Social Network Analysis

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?
- 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

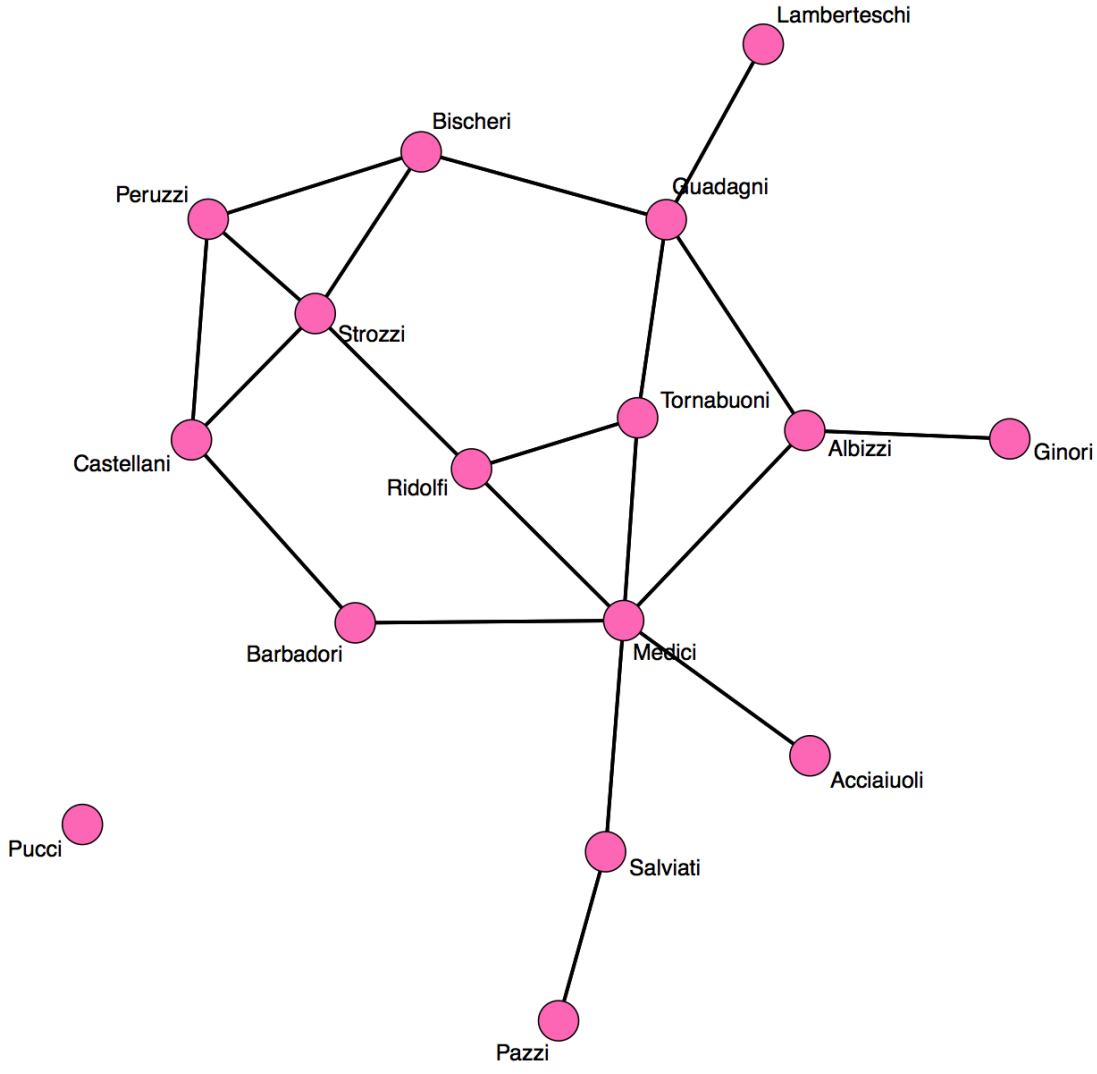
- Theory: Math, physics, and computer science.
- Application:
  - Engineering (e.g., power grid and signal exchanges)
  - Natural and life sciences (e.g., disease transmission; yeast Interactome among proteins)
  - Humanities and social sciences (e.g., market transactions and financial exchanges; friendships;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

What is a network?

