

Network Analysis I: Statistical Perspectives

PS690 Computational Methods in Social Science

Jiawei Fu

Department of Political Science
Duke University

November 20, 2025

1. Representation
2. Descriptive Analysis
 - 2.1 Degree
 - 2.2 Clustering
 - 2.3 Distance
 - 2.4 Homophily
 - 2.5 Centrality
3. Sampling and Inference
4. Graph and Deep Learning
5. Random Graph
 - 5.1 Bernoulli Random Graph
 - 5.2 Exponential Random Graph Model
 - 5.3 Preferential Attachment
 - 5.4 Small World
 - 5.5 Network-Based Linking

Representation

- We shall think of a network as a collection of nodes and edges; Mathematically, it is studied by graph theory.

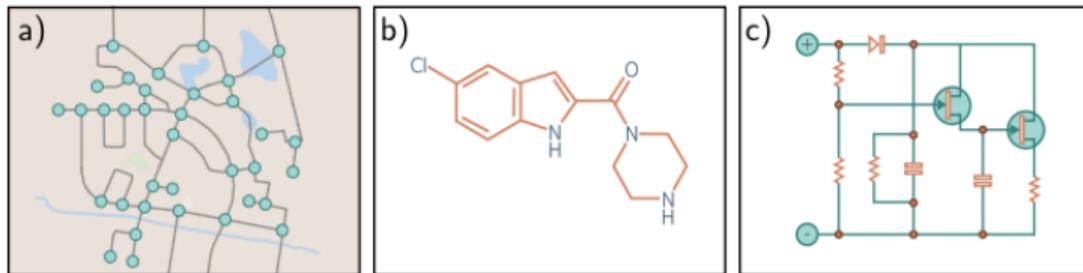


Figure: [Prince, 2023]

Representation

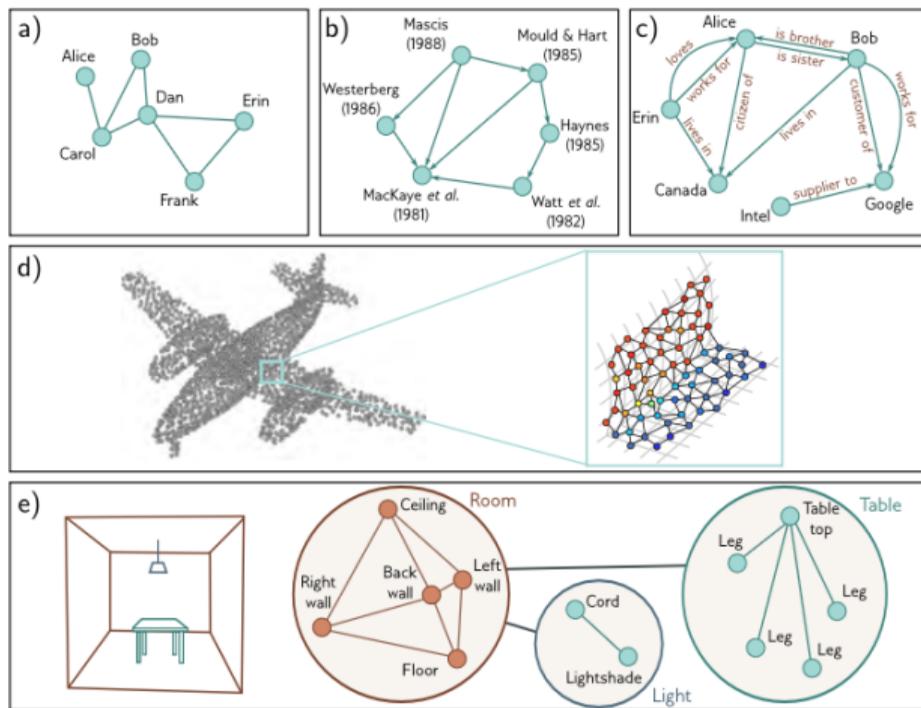


Figure: [Prince, 2023]

Representation

- Let the set $N = \{1, 2, 3, \dots, n\}$ be the set of nodes/vertices; Nodes will also be referred to as vertices, individuals, agents, players, depending on the setting.
- An edge between two nodes signifies a direct relation between them. People also call it link, denoted by g_{ij} .
- We can model link g_{ij} as 0 and 1 to represent whether the link exists or not; we call it is unweighted.
- It can also take other values to represent the intensity of the relationship; we call it is weighted.
- Some links $g_{ij} \neq g_{ji}$ are directed: for example, in Twitter, g_{ij} means i follows j .
- Undirected link has no directionality $g_{ij} = g_{ji}$: for example, friendship, a research collaboration.
- Formally, we call (N, g) as a graph. We will also use N_v and N_e to denote the number of vertices and edges.

Representation

- In addition to the graph structure itself, information is typically associated with each node.
- For example, in a social network, each individual might be characterized by a fixed-length vector representing their interests.
- Sometimes, the edges also have information attached.
- For example, in the road network example, each edge might be characterized by its length, number of lanes, frequency of accidents, and speed limit.
- The information at a node is stored in a *node embedding*, and the information at an edge is stored in an *edge embedding*.

Representation

- The graph can be encoded by three matrices A , X , and E , representing the graph structure, node embeddings, and edge embeddings, respectively.

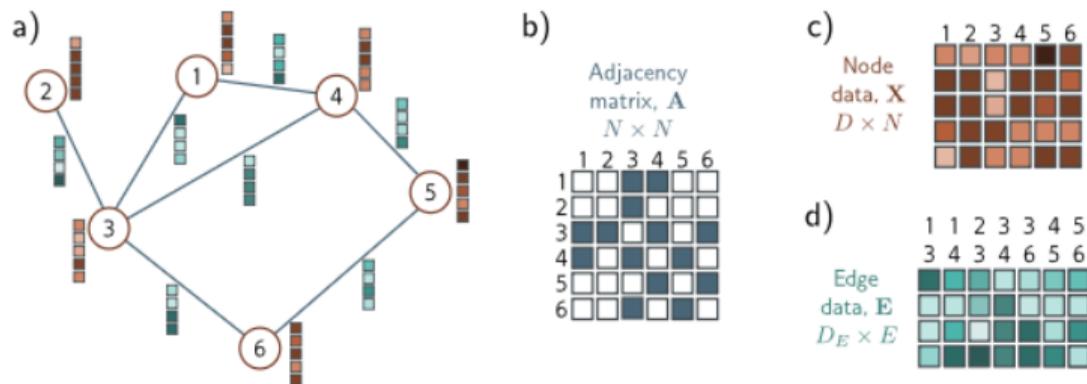
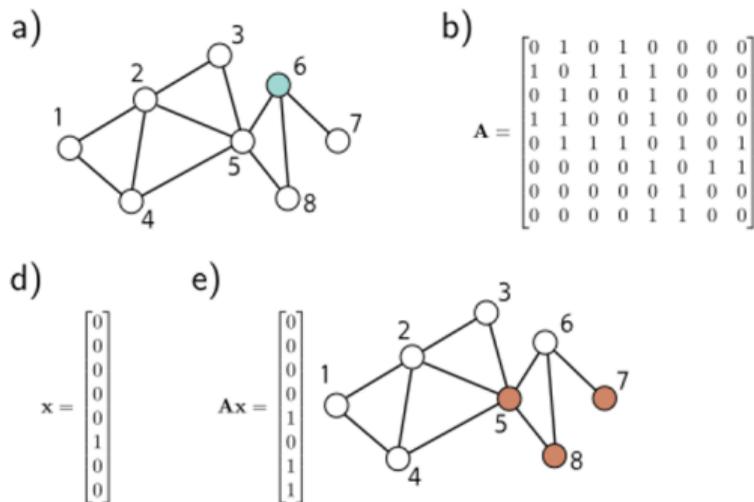


Figure: [Prince, 2023]

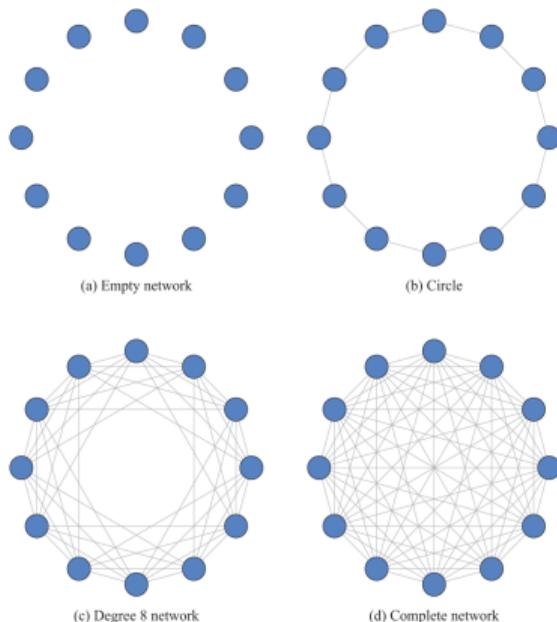
Degree

- Let $N_i(g) = \{j \neq i \mid g_{ij} = 1\}$ be the neighbors of node i in the network g .
- We can use adjacency matrix to find the neighbors of a node using linear algebra.
- Consider an one-hot column vector encoding the n^{th} node; When we pre-multiply it by the adjacency matrix, it returns a vector with ones at the positions of the neighbors.



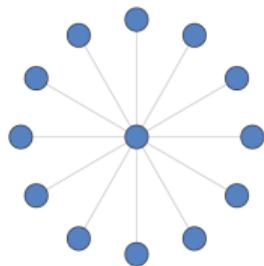
Degree

- The **degree** of node i in network g , $d_i(g) = |N_i(g)|$ is the number of neighbors of i .
- A network is said to be *regular* if every node has the same number of links.

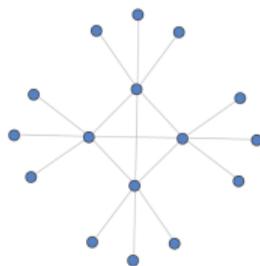


Degree

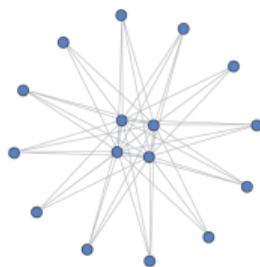
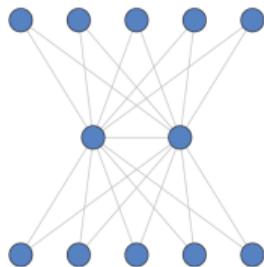
- *Irregular* networks: degree of at least one pair of nodes is different.
- A prominent member is the *core-periphery* network: contains two types of nodes—the core and the periphery.
- The *star* network (hub-spoke network) is a special of core-periphery network, with a singleton core member.



(a) Star



(b) Core periphery: single-link peripheries



- A natural way to describe the links in large networks is to consider their degree distribution.
- Two vertices u, v are said to be *adjacent* if joined by an edge.
- A vertex is *incident* on an edge if the vertex is an endpoint of the edge.
- The *degree* of a vertex v is also defined as the number of edges incident on v .
Out-degree: count the number of edges pointing out from a vertex.
In-degree: count the number of edges pointing towards a vertex.
- Let $P(d)$ be the frequency or fraction of nodes with degree d .
- The degree distribution of a regular network will take on a simple form— $P(d) = 1$ for a single degree and zero for all other degrees.
- For the star network, the degree distribution is $P(n - 1) = \frac{1}{n}$, and $P(1) = \frac{n-1}{n}$, and $P(d) = 0$ for all other degrees.

