Case Study: Network

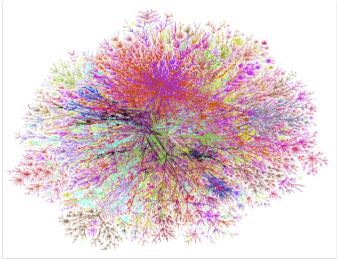
BIOSTAT830: Graphical Models

December 13, 2016

Network Fundamentals

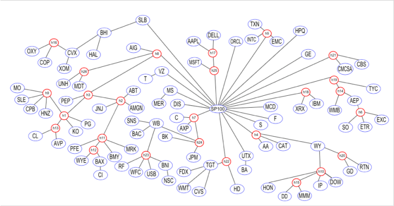
- One of many classifications:
  - Techonological networks (e.g.,)
  - Social networks (e.g., Twitter, Facebook, WeChat)
  - Information networks (e.g., World Wide Web)
  - Biological networks (e.g., gene regulation network, human brain functional connnection network, contact network epidemiology)

Examples of Networks



Internet: Bill Cheswick
http://www.cheswick.com/ches/map/gallery/index.html

Airline Network: Northwest Airlines WorldTraveler Magazine



Anandkumar and Valluvan (2013) Annals of Statistics. Figure: Tree graph learned on S&P100 monthly stock return data

New York City Subway. http://web.mta.info/maps/submap.html

General Themes:

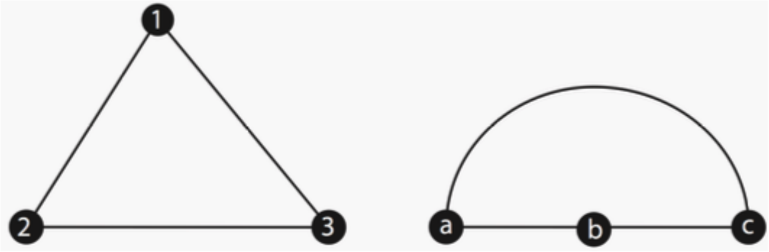
- Formulate mathematical models for network patterns, phenomena and principles
- Reason about the model's broader implications about networks, e.g., behavior, population-level dynamics, etc.
- Develop common analytic tools for network data obtained from a variety of settings

Basics

- Network is a graph
- Graphs
  - Mathematical models of network structure
  - Graph: Vertices/Nodes+Edges/Ties/Links
  - A way of specifying relationships among a collection of items

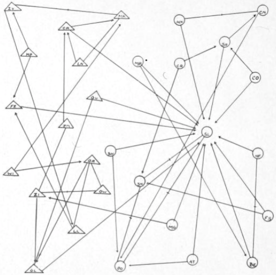
- Graph: Ordered pair $G = (V, E)$
- $V(G)$: vertex set; $E(G)$: edge set
- The vertex pairs may be ordered or unordered, corresponding to directed and undirected graphs
- Some vertex pairs are connected by an edge, some are not
- Two connected vertices are said to be (nearest) neighbors

- Two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ are equal if they have equal vertex sets and equal edge sets, i.e., if $V_1 = V_2$ and $E_1 = E_2$ (Note: equality of graph is defined in terms of equality of sets)
- Two graph diagrams (visualizations) are equal if they represent equal vertex sets and equal edge sets



- Consider a subset of vertices $V'(G) \subset V(G)$
- An **induced subgraph** of G is a subgraph $G' = (V', E')$ where $E(G') \subset E(G)$ is the collection of edges to be found in G among the subset $V(G')$ of vertices
- For examlple, consider Moreno's sociogram. If V' denotes the boys' vertices, what is the graph G' induced by V'?

EVOLUTION OF GROUPS



Class Structure, 2nd Grade

14 boys and 14 girls. *Unchosen*, 9, WI, KP, MG, AT, FS, CN, CR, MR, SH; *Pairs*, 11, ZV-MK, MK-LN, OW-ZI, GR-LL, ZI-JM, HN-CM, SL-JN, JN-PO, PO-SL, HF-BE, GL-GU; *Stars*, 2, SL, PO; *Chains*, 0; *Triangles*, 1, SL-JN-PO; *Inter-sexual Attractions*, 5.