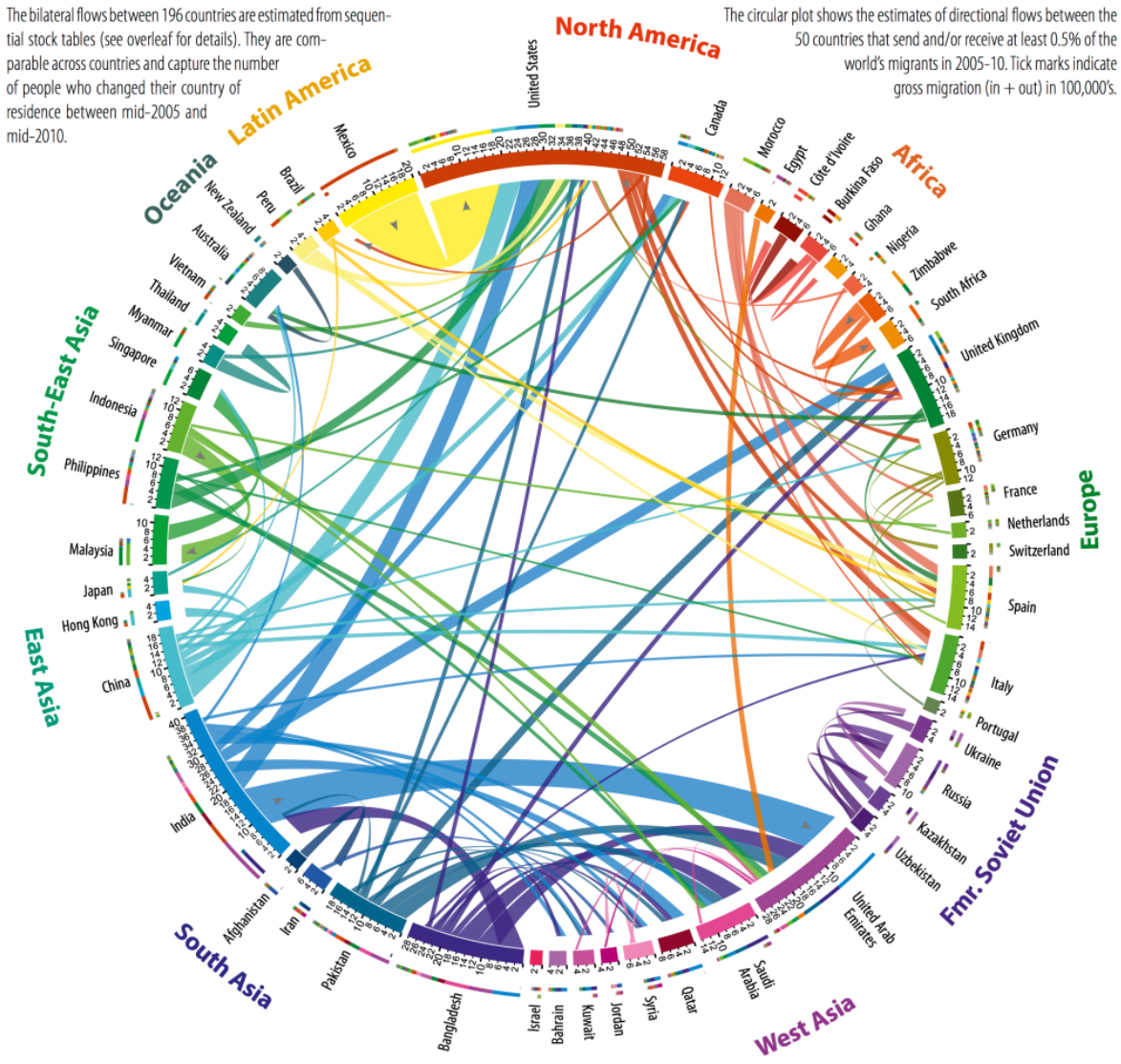
- A **network** (**graph**) is a composed of a set of objects, called **nodes** (or **vertices**), with some or all pairs of these objects connected by **ties** (**edges**).
- Two nodes are **neighbors** if they are connected.