Degree

- The mean (or average) degree distribution is $d(g) = \frac{1}{n} \sum_{i \in N} d_i(g) = \sum_d P(d)d$.
- The variance of the degree distribution is $var(g) = \sum_{d=0}^{n-1} P(d)[d(g) - d]^2$.
- Suppose the distribution follows Poisson distribution: $\frac{e^{-\lambda} \lambda^d}{d!}$
- An important feature of the Poisson distribution is that most of the nodes will have degrees close to the mean degree λ .

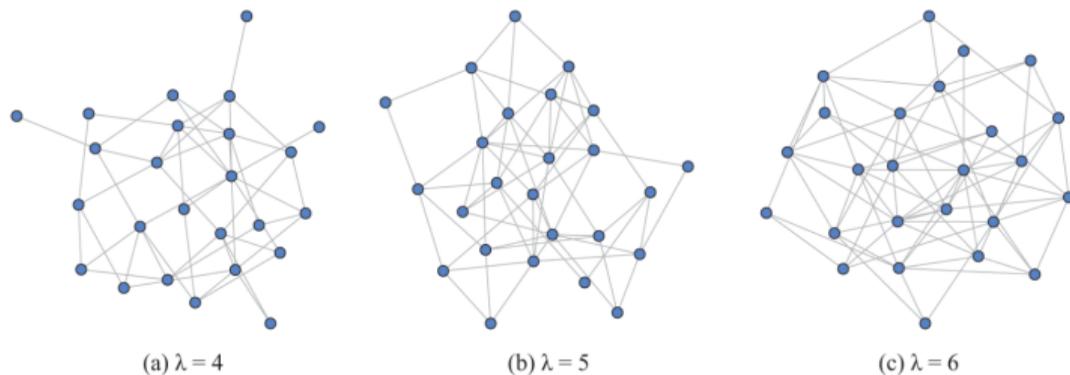


Figure: [Goyal, 2023]

Degree

- Suppose the distribution follows: $f_d \propto cd^{-\gamma}$
- As we raise the value of γ , we see that this leads to a network with a few very highly linked nodes and a large number of poorly linked nodes.

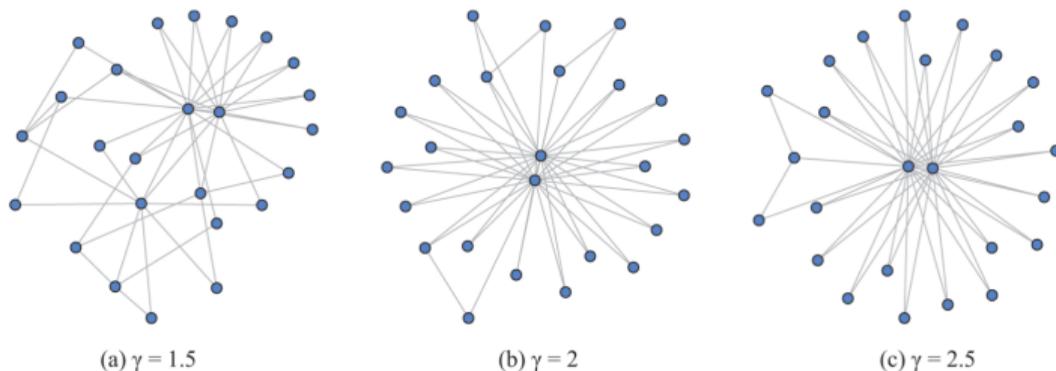
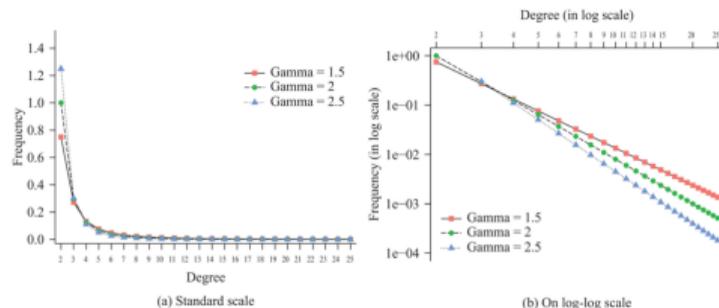


Figure: [Goyal, 2023]

Degree

- Consider power law: $P(d) = cd^{-\gamma}$, where c is a positive constant that normalizes the sum of probabilities to 1.
- A scale-free network is a network whose degree distribution obeys a power law.
- If we take logs on both sides, we get $\log P(d) = \log(c) - \gamma \log(d)$.



- Power law is a functional relationship between two quantities, where a relative change in one quantity results in a relative change in the other quantity proportional to the change raised to a constant exponent: one quantity varies as a power of another.

Degree

- In just the past 10 or 15 years, it has been found that approximate power-law degree distributions appear to be ubiquitous in networks across many areas of the sciences.

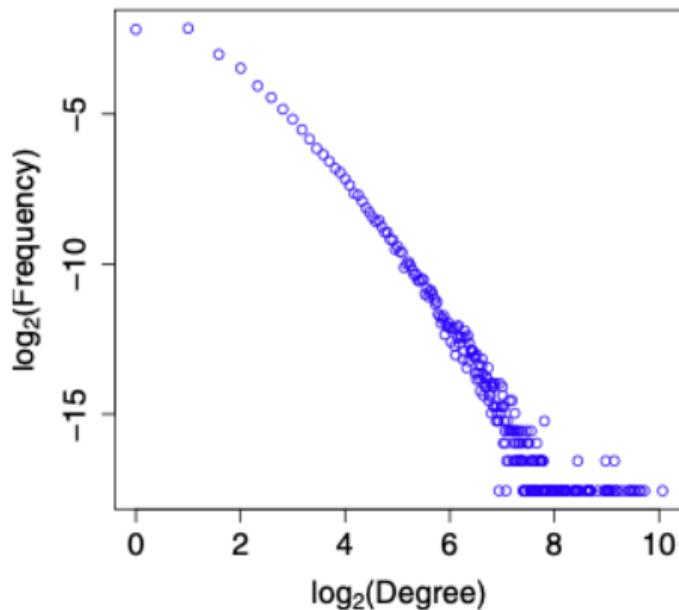


Figure: Router-level Internet network graph [Eric and Kolaczy, 2009]

Heavy Tail

- Recall, a normally distributed random variable has density $\sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(\frac{-(d-\mu)^2}{2\sigma^2}\right)$.
- It goes to zero like $\exp(-d^2)$ as $d \rightarrow \infty$, which is extremely fast.
- A random variable X on \mathbb{R}^+ is called exponentially distributed if, for some $\lambda > 0$, X has density

$$p(d) = \lambda e^{-\lambda d} \quad d \geq 0$$

- The tails of the exponential density go to zero like $\exp(-d)$ as $d \rightarrow \infty$, which is also relatively fast.
- For those random variable X with light-tail, $m(d) = \mathbb{E}e^{dX}$ is finite for at least one d , we rarely draw values deviate more than a few standard deviations from the mean.
- However, for heavy-tailed random variable, extreme outcomes occur relatively frequently.

Heavy Tail

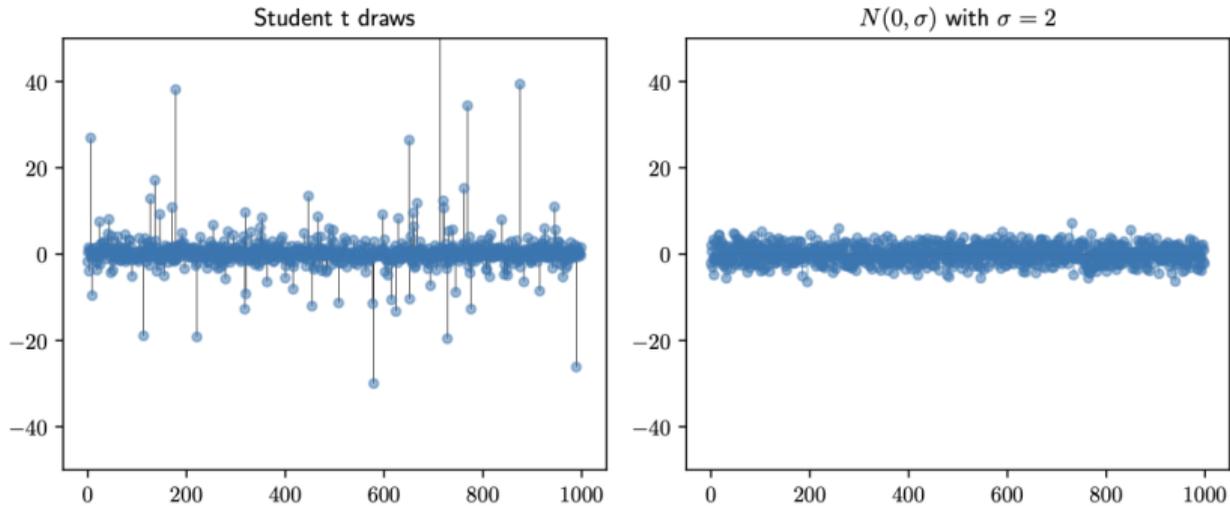


Figure: [Sargent and Stachurski, 2024]

Pareto Tail

- Define counter CDF (CCDF) as $G(d) := P[X > d] = 1 - F(d)$.
- Given $\alpha > 0$, a nonnegative random variable X is said to have Pareto tail with tail index α if there exists a $c > 0$ such that

$$\lim_{d \rightarrow \infty} d^\alpha P[X > d] = c$$

- In other words, the CCDF G of X satisfies

$$G(d) \approx cd^{-\alpha} \text{ for large } d$$

- If X has a Pareto tail for some $\alpha > 0$, then X is also said to obey a *power law*.
- For example, X has a Pareto distribution with parameter \bar{x} , $\alpha > 0$ if ccdf obeys $G(d) = 1$ if $d < \bar{x}$ and $G(d) = (\frac{\bar{x}}{d})^\alpha$ if $d \geq \bar{x}$.
- The density on the set $[\bar{x}, \infty)$ is $p(d) = cd^{-\gamma}$, with $c = \alpha\bar{x}^\alpha$ and $\gamma = \alpha + 1$.
- Every Pareto-tailed random variable is heavy-tailed.

- The most important parameter in the power law is γ . How to estimate it?
- First, we could use linear regression model: $\log P(d) \sim C - \gamma \log(d)$.
- In practice, however, this method is not advisable, due to the disproportionate level of 'noise' in the data at the high degrees.
- We can then again consider using a CDF regression-based approach to estimate γ .

$$G(d) = P(D > d) = 1 - F(d) \sim d^{-(\gamma-1)}$$

- However, the data are not mutually independent now.

Table 1. Sample results of parameter estimation using various methods for 10,000 samples of power-law distribution with $\gamma = 2.500$. Sample result based on 50 runs.