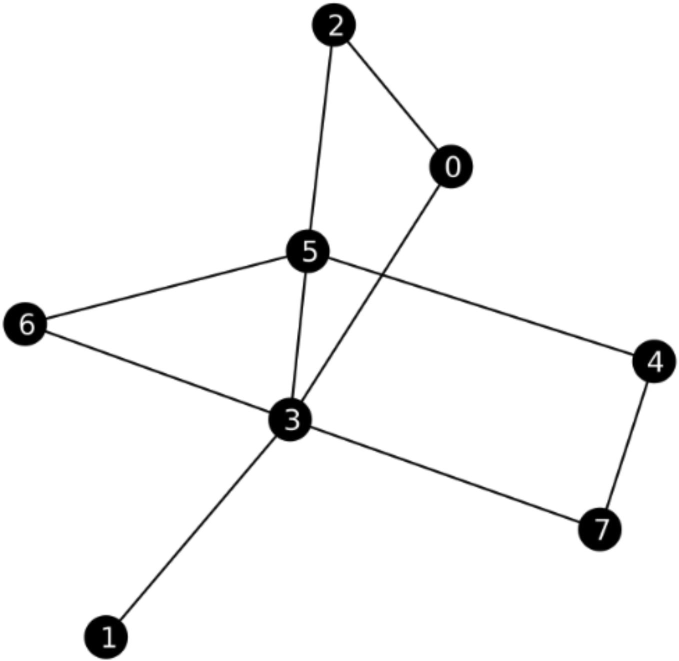
- Edges, depending on context, can signify a variety of things
- Common interpretations
  - Structural connections
  - Interactions
  - Relationships
  - Dependencies
- Often more than one interpretation may be appropriate

Local structure of networks, directed or undirected, can be summarized by **subgraph censuses**; Network motif discovery - A dyad is a subgraph of two nodes - Dyad census: count of all (3) isomorphic subgraphs - A triad is a subgraph of three nodes - Traid census: count of all (16) isomorphic subgraphs

- The **degree** of a node in a graph is the number of edges connected to it
- We use d_i to denote the degree of node i
- M edges, then there are $2M$ ends of edges; Also the sum of degrees of all the nodes in the graph: $\sum_i d_i = 2M$
- Nodes in directed graph have **in-degree** and **out-degree**

- A **walk** in a graph is a sequence $(v_1, v_2, v_3, ..., v_{n-1}, v_n)$ of not necessarily distinct vertices in which v_1 is joined by an edge to v_2, v_2 is joined by an edge to v_3, ..., v_{n-1} is joined by an edge to v_n
- A walk is sometimes presented as an alternating sequence of vertices and edges, such that every edge joins the vertices immediately preceding and following it
- A walk$(v_1, v_2, v_3, ..., v_{n-1}, v_n)$ in a graph is a closed walk if v_1 and v_n are the same vertex; otherwise it is an open walk
- A **path** is a walk without repeated vertices
- A **trail** is a walk without repeated edges
- Every path is a trail, but not every trail is a path

- A vertex v in a graph is **reachable** from another vertex u if there exists a path from u to v
- A graph is **connected** if every vertex is reachable from every other vertex
- If a graph is not connected it is **disconnected**
- There is often no a priori reason to expect a graph to be connected
- The **length of a path** is the number of edges in the sequence that comprises it
- The **(geodesic) distance** between two nodes is the length of the shortest (geodesic) path between them
- The **diameter of a graph** is the longest of all pairwise shortest paths in a graph


2
0
5
6
4
3
7
1

Link Density

- Consider an undirected network with N nodes
- How many edges can the network have at most?
  - The number of ways of choosing 2 vertices out of N: $N(N-1)/2$
- A graph is fully connected if every possible edge is present

- Let M be the number of edges
- **Link density**: the fraction of edges present, and is denoted by ρ

$$\rho = \frac{2M}{N(N-1)}$$

- Link density lies in $[0, 1]$
- Most real networks have very low ρ
- Dense network: $\rho \to \textit{constant}$ as $N \to \infty$
- Sparse network: $\rho \to 0$ as $N \to \infty$

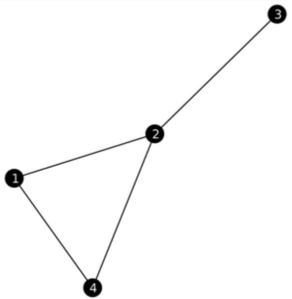
- An **adjacency matrix** is an $N \times N$ matrix $\mathbf{A}$ where A_{ij} encodes information about the edge between nodes i and j

$$\text{e.g. } \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

- **Weighted networks** have weights, covariates, or strength associated with the ties

$$\mathbf{A} = \begin{bmatrix} 0 & .5 & 0 & 2 \\ .5 & 0 & 9 & 3 \\ 0 & 9 & 0 & 0 \\ 2 & 3 & 0 & 0 \end{bmatrix}$$

