Social Network Analysis in Comparative Politics

Chao-Yo Cheng
University of California, Los Angeles

December 2, 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? Read Siegel 2011.

Motivation

- What is a network?
- How is it different from a *social* network? Complex systems?

Motivation

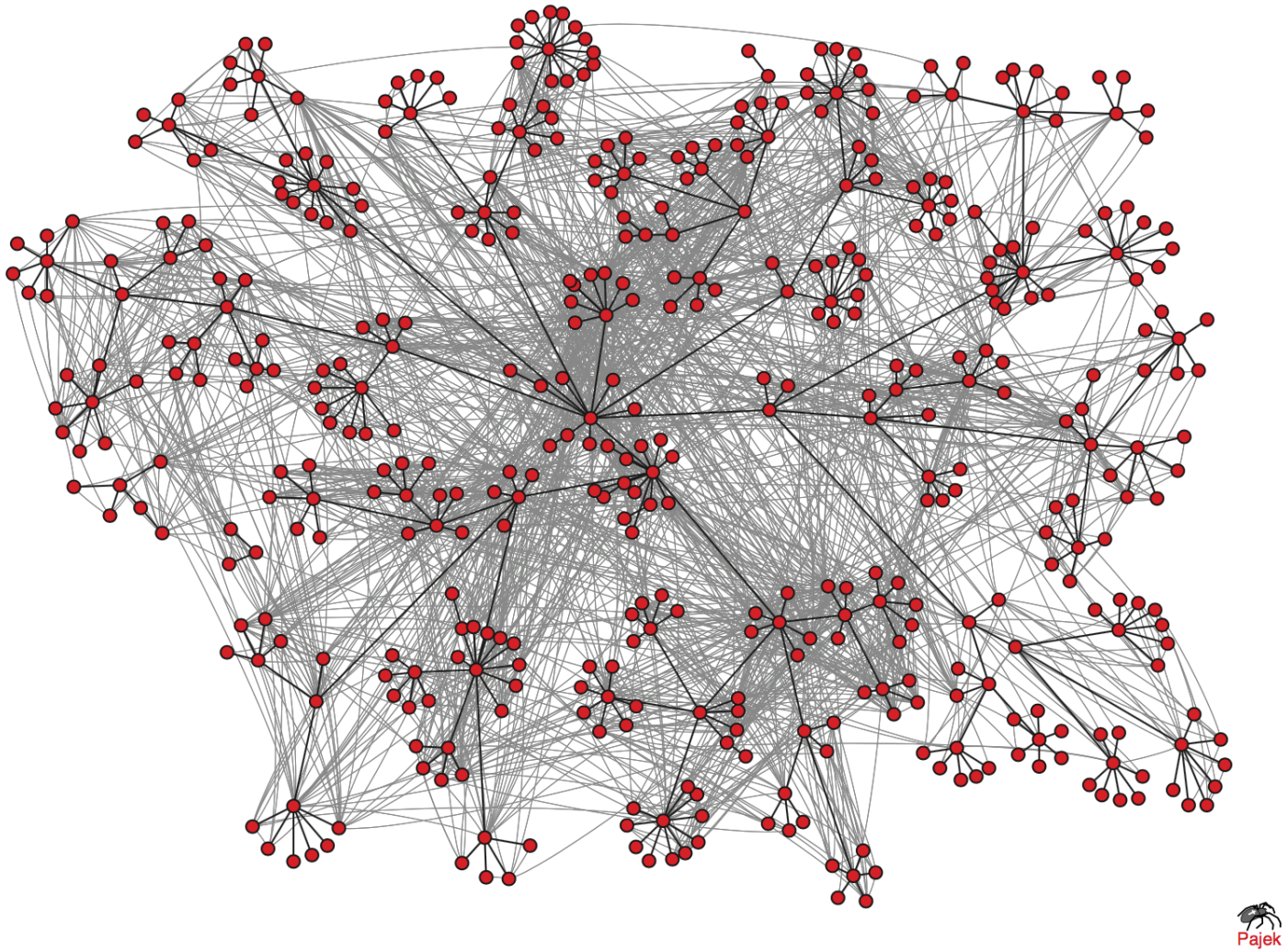
- Network: Things + connections.
- Social network: People + relationships.
- Complex social systems: Many different parts with strong interactions; collective behavior is surprising, hard to predict, namely "emergent."

Motivation

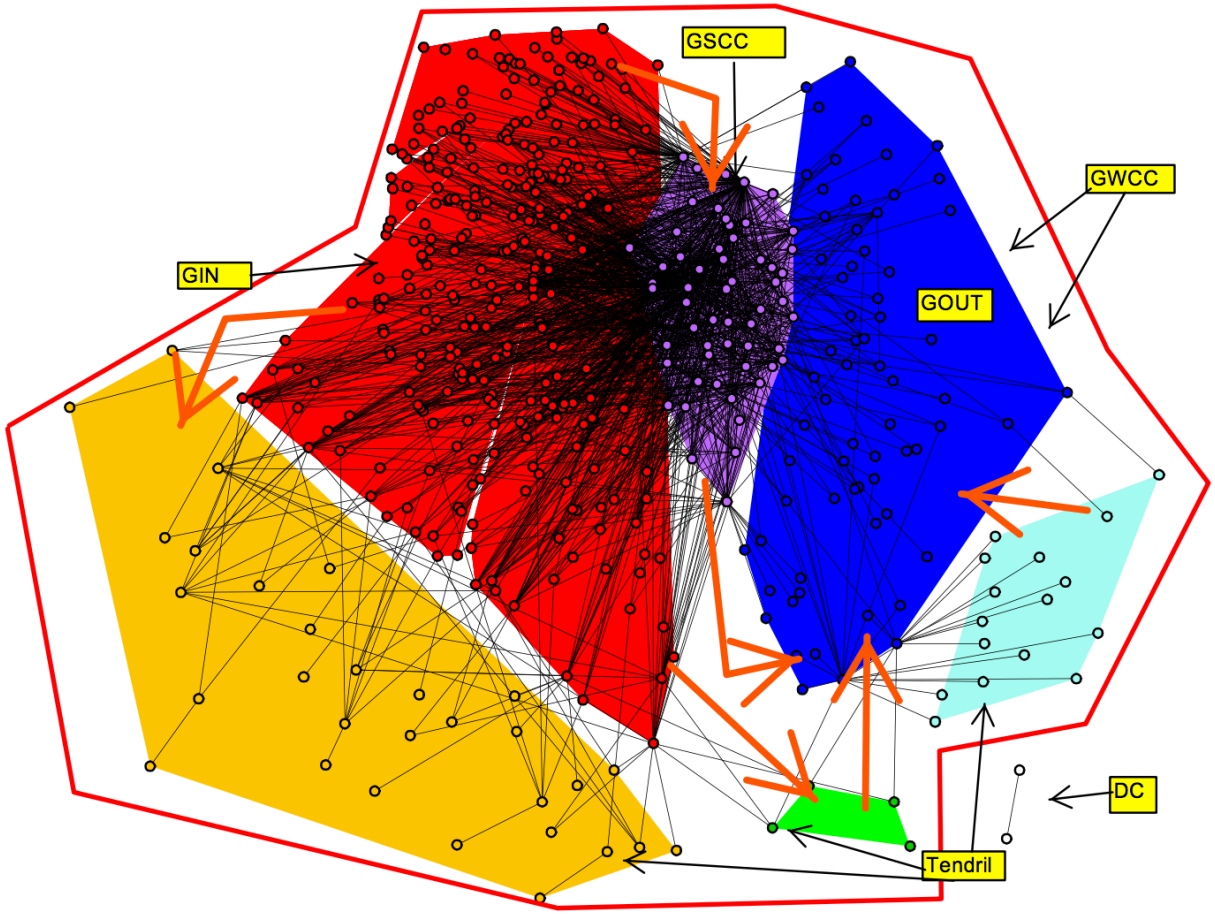
Network analysis is an interdisciplinary enterprise.