Estimation method	Mean estimated		Bias error
	γ	σ	
Linear	1.590	0.184	36%
Linear 5-points	2.500	0.045	0
Log-2 bins	1.777	0.038	29%
MLE	2.500	0.017	0

- A more rigorous alternative, and one widely used in many fields that routinely encounter power-laws, are estimators of the form, called Hill's estimator

$$\hat{\gamma} = 1 + \hat{\lambda}_k^{-1}, \quad \hat{\lambda}_k^{-1} = \frac{1}{k} \sum_{i=0}^{k-1} \log \frac{d_{(n-i)}}{d_{(n-k)}}$$

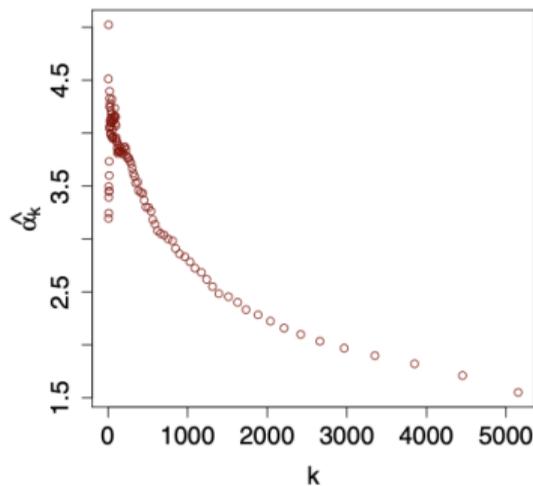
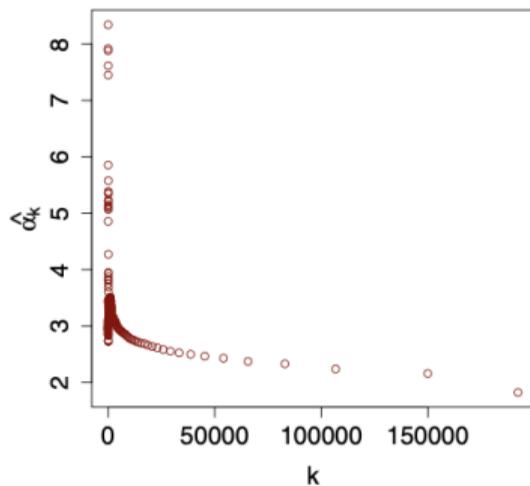
where $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n)}$.

- It comes from MLE. Recall the pdf is $f(d_i; \gamma) = (\gamma - 1)\bar{x}^{\gamma-1}d^{-\gamma} = \frac{\gamma-1}{\bar{x}}\left(\frac{d}{\bar{x}}\right)^{-\gamma}$. The log likelihood is

$$l(\gamma) = \sum_{i=1}^{N_v} \log f(d_i; \gamma) \propto N_v \log(\gamma - 1) - \gamma \sum_{i=1}^{N_v} \log\left(\frac{d_i}{\bar{x}}\right)$$

- Take derivative and set to be zero, we get $\hat{\gamma} = 1 + \left[\frac{1}{N_v} \sum_{i=1}^{N_v} \log\left(\frac{d_i}{\bar{x}}\right)\right]^{-1}$
- How to choose get rid of hand-picking \bar{x} ? For each $k \in \{1, 2, \dots, N_v - 1\}$, let $\bar{x} = d_{N_v-k}$.

- How to choose k ?
- Draw the hill plot: If a power law is credible, the Hill plot should 'settle down'.



Clustering and Clique

- Intuitive idea: if A has two close friends, B and C, then sooner or later, A will introduce them to each other, thereby making it likely that B and C will also become friends.
- The *clustering-coefficient* of a node i (that has two or more links) is defined as

$$Cl_i(g) = \frac{\sum_{l \neq k \in N_i} g_{lk}}{d_i(g)(d_i(g) - 1)}$$

- The numerator is the number of pairs of neighbors of i who have a link while the denominator is the number of all possible pairs among the neighbors.

Clustering and Clique

- The clustering of a network g can be expressed as the mean of clustering across all nodes that have degree two or more

$$Cl(g) = \sum_{i \in N} \frac{Cl_i(g)}{n}$$

- The clustering in the star is therefore zero.
- Clustering coefficients have become a standard quantity used in the analysis of network structure.
- Interestingly, their values have typically been found to be quite large in real-world networks, in comparison to what otherwise might be expected based on classical random graph models.
- In large-scale networks with broad degree distributions, it has frequently been found that the local clustering coefficient varies inversely with vertex degree.

- A *complete* graph is a graph where every vertex is jointed to every other vertex by an edge.
- A *clique* in a network g is a complete subgraph of g : a set of nodes I that for every pair $i, j \in I$, $g_{ij} = 1$.
- A case of common practical interest, particularly in social network analysis, is that of 3-cliques (i.e., triangles).
- In practice, large cliques are relatively rare, as they necessarily require that G itself be fairly dense. order as N_v .

Clique

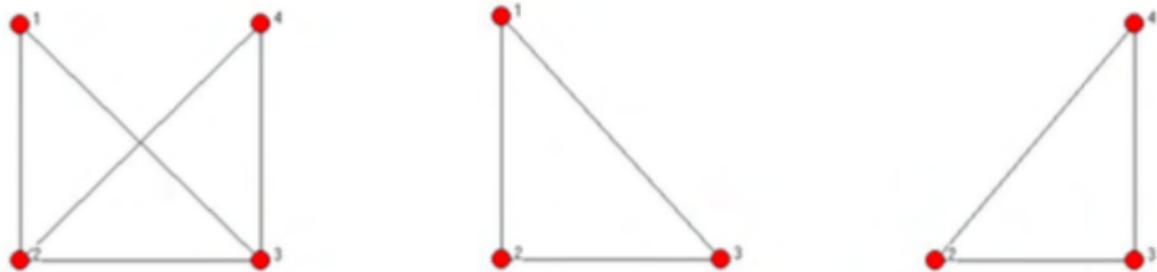


Figure: A Network on Four Nodes and its Two Cliques [Jackson et al., 2008]

Distance

- In many contexts of interest—spread of information or disease is one example—we are interested in how far nodes are from each other.
- An elementary notion is a walk, which is a sequence of nodes in which two nodes have a link between them in the network (i.e., they are neighbors).
- For example, a walk from n_1 to n_k

$$n_1, g_{1,j_1}, n_{j_1}, g_{j_1,j_2}, n_{j_2} \cdots g_{j_{k-1},j_k}, n_k$$

- A node or a link may appear more than once in a walk: a walk is the most general sequence of nodes and links possible in a network.
- The *length* of a walk is simply the number of links it crosses.
- *Trails* are walks without repeated edges; *Paths* are trails without repeated vertices.
- A trail for which the beginning and ending vertices are the same is called a *circuit*.

- A walk with three or more nodes, with no duplication of links, and where the initial and the end nodes are the same is called a *cycle*.
- A graph containing no cycles are called *acyclic*.
- A vertex v in a graph G is said to be reachable from another vertex u if there exists a walk from u to v .
- A network is *connected* if every vertex is reachable from every other.
- A *component* of a graph is a maximally connected subgraph.

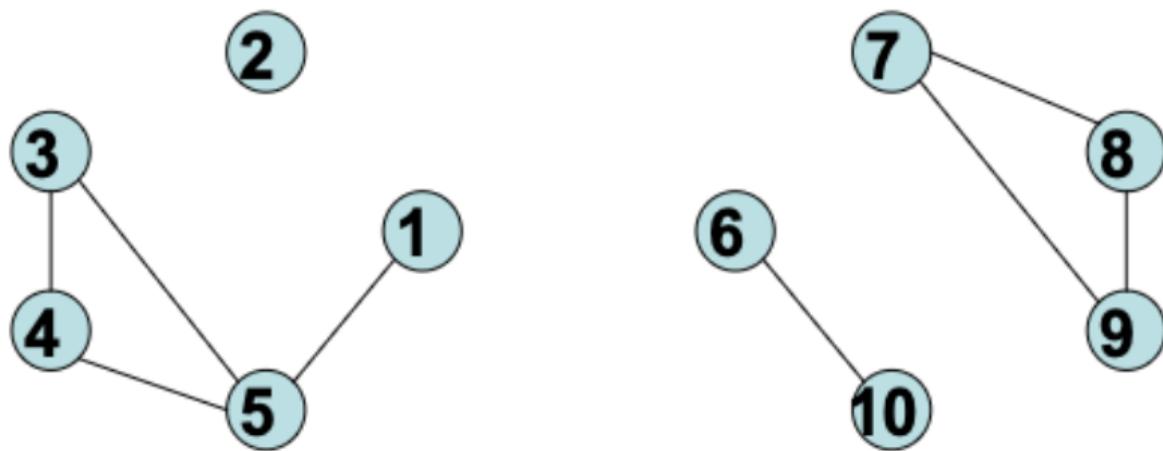


Figure: A network with four components. [Jackson et al., 2008]

Distance

- The *geodesic distance* between two nodes i and j , $d(i, j; g)$ is the length of the shortest path between them.
- The *diameter* of a connected network is equal to the geodesic distance between the pair of nodes that are farthest apart in that network.
- The mean distance is the arithmetic mean of distances across all pairs of nodes
$$\frac{\sum_{i \in N} \sum_{j \in N-i} d(i, j; g)}{n(n-1)}$$
- A celebrated characteristic observed in the giant component of many real-world networks is the so-called *small world* property: large networks tend to have small mean distance, scales as $O(\log n)$.
- The origins of the small-world idea may be traced to the Hungarian writer Frigyes Karinthy.
- Frigyes wrote a short story called “Lancszemek,” in which two characters believed that any two individuals on Earth could be connected to each other through a chain of no more than five acquaintances.

Distance

- Stanley Milgram pioneered the study through a clever experiment where people had to route a letter to another person who was not directly known to them.
- Letters were distributed to subjects in Kansas and Nebraska, who were told the name, profession, and some approximate residential details about a target person who lived in Massachusetts.
- The subjects were asked to pass the letter on to someone whom they knew well and would be likely to know the target or to be able to pass it on to someone else, etc., with the objective of getting the letter to the target.
- While roughly a quarter of the letters reached their targets, the median number of hops for a letter to reach a target was 5 and the maximum was 12.
- Watts and Strogatz report a mean distance of 3.7 in a network among actors where a link indicates that two actors have been in a movie together.

Distance

- Let us see that $O(\log n)$ should not be that surprised.
- Consider a network in which most nodes have similar degrees.
- A connected graph with no cycles is called a *tree*; the disjoint union of such graph is called a *forest*.
- We study diameter in a tree network in which every node has exactly d degrees or degree 1.
- Furthermore, to make the computation simpler, suppose that there is a root node that is exactly distance l from all the leaves. Start from this root node i .
- Each of its neighbors has d links. So each node has 1 edge back to its parent, and $d - 1$ edges going outward to children.
- This means that there are $d + d(d - 1)$ nodes within distance 2 of node i .
- Therefore, the number of nodes within distance k of root node i is

$$d + d(d - 1) + d(d - 1)^2 + \dots + d(d - 1)^{k-1}$$

Distance

- Simplify $d + d(d - 1) + d(d - 1)^2 + \dots + d(d - 1)^{k-1}$ as

$$d \left[\frac{(d - 1)^k - 1}{d - 1 - 1} \right] = \frac{d}{d - 2} ((d - 1)^k - 1)$$

- So it follows that if we want to cover $n - 1$ nodes, it would suffice to have an l :

$$\frac{d}{d - 2} ((d - 1)^l - 1) \geq n - 1$$

- Approximately, we solve $(d - 1)^l = n - 1$ and taking logs, we find l is of order $\frac{\log(n-1)}{\log(d-1)}$; The key point to note is that the diameter grows very slowly as n grows.
- For example, Suppose that the degree of every node is 11. The diameter for a network with 1, 000 nodes is 6, and for a network with 100, 000 nodes, it is 10.

Homophily

- *Homophily* is the tendency of nodes to be linked to others like themselves.
- For example, individuals with an interest in the same sport would like to link with each other.
- For simplicity, let us define the notion of homophily with reference to gender.
- Denote the fraction of men in the population by w_m and the share of women by $w_f = 1 - w_m$. Let H_m denote the mean share of male links among links of men.
- *Relative homophily* captures a straightforward idea: we say that a group of men displays relative homophily if the fraction of links that men have with other men is larger than the fraction of males in the population: $RH_s = H_s - w_s$ for $s \in \{f, m\}$.
- $RH > 0$ indicates homophily; $RH < 0$ indicates heterophily.