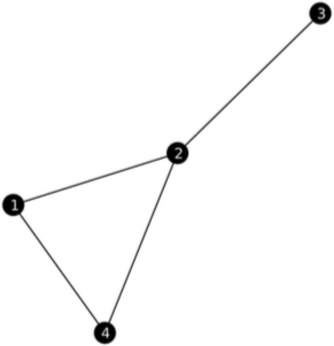
- The paths of length 2 are given by $\mathbf{A}^2$:



$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \mathbf{A}^2 = \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & 3 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix}$$

- The paths of length r are given by $\boldsymbol{A}^r$

- The shortest path between i and j is the **geodesic path**.
- Its length is the smallest r such that $[\mathbf{A}^r]_{i,j} > 0$
- What is the diameter of this network?



Network Descriptors

- **Centrality**: measures hwo central or important nodes are in the network
- Proposing new centrality measures and developing algorithms to calculate them is an active field of research
- **Degree centrality** is just another name for degree; Simplest centrality measure

Eigen-Centrality

- **Eigenvector centrality** gives more centrality to nodes whose neighbors are themselves more central: it's more important to be connected to influential neighbors than isolated ones.
- Specifically, each node's centrality score is proportional to the sum of its neighbors' centrality score.

$$\mathbf{Ac} = \kappa\mathbf{c} \Longrightarrow c_i = \frac{1}{\kappa}\sum_{j=1}^{N} A_{ij}x_j$$

Closeness Centrality

- **Closeness centrality** is based on a node's average distance to every other node.

$$l_i = \frac{1}{N}\sum_{j=1}^{N} d_{ij}$$

- This is small for nodes that are highly connected, so centrality is the inverse:

$$c_i = \frac{N}{\sum_{j=1}^{N} d_{ij}}$$

- Problem: this measure usually has a small range and is highly sensitive to small changes in the network; it is 0 whenever a network has multiple components.
- Alternative:

$$c_i = \frac{1}{N-1}\sum_{j \neq i} \frac{1}{d_{ij}}$$

Clustering

- We often want to know how densely the neighbors of a given node are connected.
- Consider a node i with degree k_i.
- Let t_i be the number of ties among the neighbors of i.
- The local clustering coefficient is defined as the number of ties that exist between the neighbors of i, divided by the number of ties that could exist between them, $k_i(k_i - 1)/2$
- This gives rise to $l_i = \frac{2t_i}{k_i(k_i - 1)}$
- The mean local clustering coefficient in a network is computed by taking the mean of l_i over all nodes in the network.

Clustering: Transitivity

- A relation $\circ$ is transitive if $a \circ b$ and $b \circ c$ together imply $a \circ c$.
- In a network, there are various relations between pairs of vertices, the simplest one being "is connected by an edge".
- If the "connected by an edge" relationship were transitive, it would mean that if vertex u is connected to vertex v, and v is connected to w, then u is also connected to w
  - "triangle closure" or "triadic closure"
- Networks showing this property are said to be **transitive.**
- Perfect transitivity implies a fully connected graph (not a very useful concept).
- In practice, many networks exhibit partial transitivity, and this is true especially for social networks: the friend of my friend is far more likely to be my friend than some randomly chosen member of the population.

Clustering: Transitivity

- We can quantify the extent of transitivity by considering paths and loops consisting of three nodes u, v, and w.
- If u knows v and v knows w, then we have a path (u, v, w) of two edges.
- If, in addition, u also knows w, then we have a loop (closed path) of 3 vertices and 3 edges.
- A **closed triad** is a set of three vertices u, v, w with edges (u, v), (v, w), and (u, w).
- A **connected triple** is a set of three vertices u, v, w with edges (u, v) and (v, w), where the edge (u, w) may or may not be present.

Clustering: Transitivity

- The **global clustering coefficient** is:

$$L = \frac{3 \times (\text{number of closed triads})}{(\text{number of connected triples})}$$

- Between 0 and 1 because every closed triad contributes 3 connected triples.
- Sometimes referred to as the "fraction of transitive triples."
- Measures how transitive a network is.