- Natural sciences (e.g., math, chemistry, and physics)
- Computer science
- Medical and life sciences
- Engineering
- Humanities and social sciences

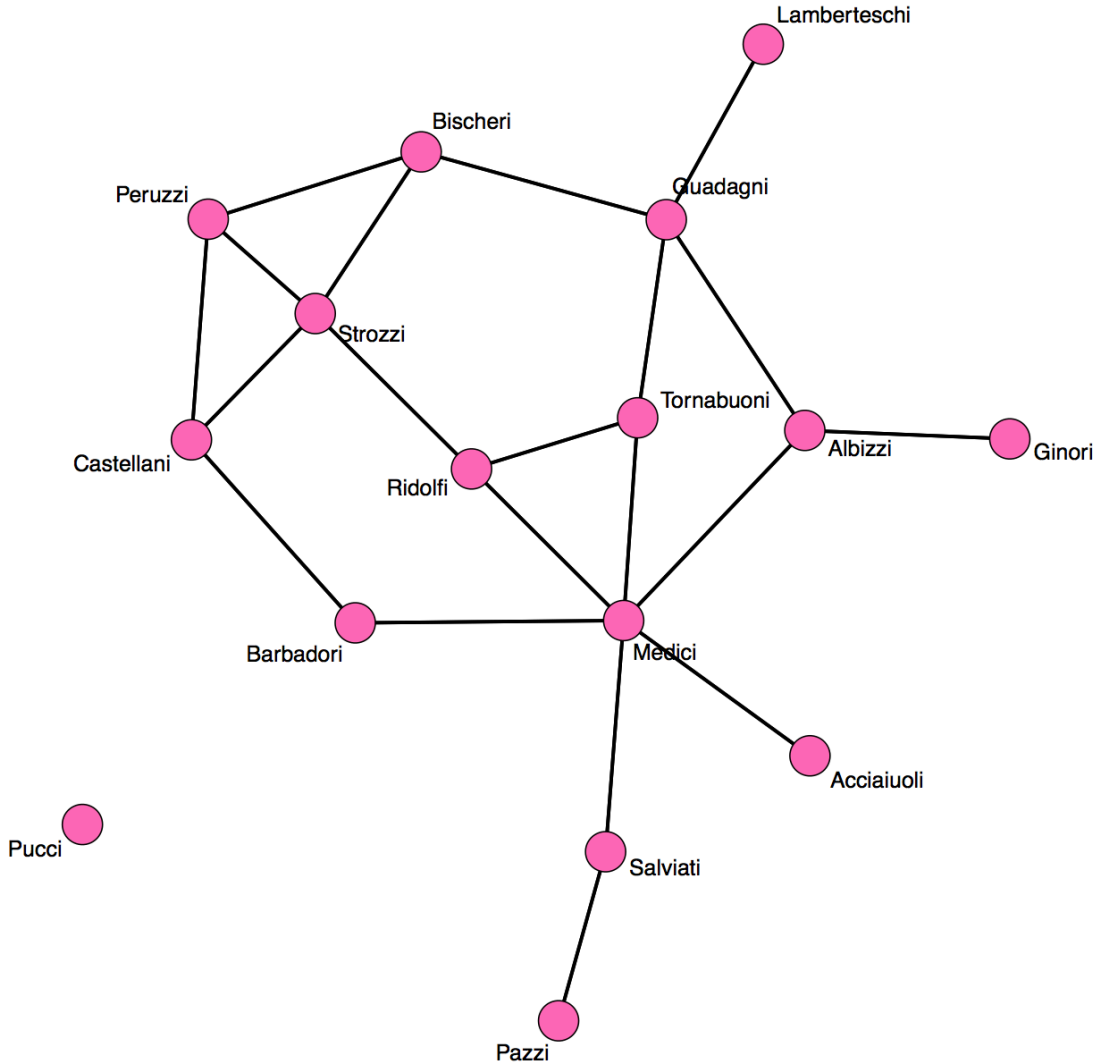