- *Inbreeding homophily* goes a step further: the proportion of links within the same group in relation to the fraction of the population that belongs to this group and then normalizes the difference by the maximum bias that a group could possess.
- It is defined as $IH_s = \frac{H_s - w_s}{1 - w_s}$ for $s \in \{f, m\}$.
- In the denominator, if all links connecting people within the same group, $H_s = 1$; then $1 - w_s$ is the maximum bias.
- $IH > 0$ indicates homophily; $IH < 0$ indicates heterophily.

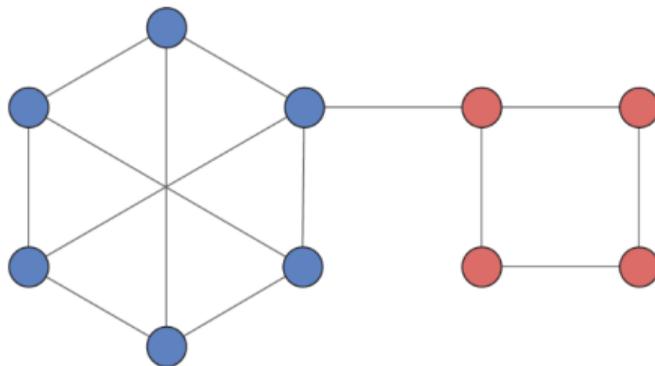


Figure 1.11
Homophily in a network.

Table 1.4
Gender homophily

H_{blue}	H_{red}	w_{blue}	w_{red}	RH_{blue}	RH_{red}	IH_{blue}	IH_{red}
0.9	0.8	0.6	0.4	0.3	0.4	0.75	0.67

Centrality

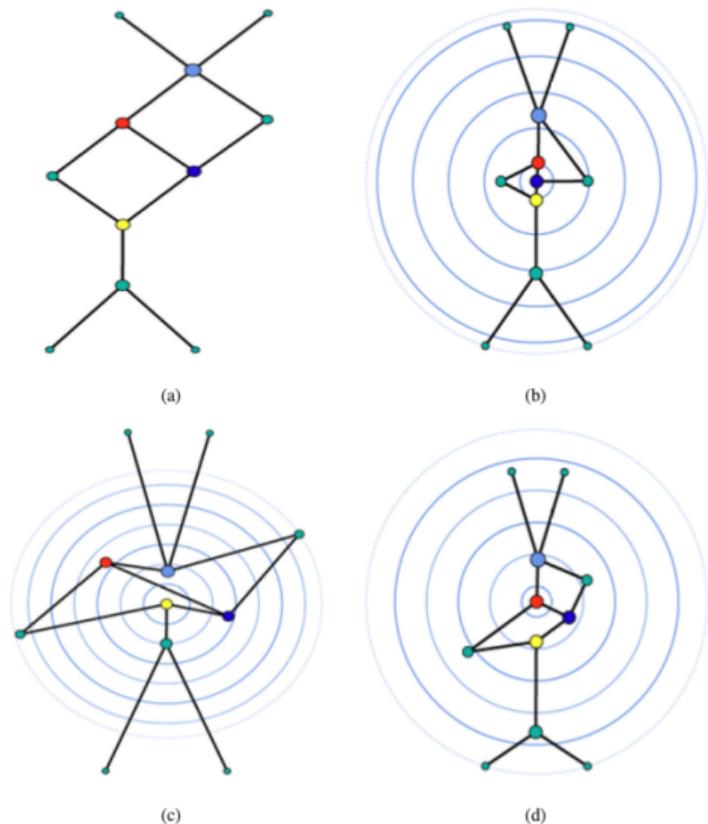
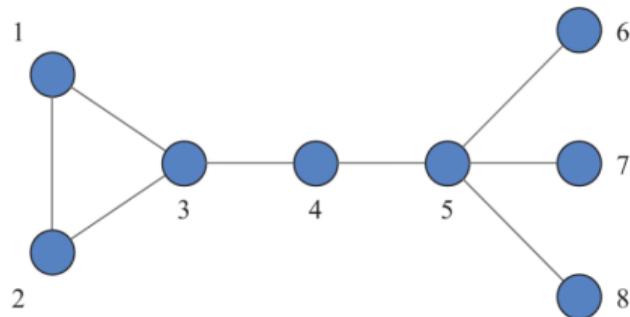


Fig. 4.4 Illustration of (b) closeness, (c) betweenness, and (d) eigenvector centrality measures on the graph in (a). Example and figures courtesy of Ulrik Brandes.

Degree Centrality

- The centrality of a node in a network captures a number of ideas relating to its prominence.
- We introduce four different centrality measures: degree, closeness, betweenness, and eigenvector centrality.
- The simplest notion pertains to the idea of how many links a node has.
- *Degree centrality* measures the relative prominence of a node vis-à-vis other nodes in terms of its degree:

$$C_d(i; g) = \frac{d_i(g)}{n - 1}$$



Closeness Centrality

- Another notion of centrality derives from the idea of proximity: a node is said to be central in a network if the distance from other nodes is small.
- The *closeness centrality* of node i in network g is defined as

$$C_c(i; g) = \frac{n - 1}{\sum_{j \neq i} d(i, j; g)}$$

where denominator is the total distance from node i to all other nodes; To account for the number of nodes, we normalize the measure by multiplying it by the minimum possible total distance.

- This measure of centrality lies between 0 and 1.

Betweenness Centrality

- In some contexts, a node's status may arise from its location between other nodes, for example, due to possibilities of intermediation and brokerage.
- Let us define betweenness for a node i with respect to a pair of other nodes, j and k :

$$b_i^{jk}(g) = \frac{\text{\#shortest paths between } j \text{ and } k \text{ on which } i \text{ lies}}{\text{\#shortest paths between } j \text{ and } k}$$

- Aggregating across all possible other pairs yields us the *betweenness centrality* of a node

$$C_b(i, g) = \frac{1}{\binom{n-1}{2}} \sum_{j, k \neq i} b_i^{jk}(g)$$

where where the denominator is the set of all possible pairs of remaining nodes in the network.

- Betweenness centrality of a node lies between 0 and 1.

Prestige

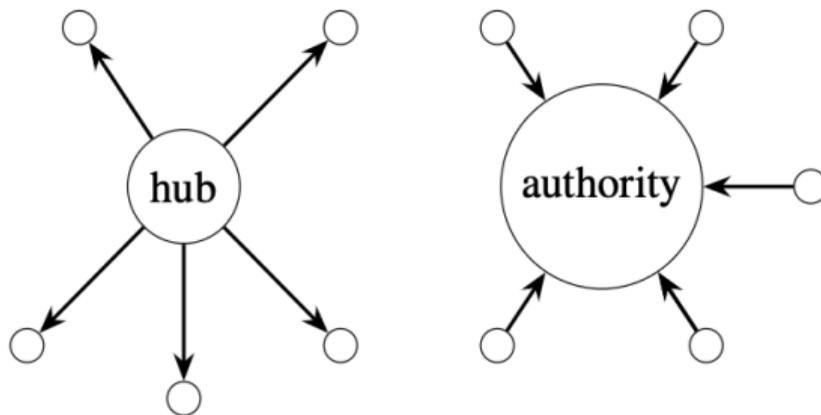
- Another natural idea is that a person's standing in a society depends on the standing of their associates.
- This leads us to consider prestige or influence recursively.
- Katz (1953) proposed that a node's *prestige* is given by

$$P_i^K = \sum_{j \neq i} g_{ij} \frac{P_j^K(g)}{d_j(g)}$$

- Note that if j has more neighbors than i obtains less prestige from being connected to j .
- Let $\hat{g}_{ij} = \frac{g_{ij}}{d_j(g)}$ be the normalized adjacency matrix so that sum across any column is normalized to 1.
- Then $P^K(g) = \hat{G}P^K(g)$; equivalently, $[I - \hat{G}]P^K(g) = 0$. In other words, we need to find the unit eigenvector of \hat{G} .

Eigenvector Centrality

- We may also define a recursive notion of prestige that does not normalize for degrees of neighbors.
- This yields us the *eigenvector centrality*. There are two-kinds of eigenvector centrality.



Eigenvector Centrality

- The hub-based eigenvector centrality $C^e(g)$ of a node is proportional to the sum of the eigenvector centrality of its neighbors:

$$C_i^e(g) = \frac{1}{r(G)} \sum_{j \neq i} g_{ij} C_j^e(g)$$

where $r(G) = \max\{|\lambda| : \lambda \text{ is the eigenvalue of adjacent matrix}\}$ is the spectral radius.

- A vertex i is highly ranked if (a) there are many edges leaving i , (b) these edges have large weights and (c) the edges point to other highly ranked vertices.
- The matrix form is

$$r(G)C^e(g) = GC^e(g)$$

- Thus, $C^e(g)$ is the eigenvector of G with respect to largest eigenvalue.

Eigenvector Centrality

- The authority-based eigenvector centrality $C^e(g)$ is

$$C^e(g) = \frac{1}{r(G)} G^T C^e(g)$$

- The difference is transpose.
- Element-by-element, this is

$$C_j^e(g) = \frac{1}{r(G)} \sum_{i \neq j} g_{ij} C_i^e(g)$$

- We see C_j^e will be high if many nodes with high eigenvector authority ranking link to j .

Eigenvector Centrality

- Katz (1953) also introduced a second measure of centrality in which the prestige of a node is a weighted sum of the walks that emanate from it, and a walk of length k is worth a^k for some parameter $0 < a < 1$.
- Katz's *second prestige measure* is given by

$$P^{K2}(g; a) = [I - aG]^{-1} aG1$$

- We can generalize Katz's second prestige measure to obtain the *Bonacich measure of centrality*:

$$CB(g; a, b) = [I - bG]^{-1} aG1$$

where $a > 0$ and $b > 0$ are scalars and b is sufficiently small; b now provides us the weights for walk length.

Centrality

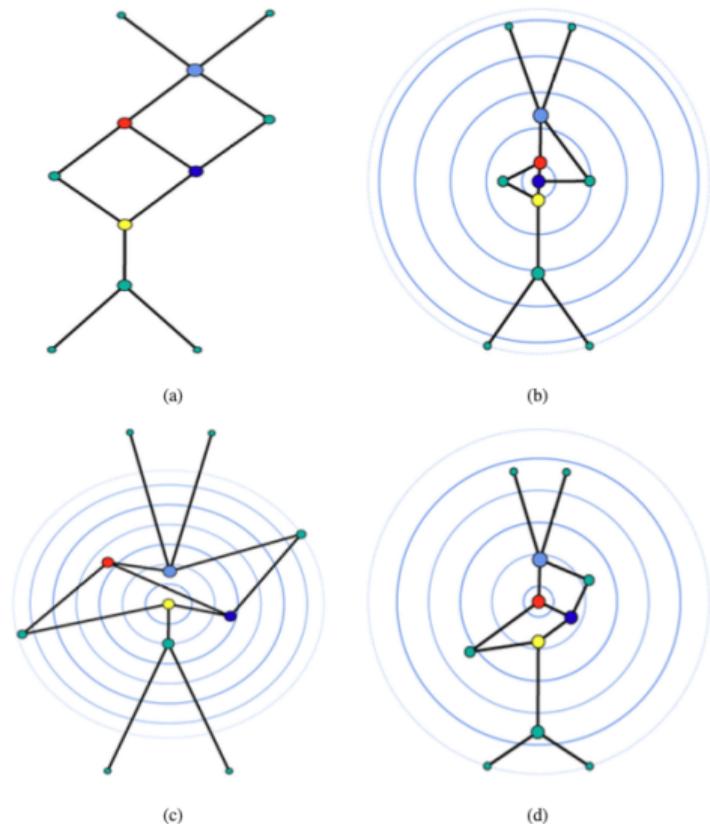


Fig. 4.4 Illustration of (b) closeness, (c) betweenness, and (d) eigenvector centrality measures on the graph in (a). Example and figures courtesy of Ulrik Brandes.

