Path, walk, and distance

Altogether, we can define d_{ij} , the geodesic distance between i and j , as follows.
Recall a path is a walk.

$$\begin{aligned} d_{ij} &= \text{length of the shortest path between } i \text{ and } j \\ &= \text{length of the shortest walk between } i \text{ and } j \\ &= \text{the first or min } k \text{ for which } A_{ij}^k > 0 \end{aligned} \tag{19}$$

Descriptive SNA: Path, walk, and distance



Y %*% Y

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
## [1,]	4	1	2	0	1	0
## [2,]	1	2	0	1	1	1
## [3,]	2	0	2	0	0	0
## [4,]	0	1	0	2	1	2
## [5,]	1	1	0	1	2	1
## [6,]	0	1	0	2	1	2

Descriptive SNA: Centrality

Centrality measures the *importance* of each node in a network.

Descriptive SNA: Centrality

Degree centrality uses the degree of a vertex (i.e., the number of neighbors it has) to measure its importance in a graph.

$$k_i = \sum_j^n A_{ij}, \quad (20)$$

where n refers to the number of nodes in the graph.

What is the caveat?

Descriptive SNA: Centrality

Degree centrality uses the degree of a vertex (i.e., the number of neighbors it has) to measure its importance in a graph.

$$k_i = \sum_j^n A_{ij}, \quad (20)$$

where n refers to the number of nodes in the graph.

What is the caveat? Degree centrality treats all neighbors as “equal.”

Descriptive SNA: Centrality

Eigenvector centrality measures the influence of a node in a network by giving each vertex a score **proportional** to the sum of the scores of its neighbors.

$$x_i = \frac{1}{\lambda} \sum_j A_{ij} x_j \quad (21)$$
$$\Rightarrow \mathbf{Ax} = \lambda \mathbf{x},$$

where λ is the leading eigenvector of \mathbf{A} .

Why do we need eigenvector centrality?

Descriptive SNA: Centrality

Closeness centrality measures the importance of a node by calculating the sum of the length of the shortest paths between the node and all other nodes in the graph.

$$c_i = \frac{1}{\frac{1}{n} \sum_j d_{ij}} = \frac{n}{\sum_j d_{ij}}, \quad (22)$$

where d_{ij} is the geodesic distance between i and j .

Use n or $n - 1$? Should we consider the situation such that $i = j$?

Descriptive SNA: Centrality

Closeness centrality measures the importance of a node by calculating the sum of the length of the shortest paths between the node and all other nodes in the graph.

$$c_i = \frac{1}{\frac{1}{n} \sum_j d_{ij}} = \frac{n}{\sum_j d_{ij}}, \quad (22)$$

where d_{ij} is the geodesic distance between i and j .

Use n or $n - 1$? Should we consider the situation such that $i = j$? Maybe not. A vertex's influence on itself is usually not our concern.

Descriptive SNA: Centrality

Betweenness centrality measures the importance of a node by considering how often a node sits on the paths between all other nodes in the graph.

$$x_i = \sum \frac{\sigma_{ijk}}{\sigma_{jk}}, \quad (23)$$

where $i \neq j \neq k$:

- ▶ σ_{jk} refers to the number of shortest paths from j to k .
- ▶ σ_{ijk} refers to the number of those paths that pass through i .

Normalization?

Descriptive SNA: Centrality

Betweenness centrality measures the importance of a node by considering how often a node sits on the paths between all other nodes in the graph.

$$x_i = \sum \frac{\sigma_{ijk}}{\sigma_{jk}}, \quad (23)$$

where $i \neq j \neq k$:

- ▶ σ_{jk} refers to the number of shortest paths from j to k .
- ▶ σ_{ijk} refers to the number of those paths that pass through i .

Normalization? One choice is to normalize the path count by dividing by the total number of vertex pairs (i.e., n^2).

Descriptive SNA: Centrality

A node in a graph can be important if

- ▶ it connects to many other nodes in the graph;
- ▶ it connects to other nodes that connects to many other nodes in the graph;
- ▶ it is on the paths between many other nodes in the graph;
- ▶ it is on average close to many other nodes in the graph;
- ▶ ... and many more.

Descriptive SNA: Centrality

- ▶ Katz centrality (to avoid zero centrality).
- ▶ PageRank (designed by Google to avoid inappropriate high centrality).
- ▶ Hubs and authorities (by taking edge directions seriously; what makes a node more important – a vertex has high centrality if those that point to it have high centrality OR a vertex high centrality if it points to others with high centrality?).

Inferential SNA

We hope to specify a stochastic model to

- ▶ understand the underlying dynamic social and interactive processes (e.g., assortative mixing) associated with the observed outcomes;
- ▶ test hypotheses based on different varieties of node and structural attributes;
- ▶ extrapolate and simulate from the specified model.

Inferential SNA

For example: “clustering” typically observed in social nets can be a result of

- ▶ Sociality: highly active persons create clusters.
- ▶ Homophily: assortative mixing by attribute creates clusters.
- ▶ Transitive triad closure: triangles create clusters.

Inferential SNA: Exponential-family random graph models

The model should fit these terms simultaneously, and identify the independent effects of each process on the overall outcome.

- ▶ Network ties are variable and network nodes are fixed;
- ▶ Assumptions about the form of local interactions (i.e., dependence) among tie variables are explicit;
- ▶ Regularities are represented by model parameters and estimated from data.

Inferential SNA: Exponential-family random graph models

Say we see a friendship network in a classroom of size n .