Email communication network in HP Labs



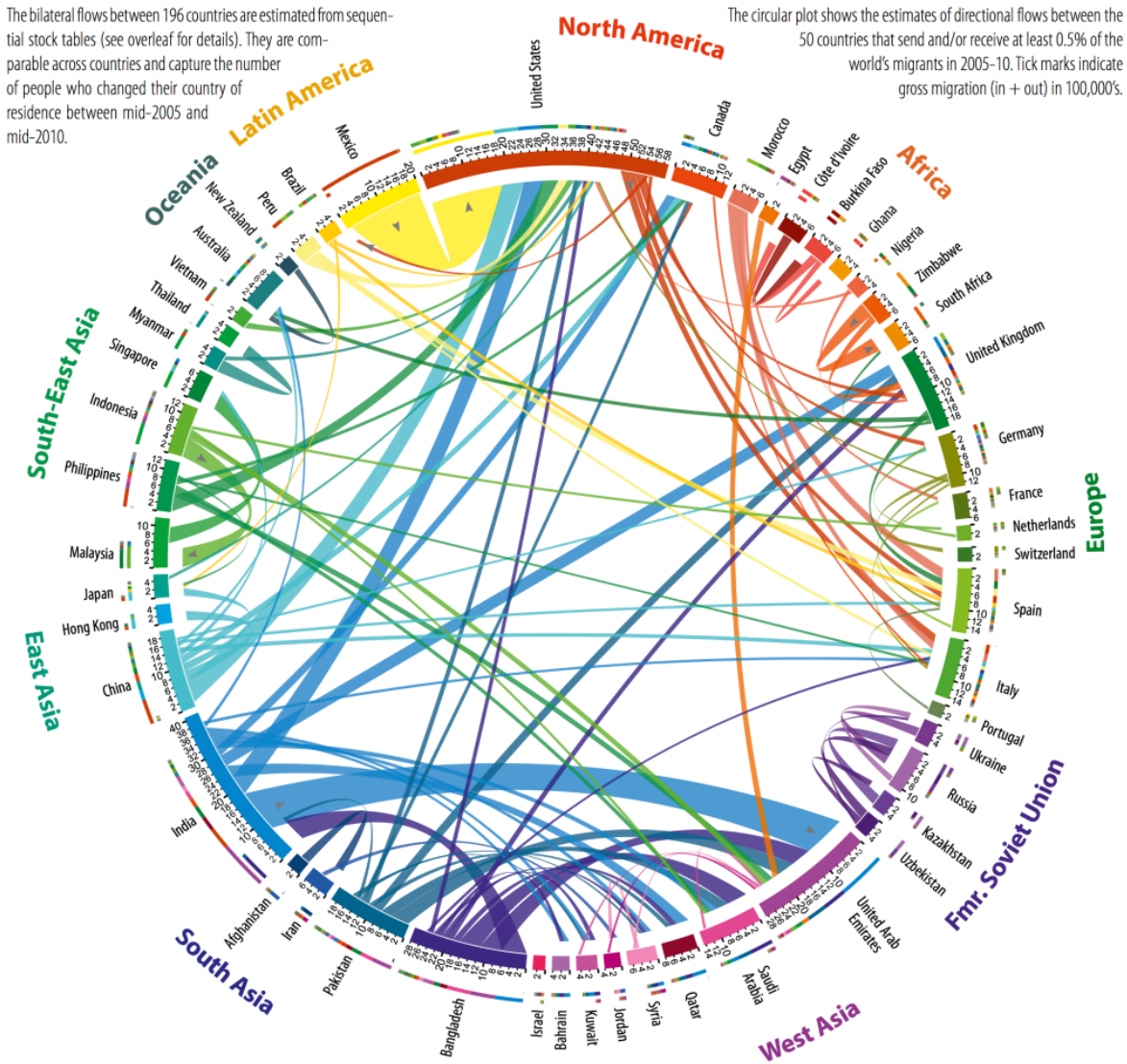
Federal funds network



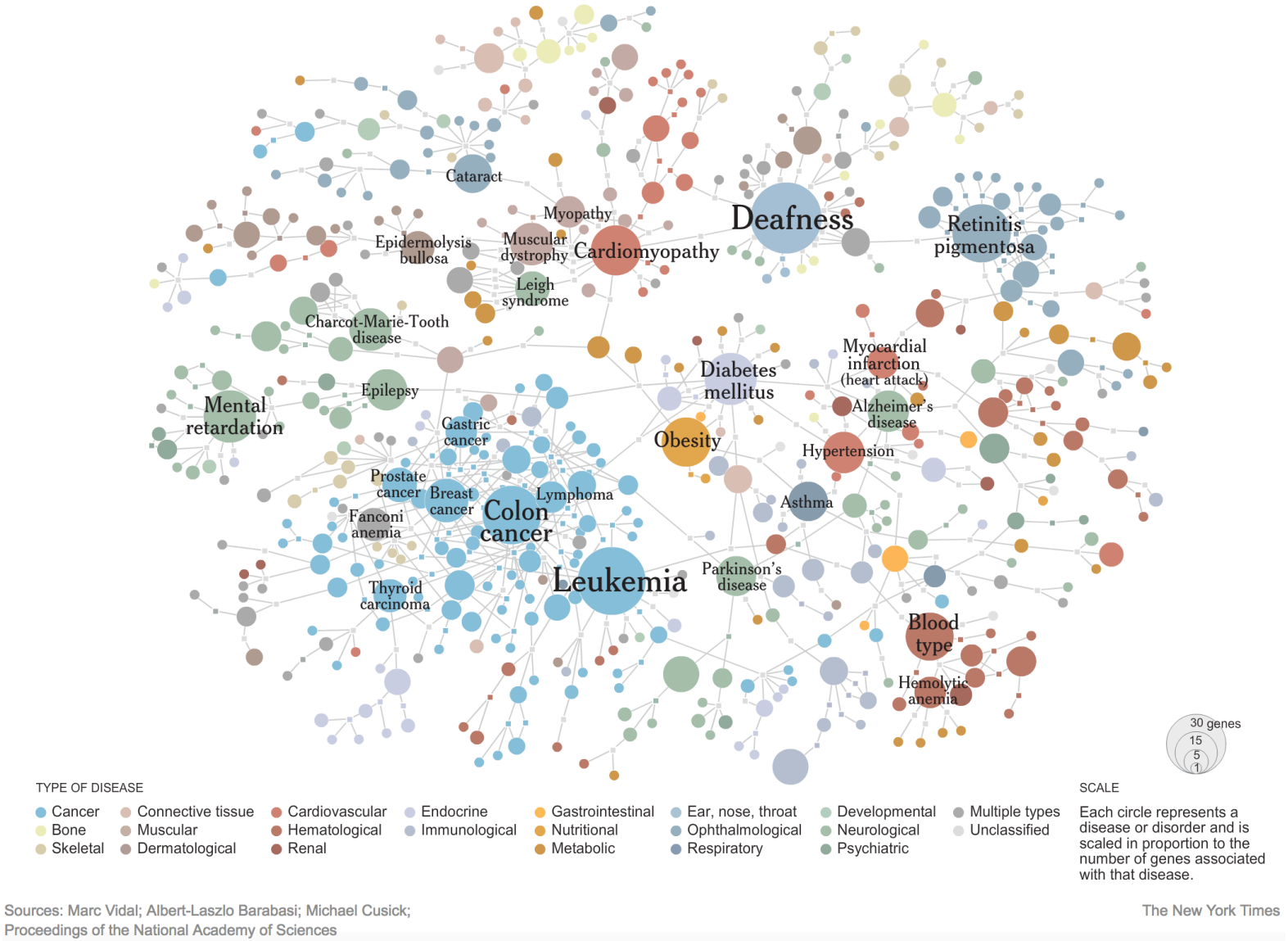
Inter-marriages between big families in Renaissance Florentine