Sampling and Inference

- In reality, we seldom observe the whole network; instead, sampled graphs may be more the rule than the exception in practice.
- For example, in social networks, while it may be possible to fully construct the friendship network among school children in a small classroom, it may be cumbersome to attempt to do so for all employees in a large corporation.
- Formally, assume there is a system under study that may be represented by a network graph G , which we will call the *population graph*.
- We take measurements that effectively yield a sample of vertices and edges, which we compile into a *sampled graph* $G^* = (V^*, E^*)$.
- Now suppose that there is a particular characteristic of G , denoted $\eta(G)$, that is of interest: number of edges, average degree, or distribution of betweenness centrality, etc.
- The question thus arises as to whether we may still obtain a useful estimate of $\eta(G)$, say $\hat{\eta}$ from G^* .

Sampling and Inference

- Intuitively, it is attractive to think that we might simply estimate $\eta(G)$ by $\hat{\eta} = \eta(G^*)$ (plug-in).
- Unfortunately, in estimating graph characteristics from sample graphs, this line of reasoning can often go awry.
- For example, suppose $\eta(G) = \frac{1}{N_V} \sum_{i \in V} d_i$, average degree of a graph G .
- Let G^* be based on the n vertices $V^* = \{i_1, \dots, i_n\}$, and denote its observed degree by $\{d_i^*\}_{i \in V^*}$.
- The plug-in estimator is $\hat{\eta} = \eta(G^*) = \frac{1}{n} \sum_{i \in V^*} d_i^*$.
- To evaluate this estimator, we consider two sampling designs by which G^* might be obtained.

Sampling and Inference

- In both cases, we begin with a simple random sample without replacement of n vertices $V^* = \{i_1, \dots, i_n\}$.
- Design 1: for each vertex $i \in V^*$, we observe all edges $\{i, j\} \in E$ involving i . each such edge becomes an element of E^* .
- Design 2: we only observe, for each pair $i, j \in V^*$, whether or not $\{i, j\} \in E$. if it is, that edge becomes an element of E^* .

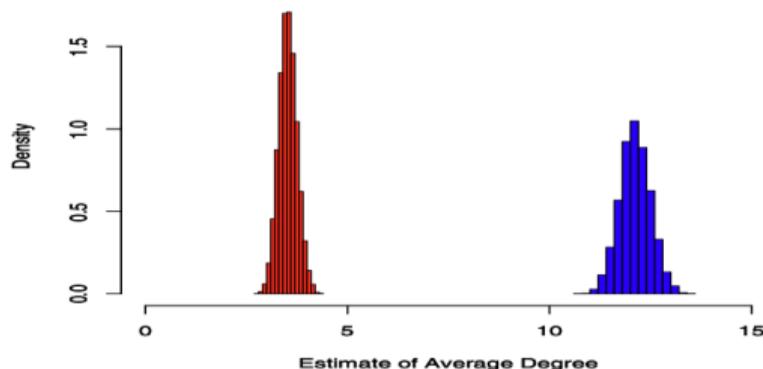


Figure: Design 2 (red) is biased.

Review of Horvitz-Thompson Estimation

- Suppose we have a population $U = \{1, 2, \dots, N_u\}$ and each $i \in U$ is associated a value y_i of interest.
- Let $\tau = \sum_i y_i$ and $\mu = \frac{\tau}{N_u}$ be the total and average values of y in the population.
- Let $S = \{i_1, \dots, i_n\}$ be the sample.
- In canonical case in which S is chosen by drawing n units uniformly from U , with replacement, a natural estimate of μ is sample mean $\bar{y} = \frac{1}{n} \sum_{i \in S} y_i$, and total is estimated by $\hat{\tau} = N_u \bar{y}$.
- They are unbiased, and variance are $V(\bar{y}) = \frac{\sigma^2}{n}$ and $V(\hat{\tau}) = \frac{N_u^2 \sigma^2}{n}$, where σ^2 is the variance of the values y in the full population.
- σ^2 can be estimated by sample analogue $\frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y})^2$ unbiasedly.

Review of Horvitz-Thompson Estimation

- What if unequal probability sampling?
- Then we should use Horvitz-Thompson estimator, for the total is

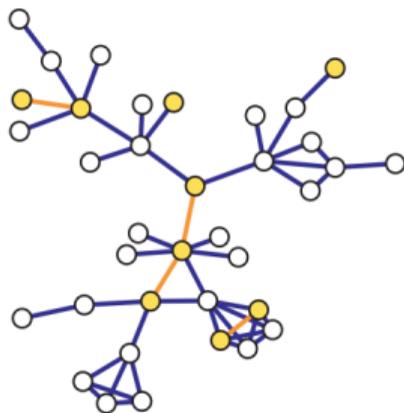
$$\hat{\tau} = \sum_{i \in S} \frac{y_i}{\pi_i}$$

where π_i is the probability that $i \in U$ is sampled in S .

- The corresponding estimate of μ is $\hat{\mu}_\pi = \frac{1}{N_u} \hat{\tau}_u$.
- Let Z_i be a binary variable to indicate whether i is in S .
- We can see that $\mathbb{E}[\hat{\tau}] = \mathbb{E}[\sum_{i \in S} \frac{y_i}{\pi_i}] = \mathbb{E}[\sum_{i \in U} \frac{y_i}{\pi_i} Z_i] = \sum_{i \in U} \frac{y_i}{\pi_i} \mathbb{E}[Z_i] = \tau$.
- The variance is $V[\hat{\tau}] = \sum_{i \in U} \sum_{j \in U} y_i y_j (\frac{\pi_{ij}}{\pi_i \pi_j} - 1)$, and can be estimated by $\sum_{i \in S} \sum_{j \in S} y_i y_j (\frac{1}{\pi_i \pi_j} - \frac{1}{\pi_{ij}})$.

Sampling and Inference

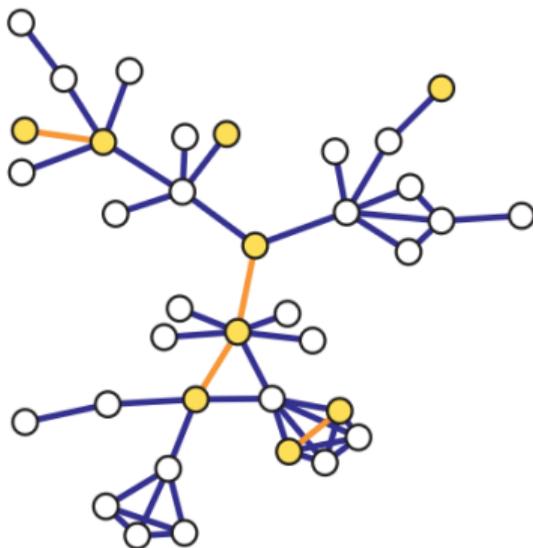
- Graph sampling designs are somewhat distinct from typical sampling designs in non-network contexts.
- Often these designs can be characterized as having two stages: a selection stage (e.g., vertices), followed by an observation stage (e.g., edges).
- Consider Design 2: induced subgraph sampling.



- $\pi_i = \frac{n}{N_v}$ and $\pi_{i,j} = \frac{n(n-1)}{N_v(N_v-1)}$

Sampling and Inference

- Similar to induced subgraph sampling, incident subgraph sampling selecting n edges first.



- $\pi_{i,j} = \frac{n}{N_e}$ and

$$\begin{aligned}\pi_i &= P(\text{vertex } i \text{ is sampled}) \\ &= 1 - P(\text{no edge incident to } i \text{ is sampled}) \\ &= \begin{cases} 1 - \frac{\binom{N_e - d_i}{n}}{\binom{N_e}{n}} & \text{if } n \leq N_e - d_i \\ 1 & \text{if } n > N_e - d_i \end{cases}\end{aligned}$$

- Design 1 is called Star sampling.
- This design may expand sampled V^* if we observe vertices to which these edges are incident.
- $\pi_i = \frac{n}{N_v}$ and $\pi_{i,j} = P(\text{neither } i \text{ nor } j \text{ are sampled}) = 1 - \frac{\binom{N_v - 2}{n}}{\binom{N_v}{n}}$.

Sampling and Inference

- Note that, given a population U and unit value y , various graph summary $\eta(G)$ can be written in a form involves a total $\tau = \sum_{i \in U} y_i$.
- For example, the average degree is $\frac{1}{N_v} \sum_{i \in V} d_i$ is $U = V$ and $y_i = d_i$.
- So, let us consider how to estimate vertex totals $\tau = \sum_{i \in U} y_i$.
- We know the Horvitz-Thompson estimator for vertex totals is

$$\hat{\tau}_\pi = \sum_{i \in V^*} \frac{y_i}{\pi_i}$$

- We can also extend it to totals on vertex pairs:

$$\hat{\tau}_\pi = \sum_{i,j \in V^*} \frac{y_{ij}}{\pi_{i,j}}$$

Sampling and Inference

- So far, we assume N_v was know.
- Many times this is simply not true. For example, there are many populations that are 'hard to find'.
- One solution is capture-recapture estimators.
- The simplest version of capture-recapture involves two stages of simple random sampling without replacement, yielding two samples, say S_1 and S_2 .
- In the first stage, the sample S_1 of size n_1 is taken, and all of the units in S_1 are 'marked.'
- All of the units in S_1 are then 'returned' to the population, and, at the second stage of sampling, a sample of size n_2 is taken from U .
- The estimator $\frac{n_2 n_1}{m}$, where m is number of marked units observed in the second sample, can be used to estimate N_u .

- We can use deep learning to learn the graph internal representation.
- There are several tasks we can consider.
 1. **Graph-level tasks:** The network assigns a label or estimates one or more values from the entire graph, exploiting both the structure and node embeddings. For example, whether a molecule is poisonous to human beings or not; the topic of the node based on hyperlinks and citations.
 2. **Node-level tasks:** The network assigns a label (classification) or one or more values (regression) to each node of the graph, using both the graph structure and node embeddings. For example, classify the nodes according to whether they belong to the wings or fuselage.
 3. **Edge prediction tasks:** The network predicts whether or not there should be an edge between nodes n and m . For example, in the social network setting, the network might predict whether two people know and like each other and suggest that they connect if that is the case.