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

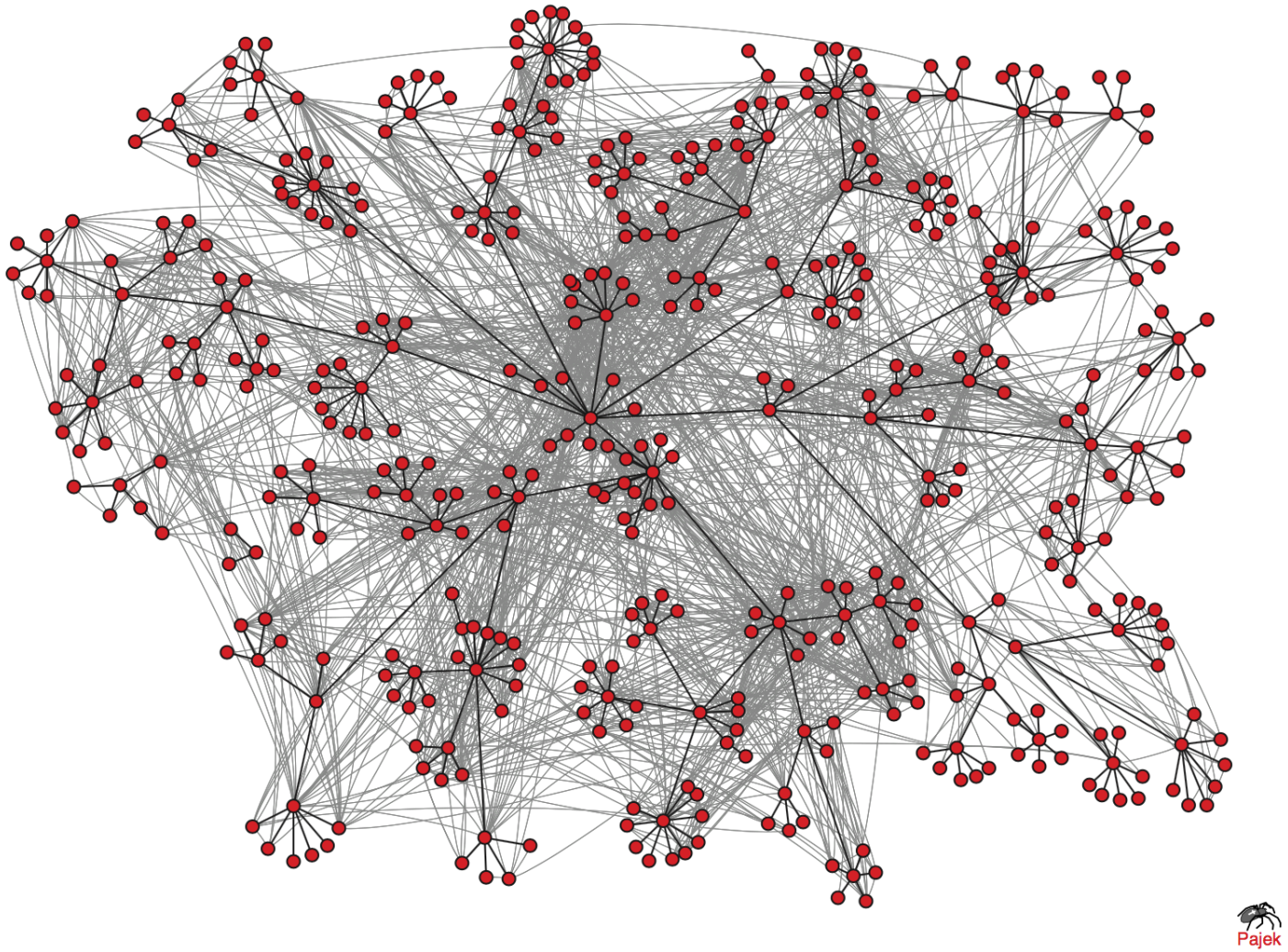
Inter-marriages between big families in Renaissance Florentine