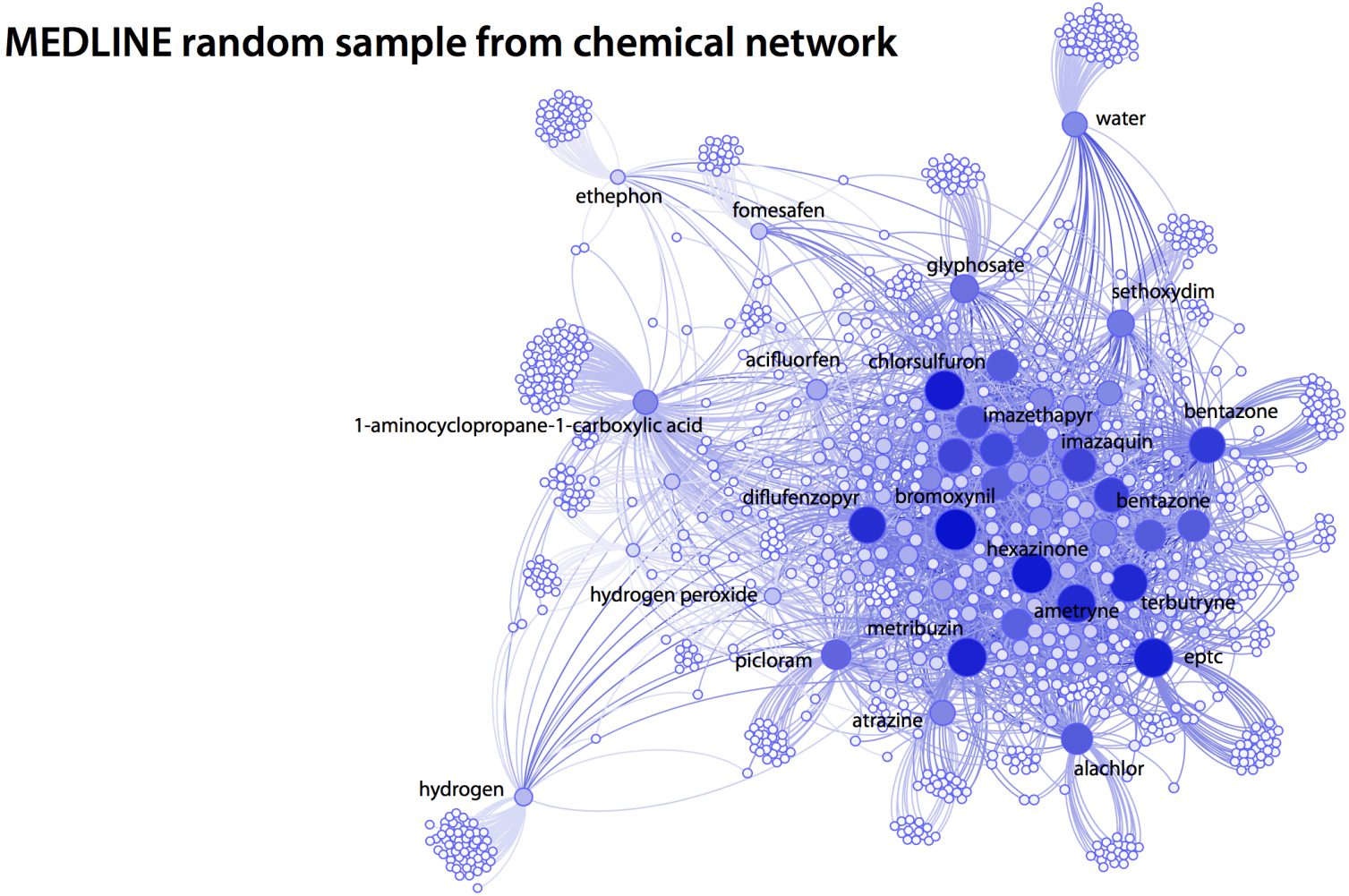
Global migration network



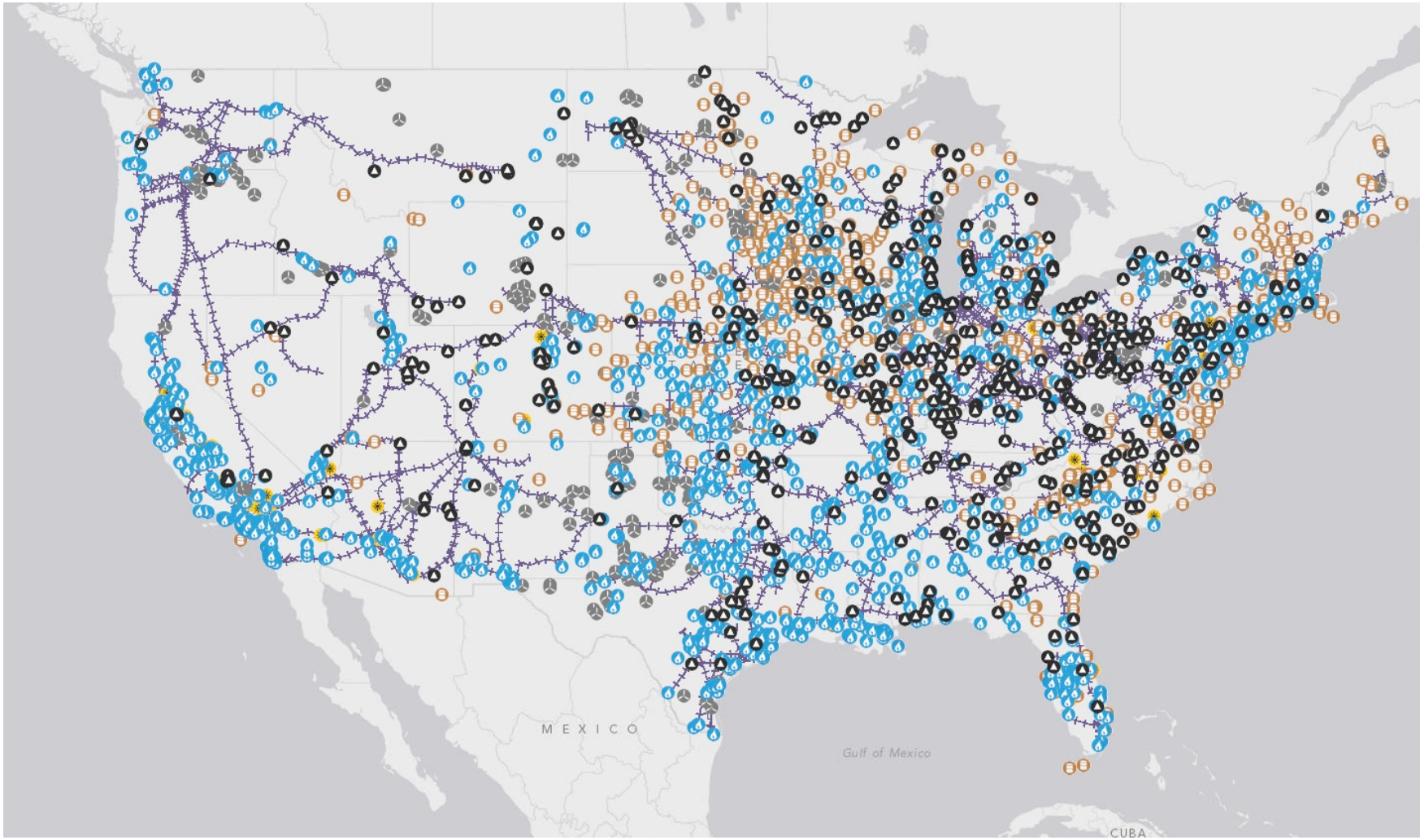
Network of diseases with shared genes



Co-mentioned chemicals in articles and patents

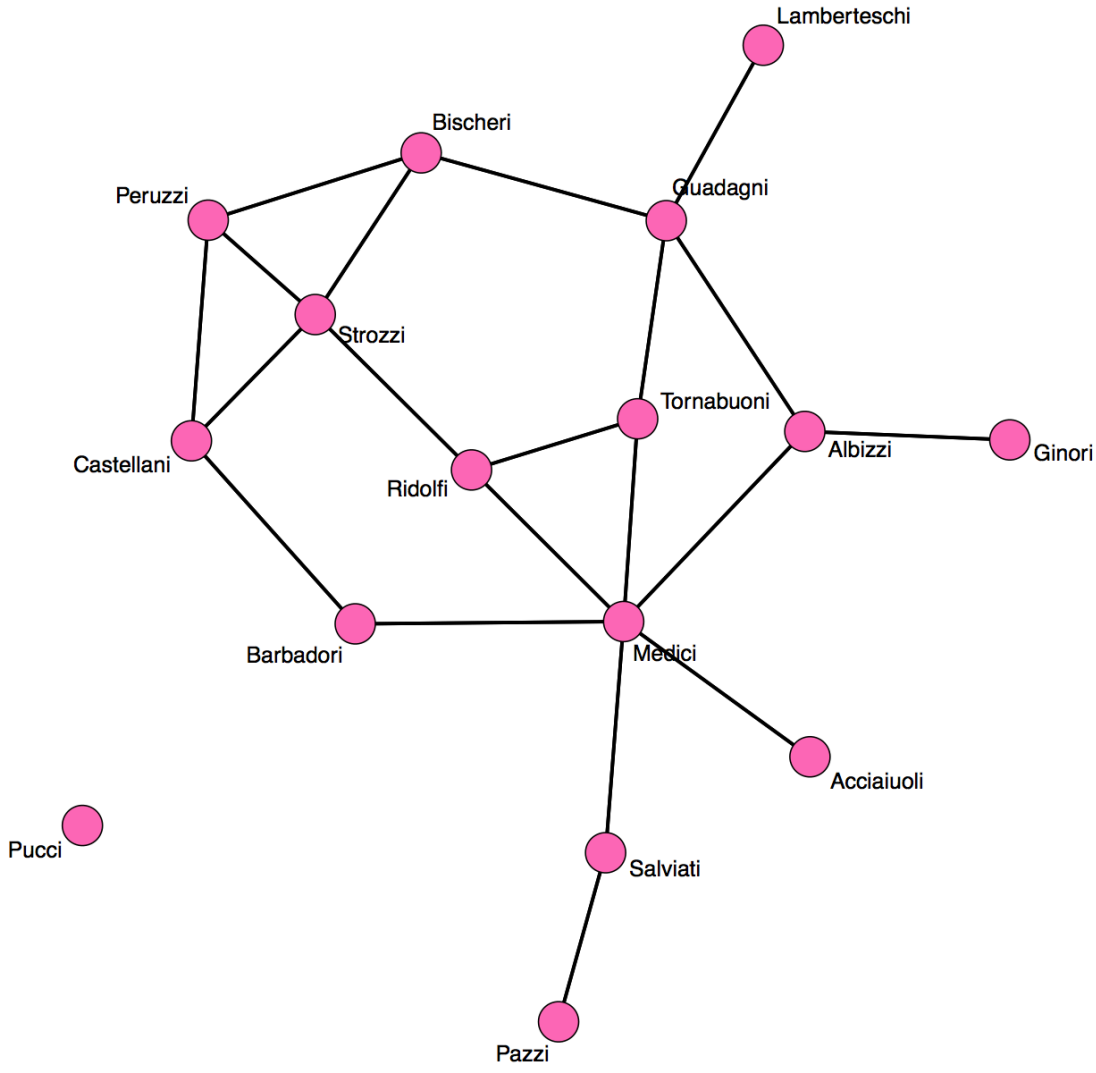


Network of energy supply in the US



A network and its components

- Graph (or network)
- Nodes (or vertices or actors)
- Edges (or links)


Lamberteschi
Bischeri
Peruzzi