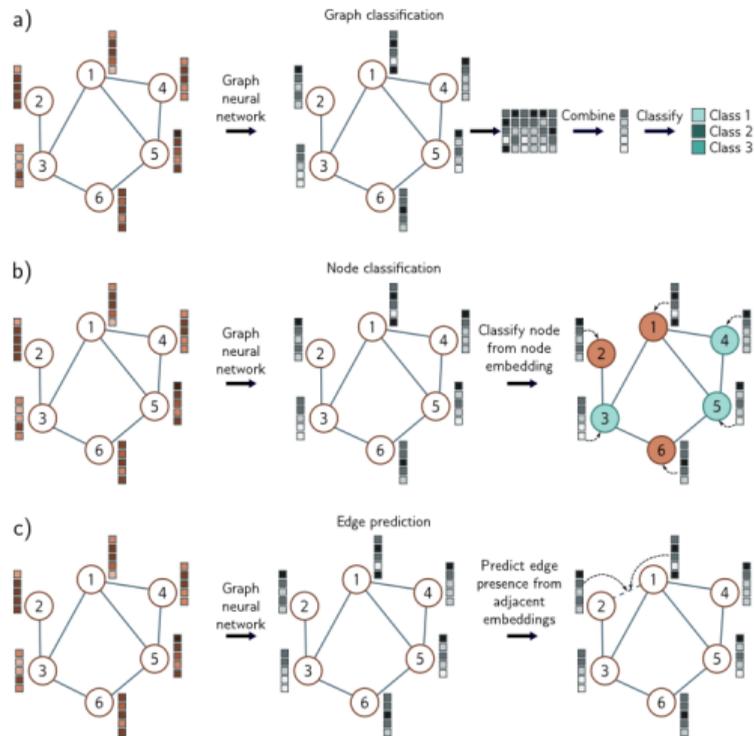


Figure: [Prince, 2023]

Graph Convolutional Networks

- An image can be viewed as a specific instance of graph, in which the nodes are the pixels and the edges represent pairs of pixels that are adjacent in the image.
- We briefly introduce convolutional graph neural networks, or GCNs.
- Each layer of GCN is a function F with parameters Φ that takes the node embeddings and adjacency matrix and outputs new node embeddings:

$$H_1 = F[X, A, \phi_0]$$

$$H_2 = F[H_1, A, \phi_1]$$

$$\vdots$$

$$H_k = F[H_{K-1}, A, \phi_{K-1}]$$

- Therefore, H_k contains the modified node embeddings at the k^{th} layer.

Graph Convolutional Networks

- Recall that a convolution updates a variable by taking a weighted sum of information from its neighbors.
- One way to think of this is that each neighbor sends a message to the variable of interest, which aggregates these messages to form the update.

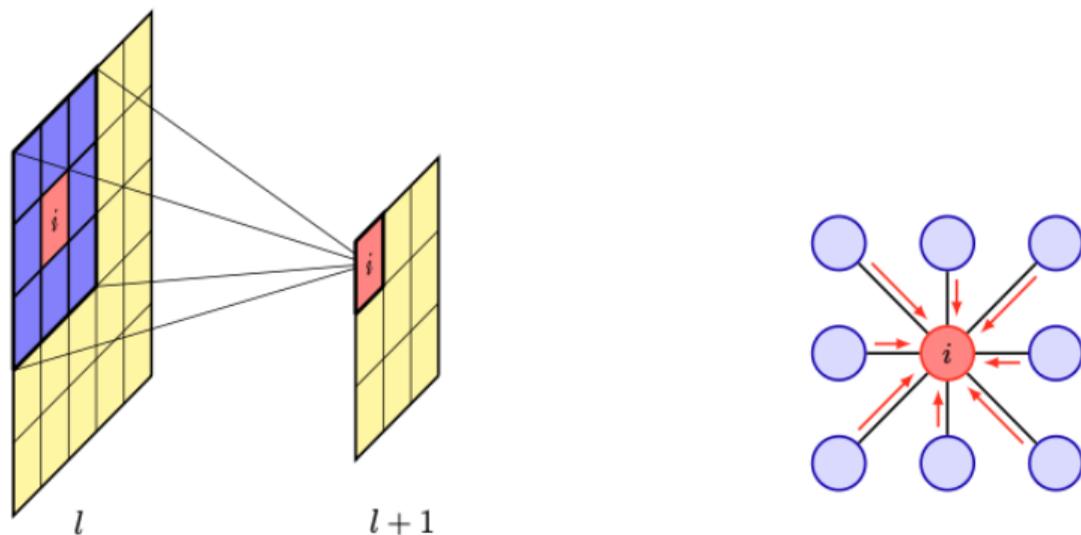
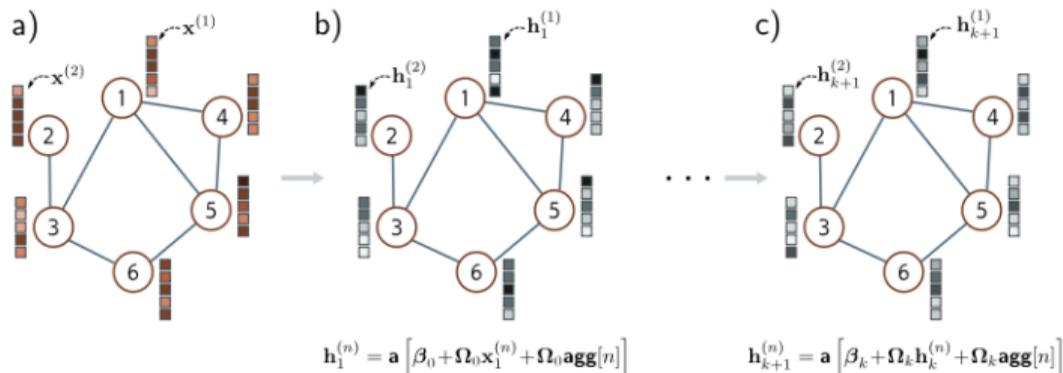


Figure: [Bishop and Bishop, 2023]

Graph Convolutional Networks

- In GCN, at each node n in layer k , we aggregate information from neighboring nodes by summing their node embeddings $agg[n, k] = \sum_{m \in N_n(g)} h_k^m$.
- Then, we update the embedding of each node by

$$h_{k+1}^n = a[\beta_k + \Omega_k h_k^n + \Omega_k agg[n, k]]$$



Graph Convolutional Networks

- Since each node in a given layer of the network is updated by aggregating information from its neighbours in the previous layer, this defines a receptive field analogous to the receptive fields of filters used in CNNs.
- As information is processed through successive layers, the updates to a given node depend on a steadily increasing fraction of other nodes in earlier layers.

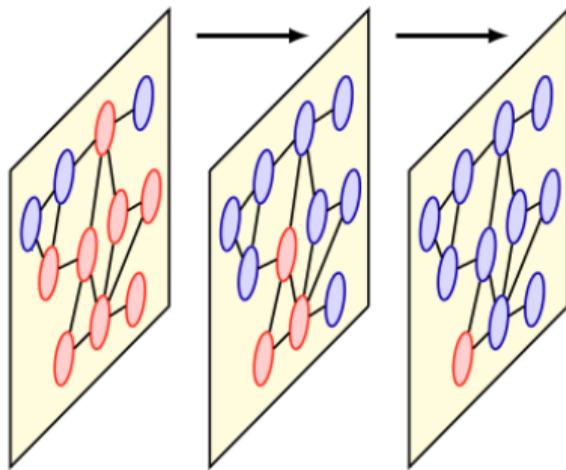


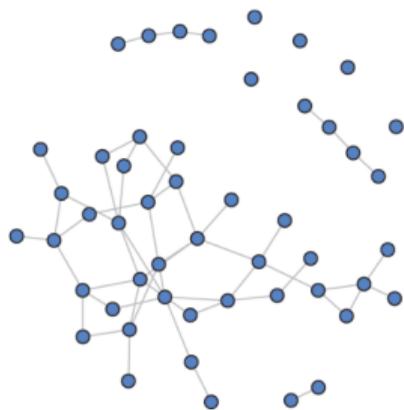
Figure: [Bishop and Bishop, 2023]

How Networks form?

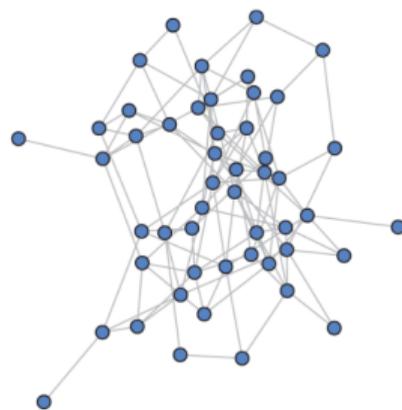
- We introduce the Erdős–Rényi model of random graphs.
- Consider n nodes and an equal probability $p \in [0, 1]$ for a link to form between any two of the nodes.
- Equivalently, each $n(n - 1)/2$ possible edges between vertices are independently present with probability p .
- There exists another version of the model, where we select m edges from $n(n - 1)/2$ at random (there is a small and annoying amount of dependence caused by picking a fixed number of edges)
- What is the structure of the network generated through this process?
- For instance, what is the distribution of connections? Are most nodes in the network connected? What is the distance between the nodes?

Random Network

- There are basically two parameters: p and n .
- Let us see some examples first. How does p affect the connectivity of the graph?



(a) Probability of linking, $p = 0.05$

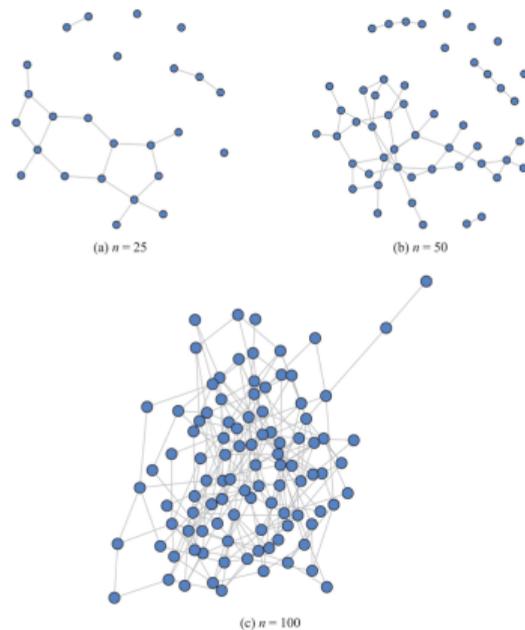


(b) Probability of linking, $p = 0.10$

- In panel (a), there are multiple components, and the largest group of connected nodes—the so-called giant component—is relatively small.
- In panel (b), the graph is connected (i.e., it contains only one component)

Random Network

- If we fix $p = 0.05$ and increase n , we observe:



- Do you have any intuition for why?

Degree Distribution

- We are interested in the Degree distribution.
- What is the probability that node i has degree k ?
- Hint: For any given node i , there are $n - 1$ other nodes.
- Therefore,

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} = \frac{(n-1)!}{k!(n-1-k)!} p^k (1-p)^{n-1-k}$$

- Then, due to the above binomial distribution, the expected degree of each node is $(n-1)p$.
- Let us consider asymptotics by $n \rightarrow \infty$ and fixing $\lambda = (n-1)p$:

$$\lim_{n \rightarrow \infty} P(k) = e^{-\lambda} \frac{\lambda^k}{k!},$$

which is the Poisson distribution.

Connectivity

- We next discuss the relationship between the number of nodes n , the probability of linking p , and the connectedness of the network.
- As we vary p and n , we would like to ask if connectedness holds.
- From previous examples, we observe that as we increase n , connectedness would be possible for lower p .
- A key building block in the analysis is the concept of threshold function.
- Let $A(N)$ be the set of networks that exhibit a property (e.g. particular nodes have some number of links or connectedness of the graph as a whole).
- A *threshold function* for property $A(N)$ is a function $t(n)$ such that

$$P[A(N)|p(n)] \rightarrow 1 \text{ if } \frac{p(n)}{t(n)} \rightarrow \infty$$

$$P[A(N)|p(n)] \rightarrow 0 \text{ if } \frac{p(n)}{t(n)} \rightarrow 0$$

Connectivity

- If such a function exists, then we shall say that there is a phase transition at the threshold: the qualitative properties of the networks generated undergo a marked transformation when we move from slightly below to slightly above the threshold.
- Take an example. Let us take up the property that node 1 has at least one link: $A(N) = \{g | d_1(g) \geq 1\}$.
- In the Poisson graph with n nodes, the probability that node 1 has zero links is $(1 - p)^{n-1}$.
- Thus, the probability of $A(N)$ is $1 - (1 - p)^{n-1}$.
- Consider the function $p(n) = \frac{r}{n-1}$ and note that $\lim_n (1 - \frac{r}{n-1})^{n-1} = e^{-r} \in (0, 1)$.
- Now Let $t(n) = \frac{1}{n-1}$. If $\frac{p(n)}{t(n)} \rightarrow \infty$, then $p(n) \geq \frac{r}{n-1}$ for any r and large enough n . And thus $\lim_n (1 - \frac{r}{n-1})^{n-1} \leq e^{-r}$ for any r , and thus is 0.
- Similarly, If $\frac{p(n)}{t(n)} \rightarrow 0$, then $p(n) < \frac{r}{n-1}$, and $\lim_n (1 - \frac{r}{n-1})^{n-1} = 1$.
- Threshold function is not unique. For example, any function $\frac{a}{n+b}$ is a threshold function.