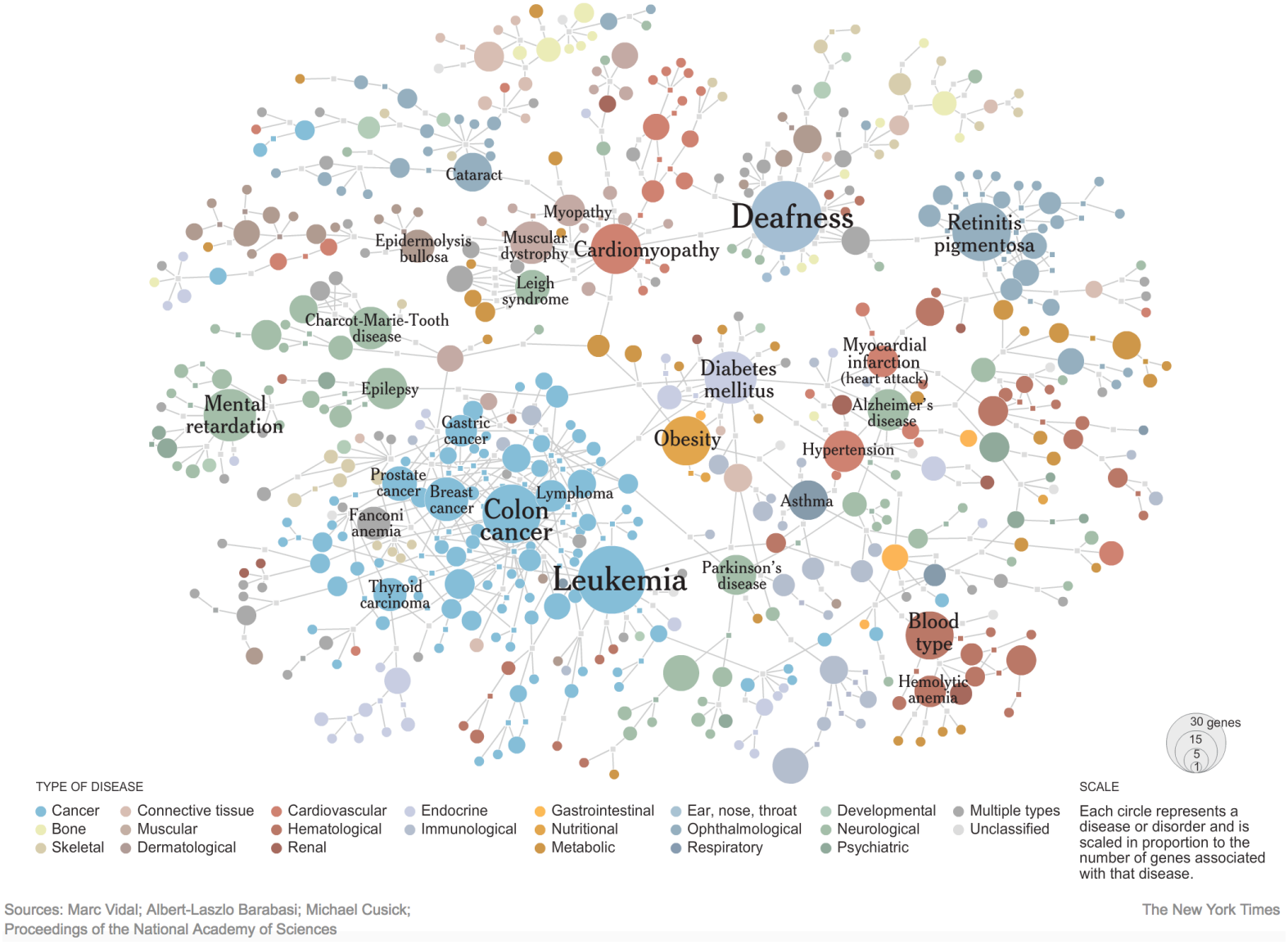
Global migration network



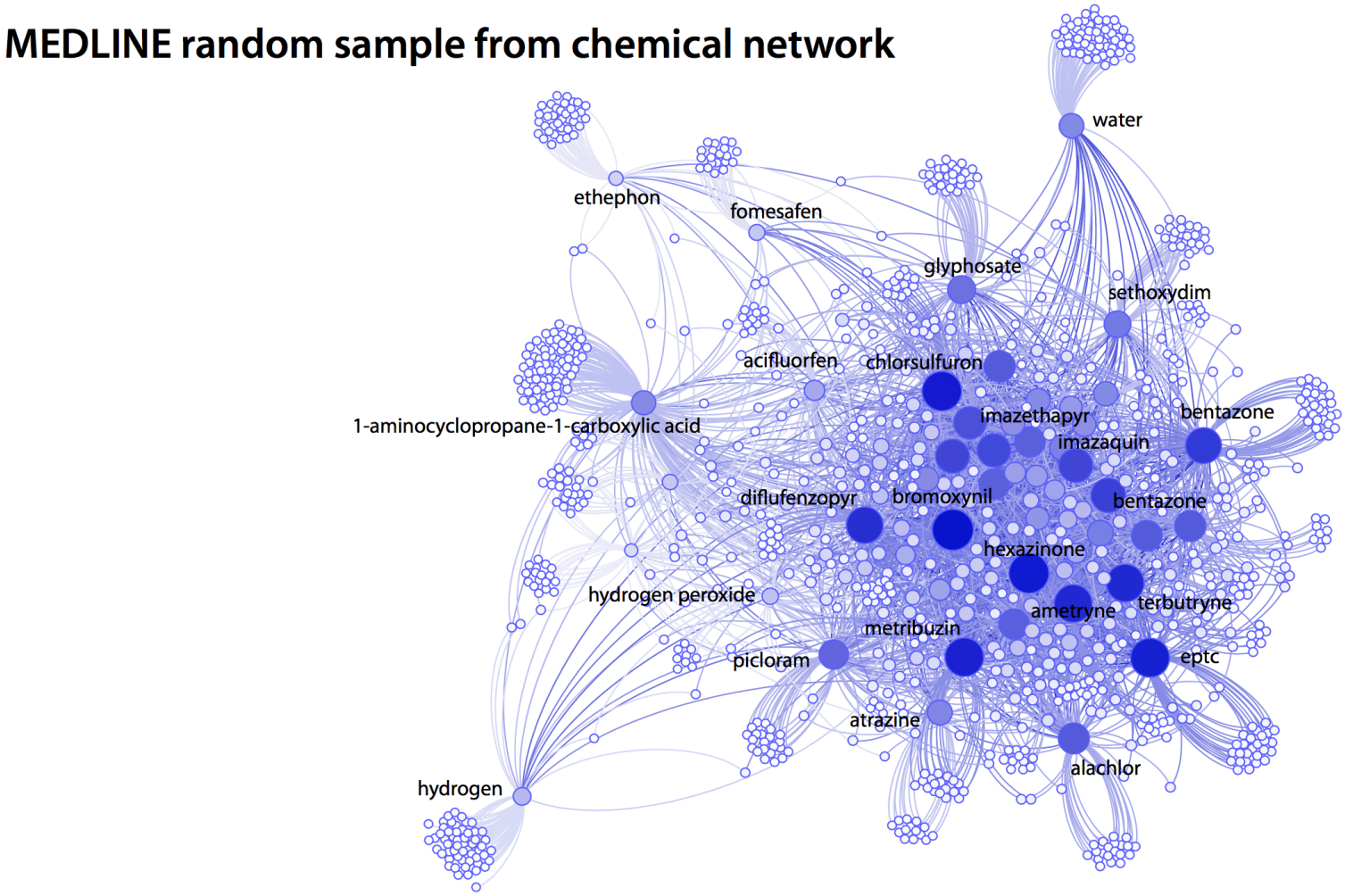
Email communication network in HP labs



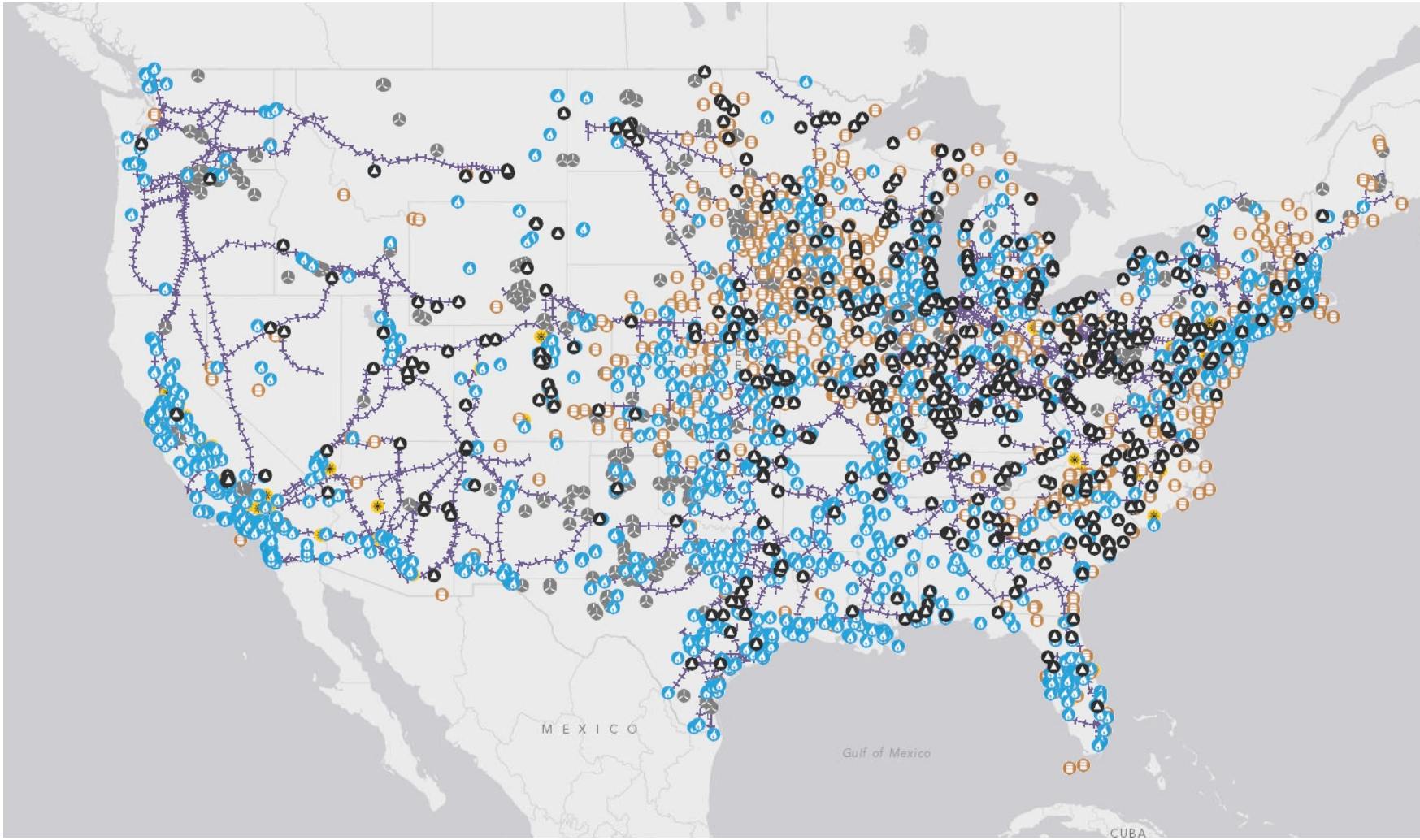