Guadagni
Strozzi
Tornabuoni
Albizzi
Ginori
Castellani
Ridolfi
Barbadori
Medici
Acciaiuoli
Pucci
Salviati
Pazzi

A network and its components

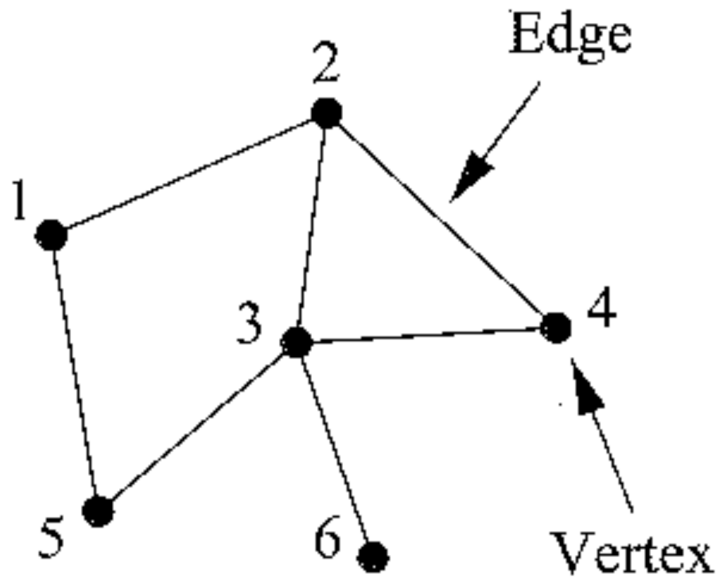
- A **graph** (**network**) is a composed of a set of objects, called **nodes** or **vertices**, with certain pairs of these objects connected by **links** called **edges**.
- Two nodes are **neighbors** if they are connected by an edge.