- For Poisson random network, there is much that is known:
- At threshold $\frac{1}{n^2}$, the first links emerge.
- Once $p(n)$ is at least $\frac{1}{n^{3/2}}$ there is a probability converging to one that the network has at least one component with at least three nodes.
- At $\frac{1}{n}$, we see cycles emerge.
- The giant component grows in size until $\frac{\log n}{n}$, where the network becomes connected.

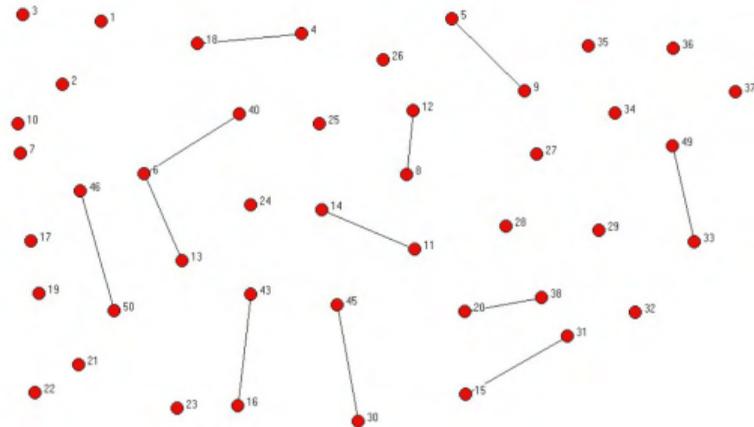


Figure: [Jackson et al., 2008] First Component with More than Two Nodes: 50 nodes, $p=0.01$

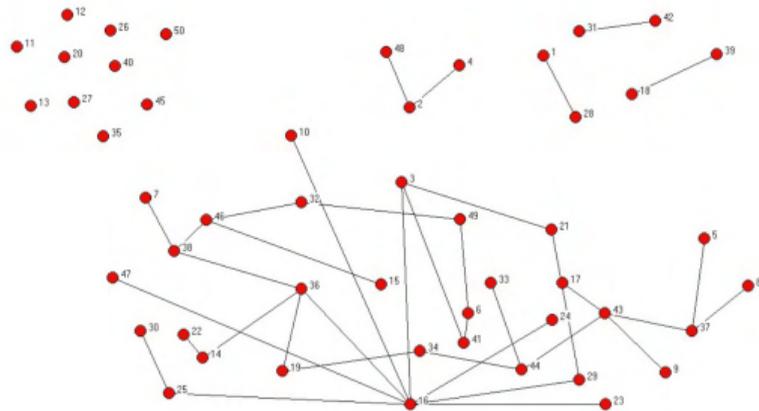


Figure: [Jackson et al., 2008] The Emergence of Cycles: A Random Network on 50 Nodes with $p=0.03$

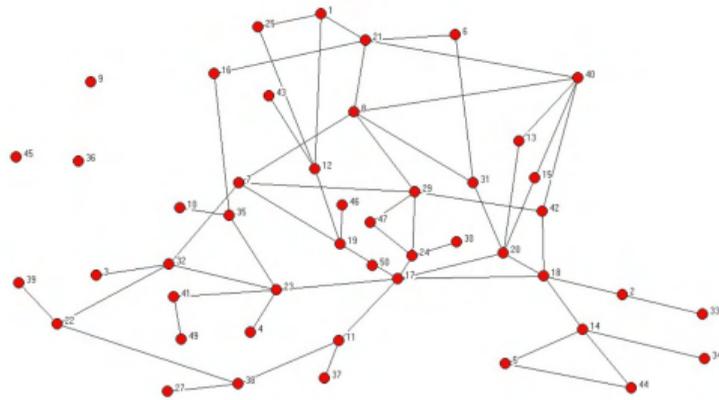


Figure: [Jackson et al., 2008] The Giant Component: A Random Network on 50 Nodes with $p=0.05$

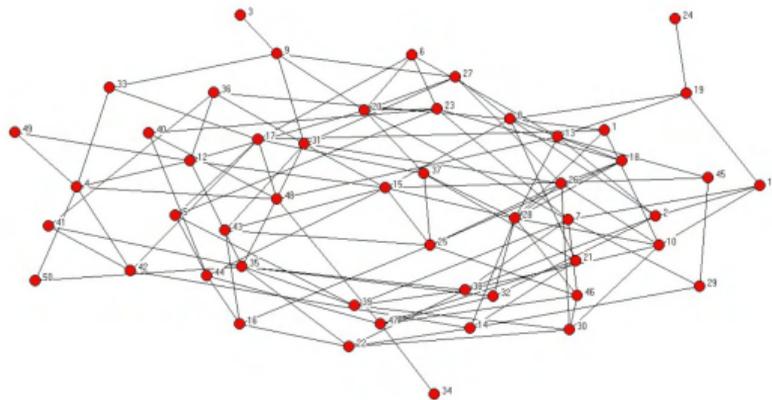
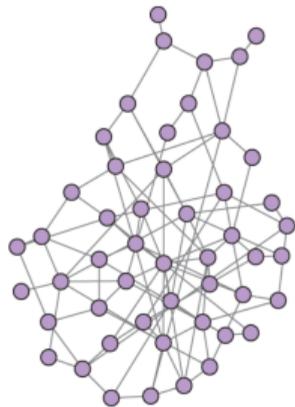


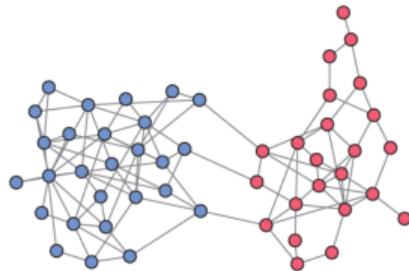
Figure: [Jackson et al., 2008] Emergence of Connectedness: A Random Network on 50 Nodes with $p=.10$

- Homophily: the tendency of individuals to form links with others of their own type.
- In the Erdős–Rényi model, the probability of linking is the same between every pair of individuals.
- We now extend the basic model to illustrate how it can accommodate homophily.
- Suppose that there are M groups, and that the probability of a link between two individuals within a group p_s is different from the probability of a link between two individuals in different groups, p_d , and $p_s > p_d$.
- These different probabilities define a random graph that is referred to as the stochastic block model.

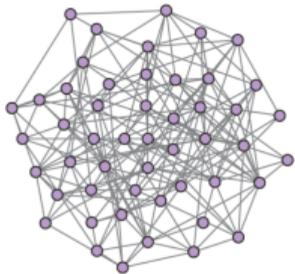
Homophily



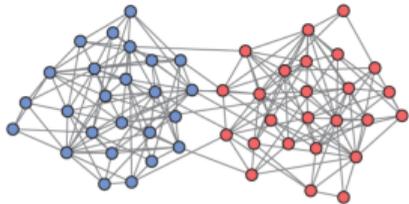
(a) Erdős-Rényi: $p = 0.08$



(b) $p_s = 0.15; p_d = 0.03$



(c) Erdős-Rényi: $p = 0.16$



(d) $p_s = 0.30; p_d = 0.01$

Exponential Random Graph Model

- It is not clear how previous model can be linked to data.
- “A good [statistical network graph] model needs to be both estimable from data and a reasonable representation of that data, to be theoretically plausible about the type of effects that might have produced the network, and to be amenable to examining which competing effects might be the best explanation of the data.”
- Exponential-family random graph models (ERGMs) are a general class of models based in exponential-family theory for specifying the probability distribution for a set of random graphs or networks.
- It facilitates the adaptation and extension of well-established statistical principles and methods for the construction, fitting, and comparison of models.
- We can also include covariates representing features like homophily, mutuality, triad effects, and a wide range of other structural features of interest.

Exponential Random Graph Model

- A random vector Z is said to belong to an exponential family if its probability mass function may be expressed in the form

$$P_{\theta}(Z = z) = \exp\{\theta^T g(z) - \psi(\theta)\}$$

where g is a function of z and ψ is a normalization term.

- This family includes many familiar distributions, such as the binomial, geometric, Poisson, Gaussian and chi-square distributions.
- Consider $G = (V, E)$ as a random graph, and $Y = [Y_{ij}]$ is the adjacency matrix for G . An exponential random graph model specifies the joint distribution of the elements in Y :

$$P_{\theta}(Y = y) = \frac{1}{\kappa} \exp\left\{\sum_H \theta_H g_H(y)\right\}$$

- each H is a configuration; $g_H(y) = \prod_{y_{ij} \in H} y_{ij}$, and is therefore either one if the configuration H occurs in y , or zero, otherwise
- non-zero θ_H means that Y_{ij} are dependent for all pairs of vertices; κ is a normalization constant.

Exponential Random Graph Model

- For example, in the Bernoulli Random Graphs, the presence or absence of an edge between that pair is independent of the status of possible edges between any other pairs of vertices.
- That is, for each pair i, j , Y_{ij} is independent of $Y_{i'j'}$. This assumption implies that $\theta_H = 0$ for all configuration H involving three or more vertices.
- The only relevant functions g_H are those of the form $g_H(y) = g_{ij}(y) = y_{ij}$. Then ERGM reduces to

$$P_{\theta}(Y = y) = \left(\frac{1}{\kappa}\right) \exp\left\{\sum_{i,j} \theta_{ij} y_{ij}\right\}$$

- This is another way write $p_{ij} = \frac{\exp(\theta_{ij})}{1 + \exp(\theta_{ij})}$.
- Note, there are N_V^2 parameters; therefore, it is common to impose an assumption of homogeneity across certain vertex pairs.

Exponential Random Graph Model

- For example, assuming homogeneity across all of G , $\theta_{ij} := \theta$:

$$P_{\theta}(Y = y) = \frac{1}{\kappa} \exp\{\theta L(y)\}$$

where $L(y) = \sum_{i,j} y_{ij} = N_e$ is the number of edges in the graph.

- Alternatively, suppose that vertices are known a priori to fall within either of two sets, say S_1 and S_2 . If we impose homogeneity within and between sets, we arrive at a model of the form

$$P_{\theta}(Y = y) = \frac{1}{\kappa} \exp\{\theta_{11}L_{11}(y) + \theta_{12}L_{12}(y) + \theta_{22}L_{22}(y)\}$$

where $L_{11}(y)$ and $L_{22}(y)$ are the number of edges within sets S_1 and S_2 , respectively, and $L_{12}(y)$ is the number of edges between S_1 and S_2 .

- With ERGM, it is also straightforward to include additional information, such as actor attributes in a social network, defined by covariate X .