Network of diseases with common associated genes



Co-mentioned chemicals in articles and patents



Network of energy supply in the US



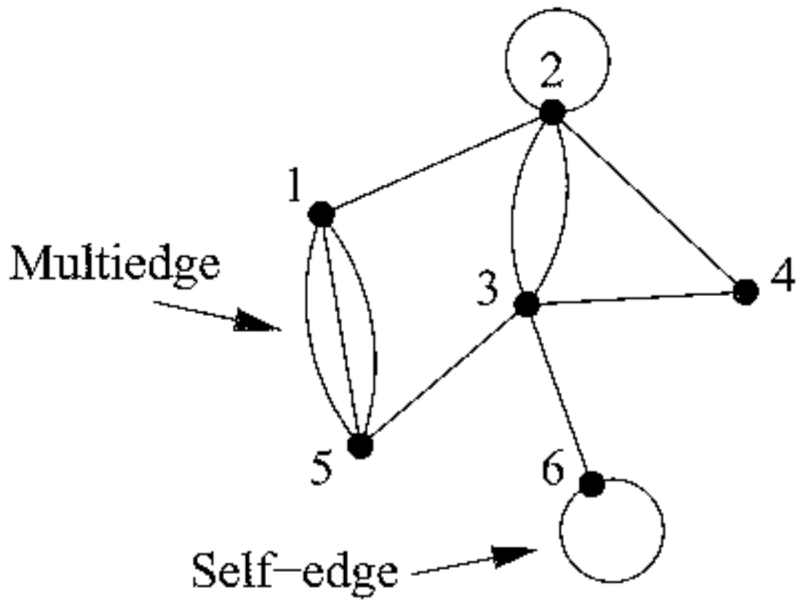
Basic types of networks (graphs)