Inferential SNA: Exponential-family random graph models

Say we see a friendship network in a classroom of size n .

- ▶ Focus on structural characteristic of interest: are there more reciprocated ties than would be expected by chance?
- ▶ Posit a stochastic network model that includes a density parameter (ties occur at random) and a reciprocation parameter (tendency for reciprocation to occur).
- ▶ The probability distribution on the set of all possible graphs of size n is constructed such that graphs with a lot of reciprocation are more probable.
- ▶ The MLE will be the parameter value such that the most probable degree of reciprocation is that which occurs in the observed network.
- ▶ Once we have estimated the parameters of the probability distribution, we can draw graphs at random and compare their characteristics with those of the observed network.
- ▶ If the model is good, then sampled graphs will resemble observed network; then we may hypothesize that the modeled structural effects could explain the emergence of the network.

Inferential SNA: Exponential-family random graph models

- ▶ Step 1: Treat each network as a random variable.
- ▶ Step 2: Propose hypothesis (or hypotheses) that define contingencies among tie variables.
- ▶ Step 3: Identify the specific form to the model based on your hypothesis (or hypotheses).
- ▶ Step 4: Simplify parameters for interpretations.
- ▶ Step 5: Estimate and interpret parameters.
- ▶ Step 6 (additional): Check goodness-of-fit and robustness.

Inferential SNA: Homogeneous Bernoulli (Erdos-Renyi) models

Suppose Y_{ij} in a graph are independent $\forall i, j$.

$$\text{logit}[P(Y_{ij} = 1 | X = x, \beta)] = \sum_k \beta_k X_{k,ij}, \quad (24)$$

given some covariates $X = \{X_1, \dots, X_k\}$.

Inferential SNA: Homogeneous Bernoulli (Erdos-Renyi) models

The log of the likelihood is

$$\ell(\beta | Y, x) \equiv \log[P(Y = y | X = x, \beta)], \quad (25)$$

where $\beta \in \mathbb{R}^k$.

Inferential SNA: Homogeneous Bernoulli (Erdos-Renyi) models

Say Y_{ij} are independent and equally likely.

$$P(Y_{ij} = 1 | X = x, \beta) = \frac{\exp(\beta)}{1 + \exp(\beta)} \quad \forall i, j. \quad (26)$$

Equivalently,

$$\text{log odds}(Y_{ij} = 1 | X = x, \beta) = \beta \quad \forall i, j. \quad (27)$$

Inferential SNA: Homogeneous Bernoulli (Erdos-Renyi) models

Say Y_{ij} are independent and equally likely.

$$P(Y_{ij} = 1 | X = x, \beta) = \frac{\exp(\beta)}{1 + \exp(\beta)} \quad \forall i, j. \quad (26)$$

Equivalently,

$$\text{log odds}(Y_{ij} = 1 | X = x, \beta) = \beta \quad \forall i, j. \quad (27)$$

Our goal is to find $\hat{\beta}$ that maximizes the log-likelihood function $\ell(\beta | Y, x)$.

Inferential SNA: French financial elite network (Kadushin 1995, AJS)}

- ▶ Each node is a member of french financial elite ($n = 28$).
- ▶ Each edge represents who-to-whom responses to questions about "who are your friends."
- ▶ The dataset also recorded other node-level attributes.

Inferential SNA: French financial elite network (Kadushin 1995, AJS)}

```
> fit = ergm(ffef ~ edges)
> summary(fit)

=====
Summary of model fit
=====

Formula: ffef ~ edges

Iterations: 5 out of 20

Monte Carlo MLE Results:
  Estimate Std. Error MCMC % p-value
edges -1.5533     0.1355      0 <1e-04 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null Deviance: 524.0 on 378 degrees of freedom
Residual Deviance: 350.1 on 377 degrees of freedom

AIC: 352.1    BIC: 356    (Smaller is better.)
```

Inferential SNA: French financial elite network (Kadushin 1995, AJS)}

\begin{frame} How do we interpret the coefficient? Recall

$$P(Y_{ij} = 1 | X = x, \beta) = \frac{\exp(\beta)}{1 + \exp(\beta)} \quad \forall i, j. \quad (28)$$

Therefore,

$$P(Y_{ij} = 1 | \hat{\beta}) = \frac{\exp(\hat{\beta})}{1 + \exp(\hat{\beta})} = 0.1746032. \quad (29)$$

What does this number mean?

Examples of SNA research

- ▶ State-building (e.g., Acemoglu et al 2015 AER)
- ▶ Regime change (e.g., Linos 2011 AJPS; Naidu et al 2015 WP)
- ▶ Information diffusion (e.g., Larson and Lewis 2017 APSR)
- ▶ Voting and electoral accountability (e.g., Arias et al 2017 WP; Cruz et al 2017 AER)
- ▶ Legislative behaviors and coalition politics (e.g., Bratton and Rouse 2011 LSQ)
- ▶ Collective action and protests (e.g., Siegel 2011 JOP)
- ▶ ... and more

Some practical guide

- ▶ Reflect on your area(s) of substantive interest.
- ▶ Pick a particular structure/phenomenon/pattern of “relations.”
 - What would the nodes represent?
 - What would the edges represent?
 - Should you use an undirected or a directed network?
 - Should you use a weighted or binary network?
 - Should you use an unipartite or bipartite?
- ▶ Describe how it might be represented using networks.
- ▶ Reflect on the advantages and disadvantages of your chosen representation.

Thank you!



ccheng615@tsinghua.edu.cn