Adjacency matrix: Math representation of networks

The adjacency matrix or sociomatrix **A** of a *simple* graph (network) is an $n \times n$ matrix with elements

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

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



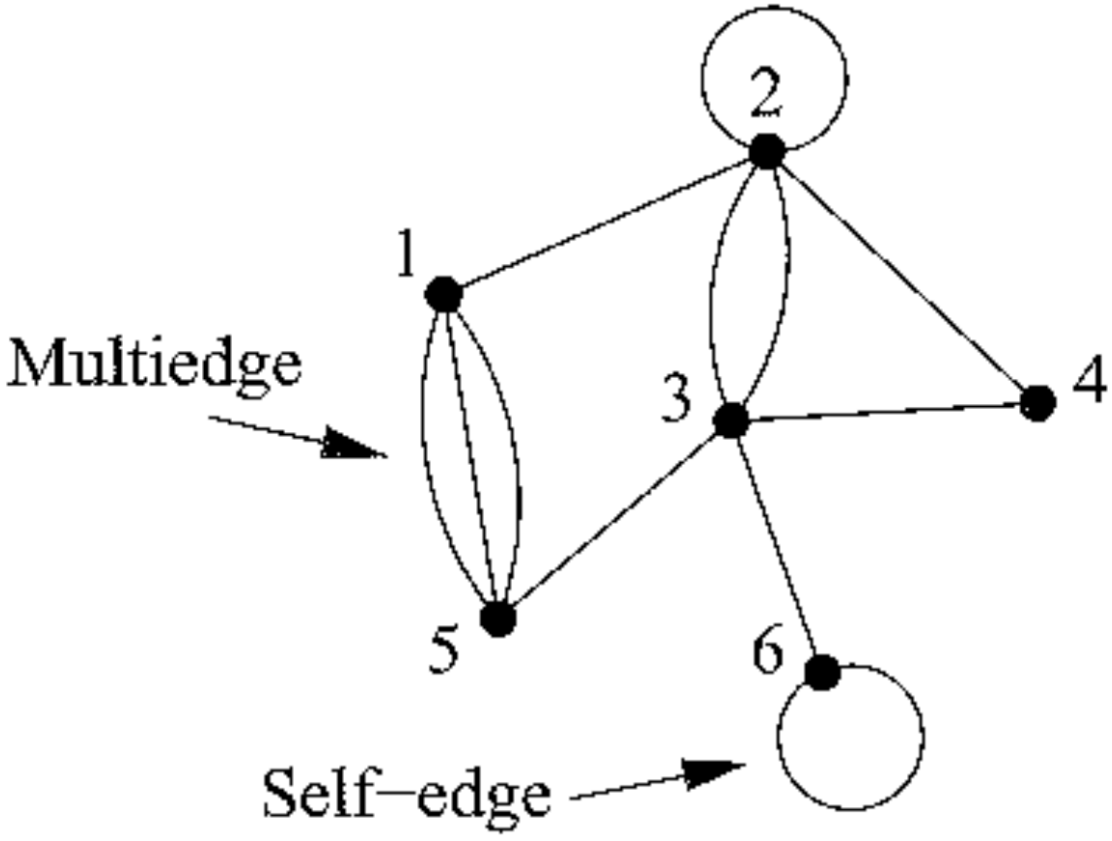
$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{2}$$

What is unique about this matrix?

Types of networks I: Simple graph

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

$$A_{ii} = 0 \quad \forall \quad i. \tag{3}$$


2
1
Multiedge
3
4
5
6
Self−edge

Types of networks II: Weighted network

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

$$A_{ij} = \begin{cases} r \in \mathbb{R} & \text{if there exists an edge between } i \text{ and } j \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

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

Examples?

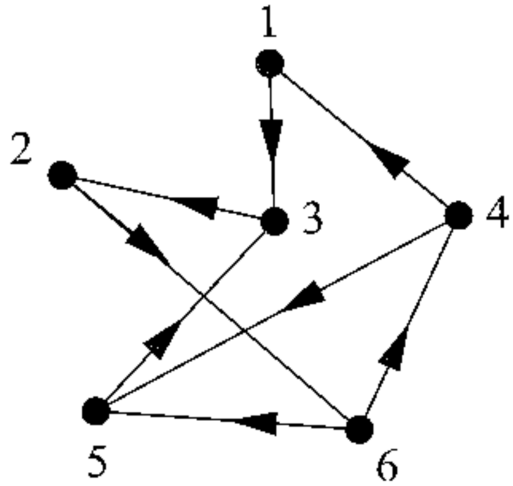
Types of networks III: Directed graph

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

$$A_{ij} = \begin{cases} 1 & \text{if there exists an edge from } j \text{ and } i \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

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

Examples?



$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{6}$$

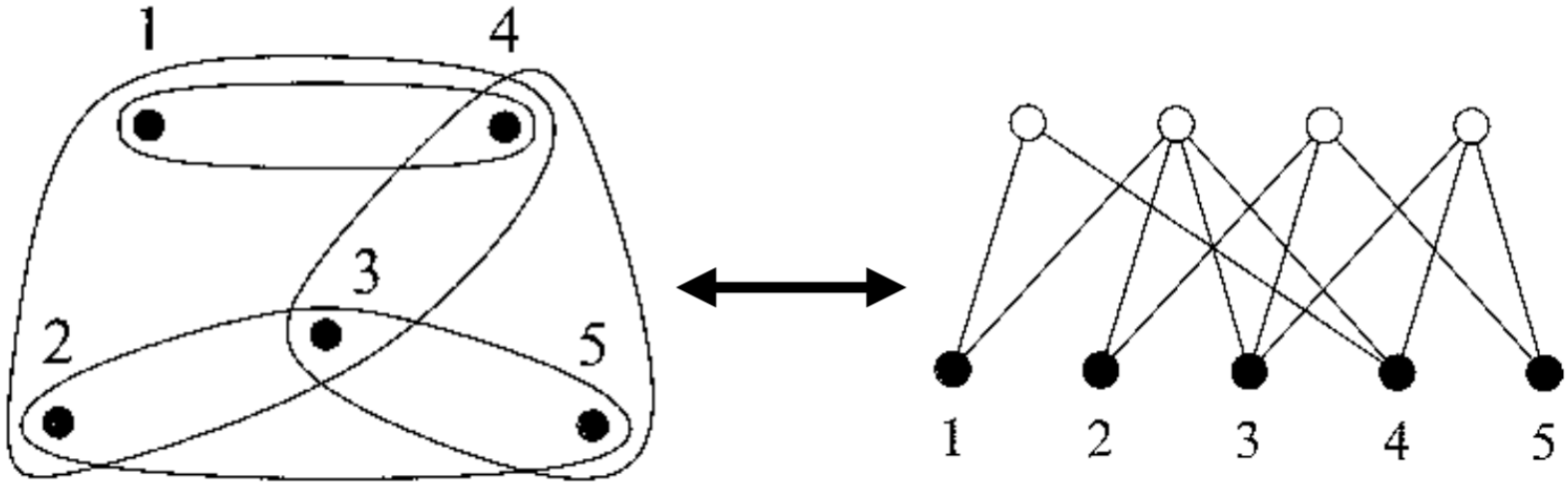
in which columns and rows refer to the starting and ending points respectively.