- 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,

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

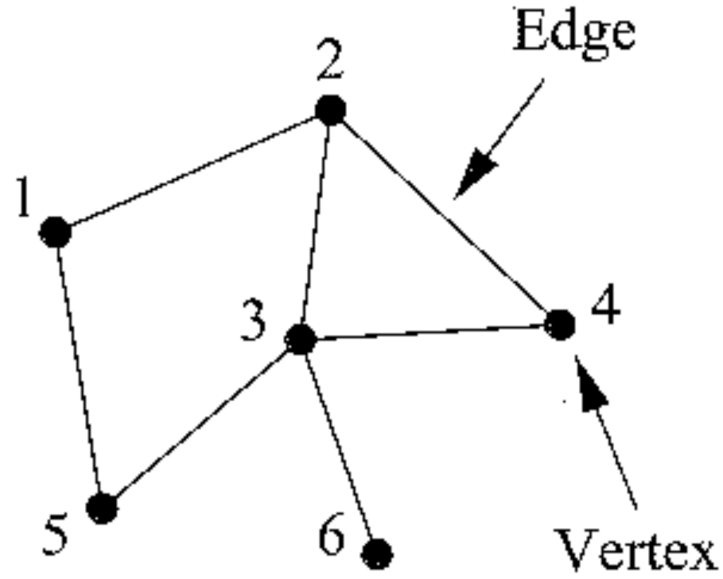


Math representation of networks

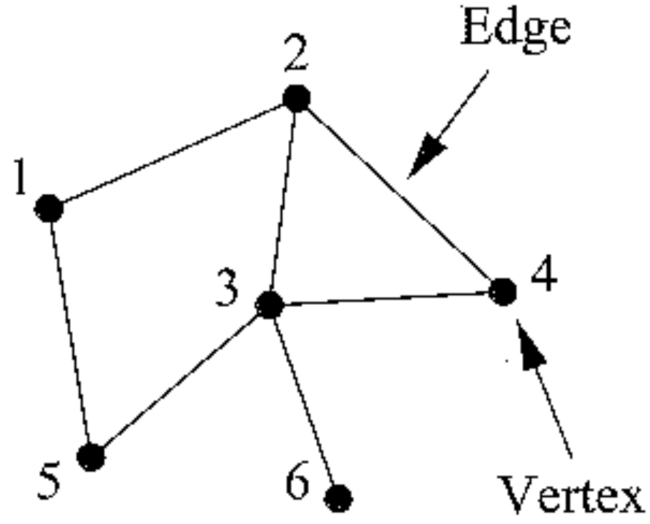
The **adjacency matrix** or **sociomatrix A** of a graph (network) can be represented as an $n \times n$ matrix **A** such that

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

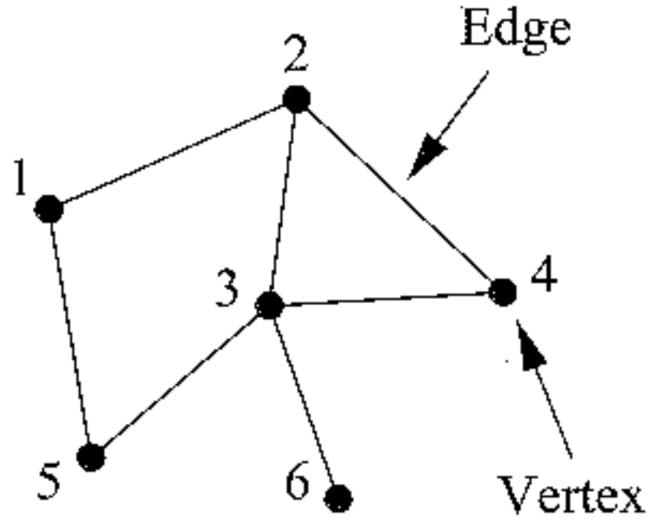
where n refers to the number of **nodes**.



$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{4}$$



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



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

Basic types of networks: Weighted graph

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

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

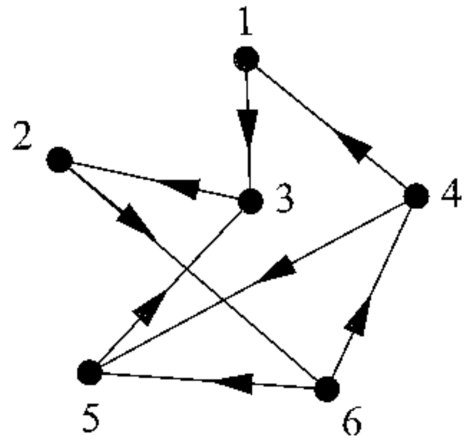
where n refers to the number of **nodes**.