Exponential Random Graph Model

- Exponential family models are generally fit using the method of maximum likelihood.
- Although MLE are well-defined, their calculation is non-trivial.
- An appropriate asymptotic theory for confidence intervals and testing, taking into account the highly dependent nature of observations in a network graph, has yet to be established.
- Therefore, this is an open area deserve investing.
- Also, people often use Bayesian methods on ERGM.

Problems

- The Erdős–Rényi random graph model is probably the most widely studied model of networks.
- The reason for its popularity is that it is easy to present and provides insights into the most fundamental questions concerning networks: the determinants of the degree distribution, the connectivity, and the diameter of the graph.
- A major attraction of this model is that the methods of analysis are transparent and prove useful when we go beyond the basic model and study variations.
- However, from an empirical point of view it has some serious weaknesses. For large graphs, the network will display negligible clustering.
- This is a general feature of social networks: they exhibit very large clustering relative to what would arise in the Erdős–Rényi network with a similar mean degree.
- Since link probability is independent across pairs of nodes, the clustering will be of the order of probability of linking and this probability gets close to zero in large graphs

Coauthorship network in economics: 1970–2010

Decade	1970s	1980s	1990s	2000s
Total authors	32,936	46,181	82,135	151,953
Average degree	0.894	1.268	1.617	1.951
Standard deviation of degree	1.358	1.793	2.204	2.539
Size of giant component	4,962	13,134	30,689	67,158
—as percentage	0.15	0.28	0.37	0.44
Clustering coefficient	0.19	0.18	0.17	0.17
Average distance	12.39	10.83	10.00	9.81

Source: www.aea.org/econlit/; Goyal, van der Leij, and Moraga-González (2006).

Real Networks

- Another critical difference is degree distribution.
- In previous table, the average degree is much smaller than the variance.
- However, in a Poisson random graph, the variance is equal to the mean.

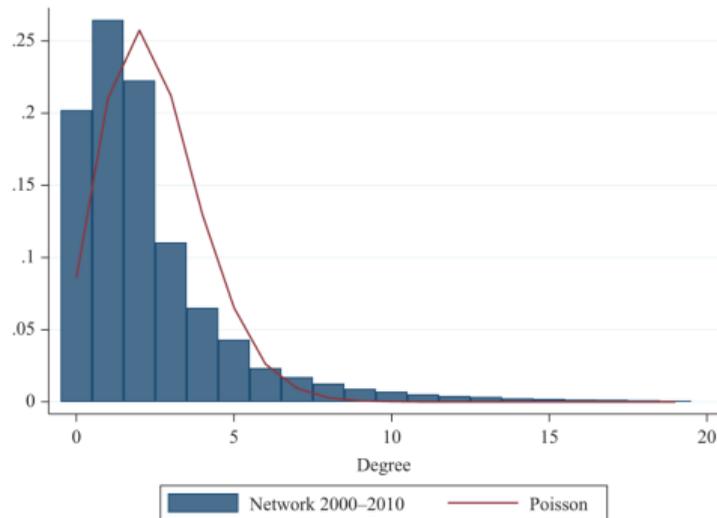
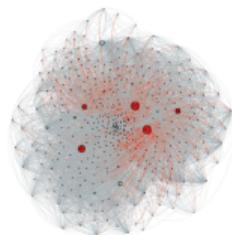


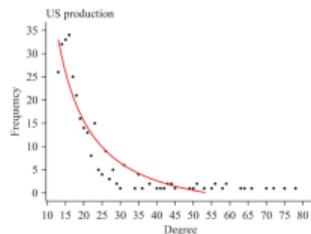
Figure: [Goyal, 2023]

Preferential Attachment

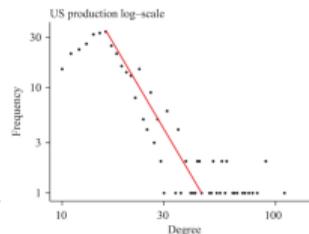
- These differences between the Poisson graph and empirical networks motivate the study of alternative models of networks.
- This is a growing random-network models that new nodes are born over time and form attachments to existing nodes.
- Preferential Attachment mechanism are designed to embody the principle that "the rich get richer."



(a) US Production Network 2002



(bi) Out-degrees of US economy 2002



(bii) Degrees on log scale

Preferential Attachment

- Let us look at Barabási and Albert (1999)'s model of undirected linking.
- Suppose that nodes are born overtime and indexed by their date of birth $i \in \{1, 2, \dots, t, \dots\}$.
- Upon birth each new node forms m links with pre-existing nodes.
- It attaches to nodes with probabilities proportional to their degrees relative to the the total degrees.
- Let $d_i(t)$ be the links of node i at time t . Then the probability that an existing node i gets a new link from the newborn node at time t is

$$m \frac{d_i(t)}{\sum_{j=1}^t d_j(t)}$$

- Since m links are created by every i , it follows that at time t , there are mt links and $2mt$ degrees: $\sum_{j=1}^t d_j(t) = 2tm$ (each link connects two nodes.)

Preferential Attachment

- Therefore, the probability may be written as $\frac{d_i(t)}{2t}$.
- In principle, this is a stochastic process; we will examine the deterministic approximation in which the rate of change of degree is equated to this probability of change in degree.
- With this in mind, we write the rate of change of degree as

$$\frac{d}{dt}d_i(t) = \frac{d_i(t)}{2t}$$

- This is a differential equation. Suppose the initial condition is $d_i(i) = m$.
- The solution is $d_i(t) = m\left(\frac{t}{i}\right)^{1/2}$
- We can use the solution to explicitly derive the long-run degree distribution.

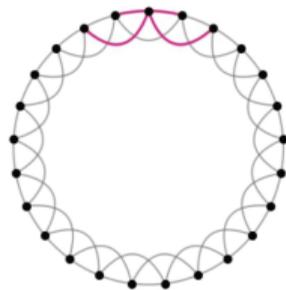
Preferential Attachment

- The degrees of nodes can be ordered by their age.
- To find the fraction of nodes with degrees that exceed some given level d at some time t , we just need to identify which node is exactly has degree d at time t .
- Denote $i_t(d)$ be such node: $d_{i_t(d)}(t) = d$.
- We can solve $\frac{i_t(d)}{t} = (\frac{m}{d})^2$.
- The nodes with a degree greater than d at time t are simply the nodes that were born before $i_t(d)$.
- This means the fraction of nodes with a degree greater than d , $P(D > d)$ is $(\frac{m}{d})^2$.
- Thus the distribution function is $F(d) = 1 - (\frac{m}{d})^2$.
- The density is $f(d) = 2m^2d^{-3}$.
- Thus, it provides a mechanism that can account for an empirically observed power law degree distribution.

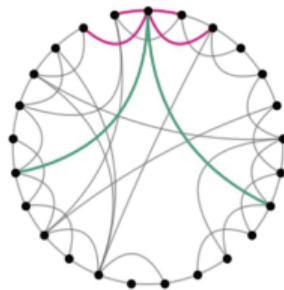
Small World

- In the real world network, we observe small average distance and high clustering coefficient.
- The Poisson random graph model and the preferential attachment model both generate small distances, but they exhibit negligible clustering.
- How can we reconcile high clustering and small distance?
- In a celebrated paper, Watts and Strogatz (1998) proposed a resolution to this tension with the help of the following simple model.
- Their approach has an initial network of n nodes arranged around a cycle, which are connected to their nearest 2 neighbors on either side.
- Observe that as n grows, the diameter ($n/4$) will grow too.
- How can we contain the growth of the mean distance as n grows?

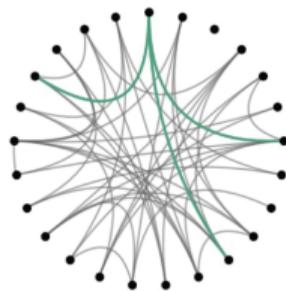
Small World



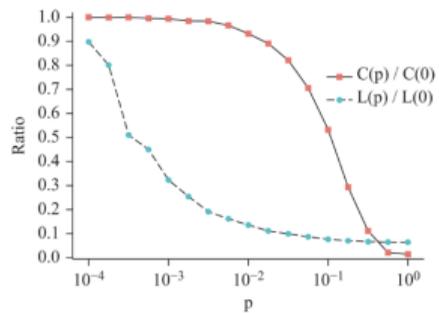
(a) Original local interaction network



(b) Small rewiring



(c) Complete rewiring



(d) Clustering versus distance

- The key idea is the “rewiring” of links: pick a link (A, B) , with a very small probability p , fix on side to node A , and then pick a new partner at random.
- For low and modest values of p , the average distance falls very sharply, while the clustering remains high and stable.
- Figure (d) shows how length and clustering coefficient changes with p .
- In the structure of small worlds, starting with a sparse graph on a cycle, the clustering remains stable for a broad range of rewiring probabilities, while the diameter comes down sharply with a small probability of rewiring.

Network-Based Linking

- The preferential attachment model delivers skewed degree distributions but fails to account for clustering;
- While the small-world model provides an account for clustering but exhibits relatively similar degrees.
- We now present a model of a growing network that combines features of preferential attachment with an additional feature—links are formed with neighbors of nodes found at random.
- This model generates networks with skewed degree distribution, as well as significant clustering levels.

Network-Based Linking

- Let us suppose that time proceeds in discrete steps ($t = 1, 2, 3, \dots$), and at each point $t > 1$, a new node enters.
- N_t is the set of nodes at time t .
- There is a contacting process followed by a linking process.
- At birth, a node picks randomly, and without replacement, m_r nodes from set $N_t - 1$ and forms links to them.
- She then picks m_n nodes randomly, without replacement, from the neighbors of the m_r nodes picked at random.
- Thus we can say that $m = m_r + m_n$ is the number of outward links formed by every entering new node.

Network-Based Linking

- To make sure that the process is well defined, let us suppose that there are enough nodes and links at the start.
- Suppose that at the start of time t , node i has $d_i(t)$ incoming links.
- Therefore, we may write the expected number of new links for node i as

$$\frac{m_r}{t} + \frac{m_n d_i(t)}{M_t}$$

where $M_t = mt$ is the total number of links in the network at time t .

- Thus the probability of getting a new link is increasing in the number of existing links. This is the preferential attachment aspect of this model.

Network-Based Linking

- Using the deterministic approximation, we may write the rate of change in links as

$$\frac{d}{dt}d_i(t) = \frac{m_r}{t} + \frac{m_n d_i(t)}{mt}$$

- Given the initial condition $d_i(i) = d_0 \geq 0$, the solution of the in-degree of node i at time $t \geq i$ is

$$d_i(t) = (d_0 + rm)\left(\frac{t}{i}\right)^{1/(1+r)} - rm$$

where $r = \frac{m_r}{m_n}$ as the ratio of random to network-based links.

- We use this formula to develop the degree distribution of the network.

$$F_t(d) = 1 - \left(\frac{d_0 + rm}{d + rm}\right)^{1+r}$$

- If r is small, almost all links are network-based, then it behaves like preferential attachment: $P(D > d) \propto d^{-1}$ and $p(d) \propto d^{-2}$.

References



Bishop, C. M. and Bishop, H. (2023).
Deep learning: Foundations and concepts.
Springer Nature.



Eric, D. and Kolaczy, K. (2009).
Statistical analysis of network data: Methods and models.



Goyal, S. (2023).
Networks: An economics approach.
MIT Press.



Jackson, M. O. et al. (2008).
Social and economic networks, volume 3.
Princeton university press Princeton.



Prince, S. J. (2023).
Understanding deep learning.
MIT press.



Sargent, T. J. and Stachurski, J. (2024).
Economic networks: Theory and computation, volume 53.