	Weighted	Binary
Directed		
Undirected		

	Weighted	Binary
Directed		
Undirected		**SIMPLE**

Types of networks IV: 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.



$$B = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \tag{7}$$

where rows correspond to groups and columns refer to actors.

Types of networks IV: Affiliation network

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

$$P_{ij} = \sum_{k=1}^{g} B_{ki} B_{kj} = \sum_{k=1}^{g} B_{ik}^{T} B_{kj}, \tag{8}$$

where k refers to affiliated groups.

Types of networks: Others

- Cocitation network (i.e., studies that are cited together in other studies).
- 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 more.

Social network analysis (SNA): Two pillars

- Descriptive: Study numerical summary measures of networks
- Generative: Study underlying dynamic process of network formation; hypothesis testing; simulation (e.g., ERGM)

Descriptive SNA

- Network connectivity
  - Degree
  - Density
  - Path and geodesic distance
- Node centrality: Measures of node importance in a network
- ... and more (e.g., cosine similarity, cliques, triads, assortive mixing)

Descriptive SNA I: Degree

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

$$k_i = \sum_{j}^{n} A_{ij}, \tag{9}$$

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

Descriptive SNA I: Degree

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

$$\begin{aligned} 2m &= \sum_{i}^{n} k_i \\ \Rightarrow \quad m &= \frac{1}{2} \sum_{i}^{n} k_i \end{aligned} \tag{10}$$

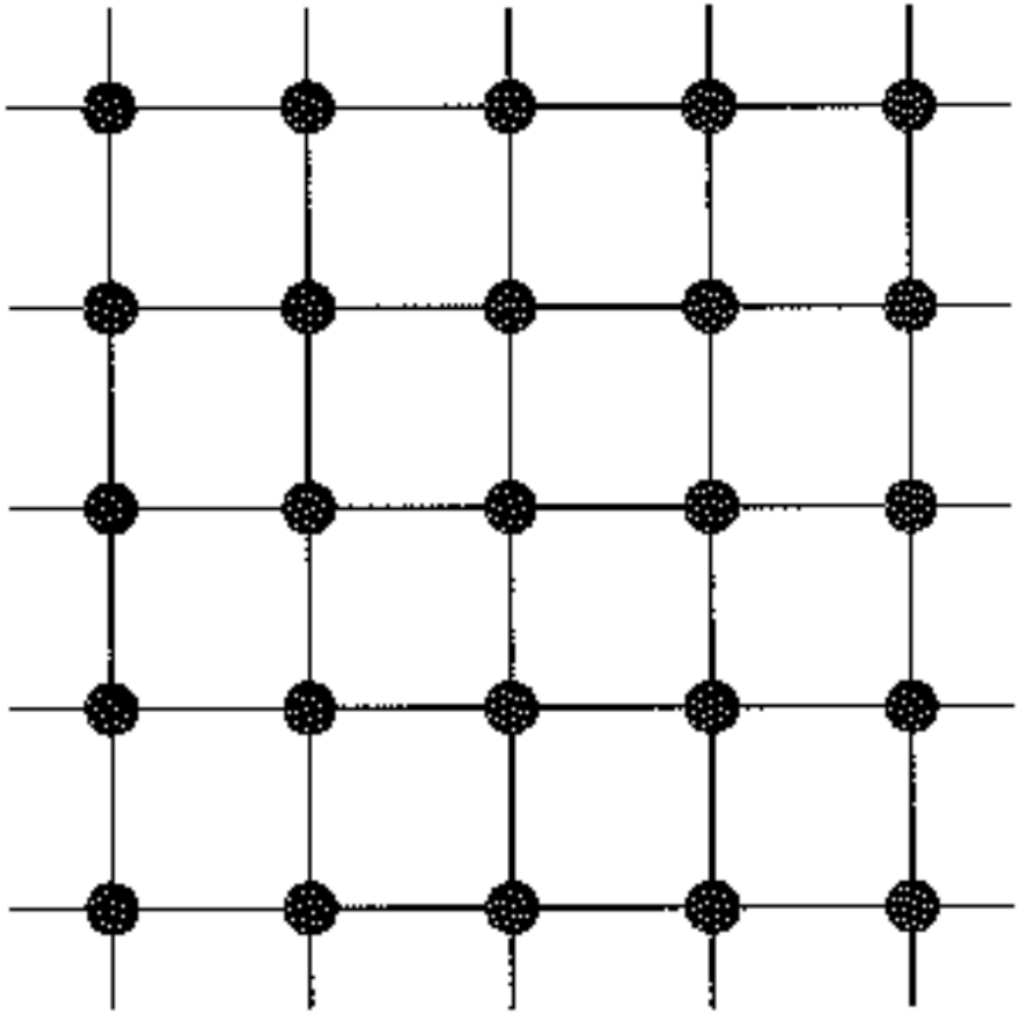
Descriptive SNA I: Degree

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

$$c = \frac{1}{n} \sum_{i}^{n} k_i = \frac{2m}{n} \tag{11}$$

Descriptive SNA I: Degree

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



Descriptive SNA II: 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 II: Density

Mathematically, it is defined as the share of the number of edges out of the max possible number of edges of a graph.

$$\rho = \frac{m}{\binom{n}{2}} = \frac{m}{\frac{1}{2}n(n-1)} = \frac{2m}{n(n-1)} = \frac{c}{n-1}, \tag{12}$$

where c is the mean degree.

What is the big deal here?

Descriptive SNA II: Density

- A network for which the density tends to a constant as the number of nodes tends to infinity is **dense**. That is, ρ remains constant as $n \to \infty$.
- A network for which the density tends to 0 as the number of nodes tends to infinity is **sparse**. That is, $\rho \to 0$ as $n \to \infty$.