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 **A** of a *directed* graph is the matrix with elements

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

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



$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{9}$$

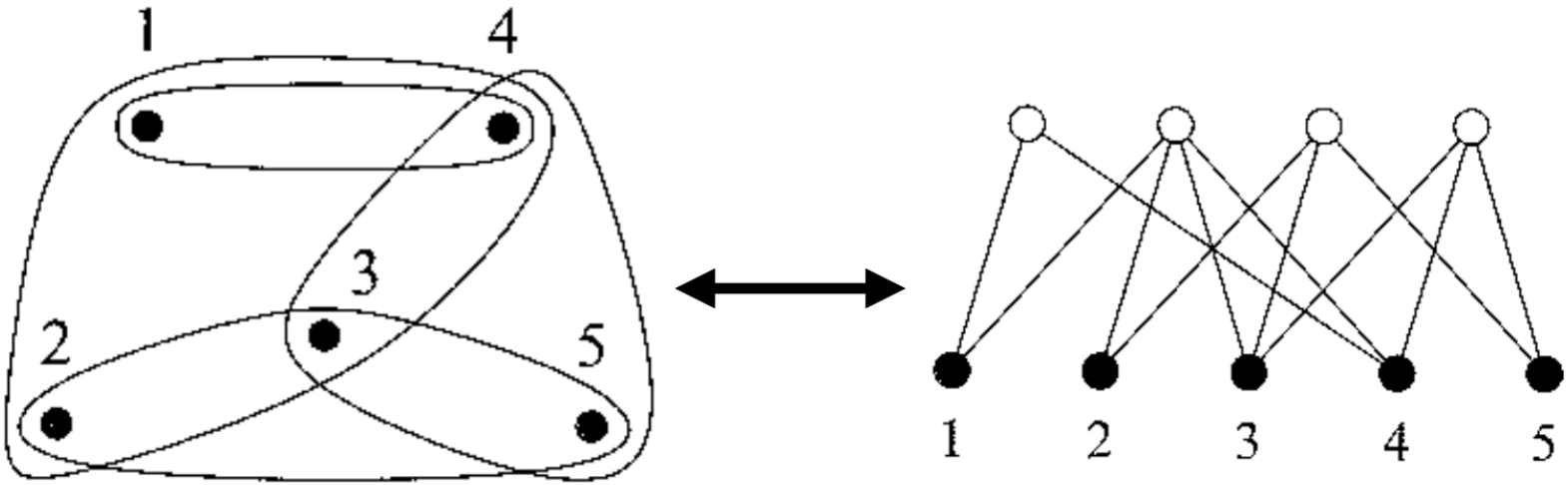
in which columns and rows refer to the starting and ending nodes respectively.

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.

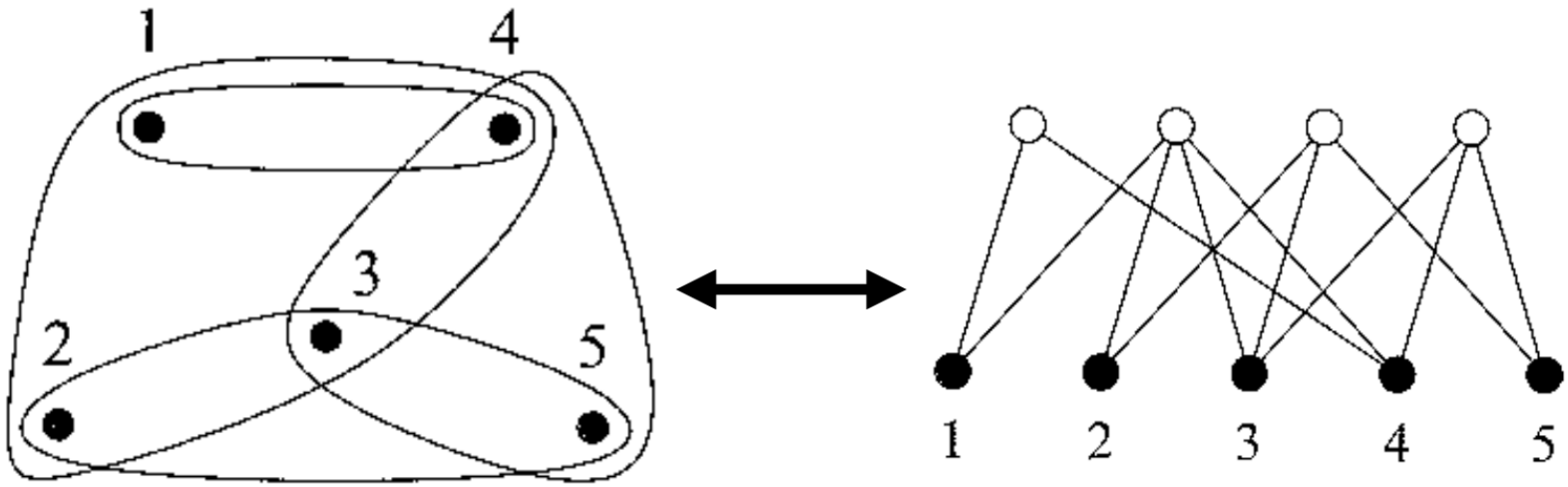
Affiliation network

- 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.



$$B = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \tag{10}$$

where rows correspond to groups and columns refer to actors.