Descriptive SNA III: 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 III: 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 III: Path, walk, and distance

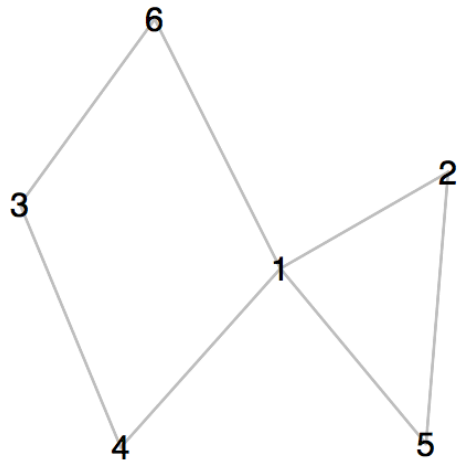
- Existence of *walks* between nodes tell us about network connectivity.
- Existence of *walks of minimal length* tells us about **geodesics**, or geodesic distance, between two nodes in the graph.

Walks of all lengths between a pair of nodes can be counted using matrix multiplication.

Descriptive SNA III: Path, walk, and distance

Define $W = A^k$, where k means we multiply the adjacency matrix of graph $\mathbf{A}$ for k times, then

$$W_{ij} = \text{the number of walks of length } k \text{ between } i \text{ and } j. \tag{13}$$




```
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
```

- How many walks of length 2 are there from i to i?
- How many walks of length 2 are there from i to j?

Descriptive SNA III: Path, walk, and distance

Define $X^{(k)}$, where $k = 1, \cdots, n-1$:

$$\begin{aligned} X^1 &= A \\ X^2 &= A + A^2 \\ &\vdots \\ X^k &= A + A^2 + A^3 + \cdots + A^k \end{aligned} \tag{14}$$

Note:

- $X^{(1)}$ counts the number of walks of length 1 between nodes.
- $X^{(2)}$ counts the number of walks of length ≤ 2 between nodes.
- $X^{(k)}$ counts the number of walks of length $\leq k$ between nodes.

Descriptive SNA III: 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.

$$\begin{aligned} d_{ij} &= \text{length of the shortest path between } i \text{ and } j \\ &= \text{length of the shortest walk between } i \text{ and } j \\ &= \text{the first or min } k \text{ for which } A_{ij}^k > 0 \end{aligned} \tag{15}$$

Can you write a function in R to compute the geodesic distances for a graph **A**?

All measures above capture the connectivity of a network. These connectivity metrics are rather coarse, though.

Descriptive SNA IV: Centrality

All centrality measures try to capture the *importance* of nodes in a network.

Descriptive SNA IV: Centrality

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

$$k_i = \sum_{j}^{n} A_{ij}, \tag{16}$$

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

What is the caveat? What does degree centrality not capture?

Degree centrality treats all neighbors as "equal."

Descriptive SNA IV: 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.

$$\begin{aligned} x_i &= \frac{1}{\lambda} \sum_j A_{ij} x_j \\ \Rightarrow \quad \mathbf{Ax} &= \lambda \mathbf{x}, \end{aligned} \tag{17}$$

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

Why do we need eigenvector centrality?

Descriptive SNA IV: 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.

$$c_i = \frac{1}{\frac{1}{n}\sum_j d_{ij}} = \frac{n}{\sum_j d_{ij}}, \tag{18}$$

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 relevant to the working of the network.

Descriptive SNA IV: 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.

$$x_i = \sum \frac{\sigma_{ijk}}{\sigma_{jk}}, \tag{19}$$

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 IV: 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 more.

Which one to use?

Descriptive SNA IV: Centrality

What if we are dealing with a directed network?

- 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?).

Your personal website.

Social network analysis (SNA): Two pillars

- Descriptive: Study numerical summary measures of networks
- Generative: Study underlying dynamic process of network formation; hypothesis testing; simulation (e.g., ERGM)

Generative SNA

- What is a model? We hope to specify a stochastic model to understand the uncertainty associated with observed outcomes (i.e., networks).
- Different (often unobserved) social and interactive processes can lead to similar networks (e.g., assortative mixing and transitivity).

Generative SNA: Bernoulli (Erdos-Renyi) models

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

$$\text{logit}[P(Y_{ij} = 1 | X = x, \beta)] = \sum_k \beta_k X_{k,ij}, \tag{20}$$

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

The log of the likelihood is then

$$\ell(\beta | Y, x) = \log[P(Y = y | X = x, \beta)], \tag{21}$$

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

Generative SNA: Bernoulli (Erdos-Renyi) models

Y_{ij} are independent and equally likely

$$P(Y_{ij} = 1|X = x, \beta) = \frac{\exp(\beta)}{1 + \exp(\beta)} \quad \forall i, j. \tag{22}$$

Equivalently,

$$\log \text{odds}(Y_{ij} = 1|X = x, \beta) = \beta \quad \forall i, j. \tag{23}$$

Generative SNA: Bernoulli (Erdos-Renyi) models

More abstractly, we can rewrite the model as follows.

$$P(Y_{ij} = 1|X = x, \beta) = \frac{\exp[\beta g(y)]}{[1 + \exp(\beta)]^n}. \tag{24}$$

where $g(y) = \sum_i^n y_{ij}$ and n refers to the number of vertices.
The log of the likelihood is then

$$\ell(\beta|Y, x) = \beta g(y) - n \log[1 + \exp(\beta)]. \tag{25}$$

This model can be easily extended when we include more covariates.

Generative SNA: Bernoulli (Erdos-Renyi) models

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

- 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.

Practice: 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 were friends."
- He also recorded a large amount of information on their individual backgrounds and characteristics.

Open `Practice.R`.

```
> 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.)
```

How do we interpret the coefficient? Recall

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

Therefore,

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

What does this number mean? Degree, density, or..?

Data sources I: Analyze existing network data

- Katya Ognyanova's collection: `http://kateto.net/2016/05/network-datasets/`
- Mark Newman's (Univ of Michigan, Ann Arbor) collection: `http://www-personal.umich.edu/~mejn/netdata/`
- Koblenz Network Collection: `http://konect.uni-koblenz.de/`
- UCI Network Data Repository: `https://networkdata.ics.uci.edu/`
- ... and more

Data sources II: Collect your own data

- Surveys and interviews
- Archives and other 3rd-party records
- Experiment

Examples of SNA research in comparative politics (or political science in general)

- 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

Before you start

- 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!

ccheng11@ucla.edu

Future directions

- Causal inferences
- Machine learning