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

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

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

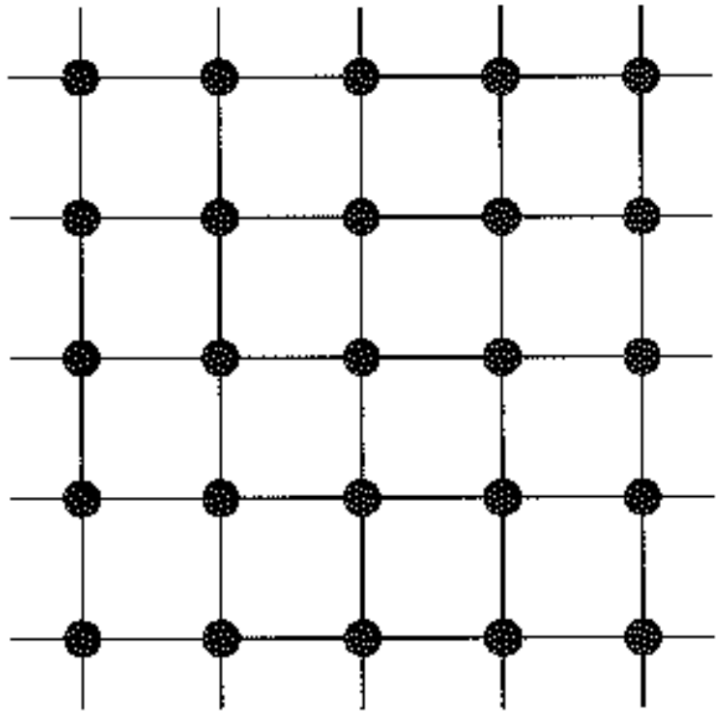
Given an adjacency matrix **A**, the degree of a vertex is the number of edges connected to it.

$$k_i = \sum_{j}^{n} A_{ij}, \tag{12}$$

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, m, in graph $\mathbf{A}$ will be

$$2m = \sum_{i}^{n} k_i \quad \Rightarrow \quad m = \frac{1}{2} \sum_{i}^{n} k_i, \tag{13}$$

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**?

$$c = \frac{1}{n} \sum_{i}^{n} k_i = \frac{2m}{n}. \tag{14}$$

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.

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

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

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

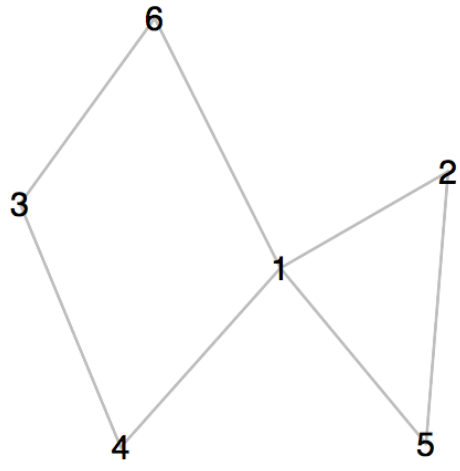
- Existence of *walks* between nodes tell us about network connectivity.
- Existence of *walks of minimal length* tells us about **geodesics**, or geodesic distance, between two nodes in the graph.

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

$$W_{ij} = \text{the number of walks of length } k \text{ between } i \text{ and } j. \tag{16}$$

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: 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.

$$\begin{aligned} d_{ij} &= \text{length of the shortest path between } i \text{ and } j \\ &= \text{length of the shortest walk between } i \text{ and } j \\ &= \text{the first or min } k \text{ for which } A_{ij}^k > 0 \end{aligned} \tag{17}$$

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.

$$k_i = \sum_{j}^{n} A_{ij}, \tag{18}$$

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

What is the caveat of degree centrality?

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.

$$\begin{aligned} x_i &= \frac{1}{\lambda} \sum_j A_{ij} x_j \\ \Rightarrow \quad \mathbf{Ax} &= \lambda \mathbf{x}, \end{aligned} \tag{19}$$

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.

$$c_i = \frac{1}{\frac{1}{n}\sum_j d_{ij}} = \frac{n}{\sum_j d_{ij}}, \tag{20}$$

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

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.

$$x_i = \sum \frac{\sigma_{ijk}}{\sigma_{jk}}, \tag{21}$$

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?

Generative 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.

Generative 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.

Generative SNA: Homogeneous Bernoulli (Erdos-Renyi) models

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

$$\text{logit}[P(Y_{ij} = 1 | X = x, \beta)] = \sum_k \beta_k X_{k,ij}, \tag{22}$$

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

Generative SNA: Homogeneous Bernoulli (Erdos-Renyi) models

The log of the likelihood is

$$\ell(\beta|Y,x) \equiv \log[P(Y=y|X=x,\beta)], \tag{23}$$

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

Generative SNA: Homogeneous Bernoulli (Erdos-Renyi) models

Say Y_{ij} are independent and equally likely.

$$P(Y_{ij} = 1|X = x, \beta) = \frac{\exp(\beta)}{1 + \exp(\beta)} \quad \forall i, j. \tag{24}$$

Equivalently,

$$\log \text{odds}(Y_{ij} = 1|X = x, \beta) = \beta \quad \forall i, j. \tag{25}$$

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

Example: 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.

Example: 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)

How do we interpret the coefficient? Recall

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

Now given that $\hat{\beta} = -1.5533$,

$$P(Y_{ij} = 1|\hat{\beta}) = \frac{\exp(\hat{\beta})}{1 + \exp(\hat{\beta})} = 0.1746032. \tag{27}$$

What does this number mean?

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