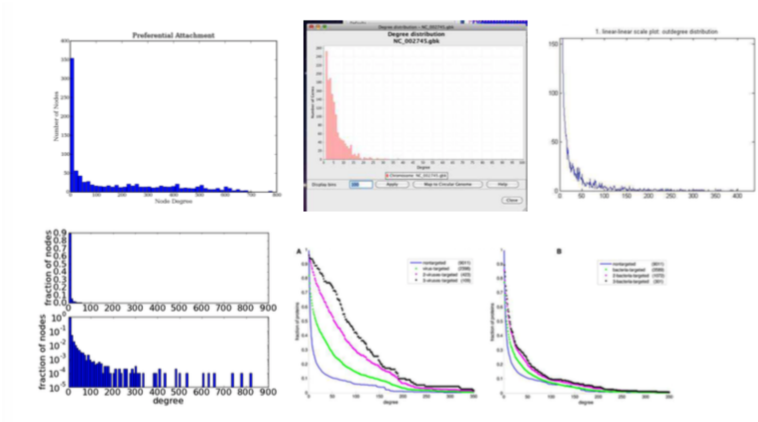
Degree Distribution

- One of the most fundamental properties of a network is the frequency of node degrees.
- Define p_d to be the fraction of nodes in the network with degree d.
- The quantities p_d for $d = 0, \dots, max$ give the **degree distribution** for the network.
- Almost all real-world networks have degree distributions that (approximately) follow a power-law distribution:

$$p_d = \beta k^{-\alpha}$$

- Networks with power-law degree distributions are called **scale-free** networks.

Degree Distribution



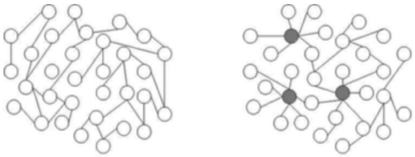
Degree Distribution

Why are these power laws so common?

- **Preferential attachment model**: the probability for a new node to connect to existing node i depends on d_i
  - "rich get richer"
- **Fitness model**: nodes compete for ties, nodes that are more "fit" for this competition increase their degree faster than nodes with less fitness.

Degree Distribution

- The long tail of the degree distribution means that there are many outliers with very high degree – **hubs**
- In general scale-free networks have a hierarchical structure: big hubs are connected to smaller hubs who are connected to the many nodes with very small degree.
- Low-degree nodes are connected to one another in dense subgraphs that are connected to each other through hubs



(a) Random network (b) Scale-free network

Small-world Phenomenon

- The **small-world phenomenon** refers to the surprising finding that the world looks "small" when you think of how to get from you to almost anyone else.
  - In the mathematical co-authorship network, Erdos is probably the biggest, most central hub.
- The average geodesic distance between two nodes in a network tends to be small.
- This follows logically from the organization of scale-free networks into hierarchical hubs: you can probably get to any other node in the network through the closest big hub.
- Captured by the notion of six degrees of separation, which comes from a play of this title by John Guare:
  - One of the characters of Guare's play utters the following line: "I read somewhere that everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everyone else on this planet."
- But how do we know that we live in a small world?

Small-world Phenomenon

- Stanley Milgram and his colleagues performed the first experimental study of this notion in the 1960s.
- Attempted to test the speculative idea that people are connected in the global friendship network by short chains of friends.
- A group of 296 randomly chosen subjects were asked to try forwarding a letter to a target person, a stockbroker living in a Boston suburb.
- The subjects were given some personal information about the target (including his address and occupation).
- The subjects were then asked to forward the letter to someone whom they knew on a first-name basis, with the same instructions, to eventually reach the target as quickly as possible.

Small-world Phenomenon

- Each letter passed through a sequence of (first-name basis) friends in succession.
- A total of 64 chains succeeded in reaching the target.
- The median chain length was six, which is the number that made its way to Guare's play two decades later.
- There are (at least) two remarkable aspects to this study • The high fraction of completed chains (64/296).
- The short length of the chains.
- Although there are a few caveats to this experiment, it is now accepted that social networks have very short paths between essentially arbitrary pairs of people.
- These short paths have substantial consequences for the potential speed with which information, pathogens, memes, behaviors, etc. spread through society.

Active Methods Research Area: Peer/Contagion Effects

- Is obesity contagious? (Christakis and Fowler, 2007, NEJM)
- Cooperative behaviour in social network (Fowler and Christakis, 2010, PNAS)
- Contact network epidemiology for studying population dynamics of infectious disease dynamics

Implication of Contagion upon Intervention

- Vaccination
  - Percolation theory: originates in statistical physics and mathematics where it is used to mainly study low-dimensional lattices, or regular networks
  - In network context, percolation referes to the process of removing nodes or edges from the network
  - Site versus bond percolation
  - “removal” referes to the elements (nodes or edges) being somehow non-functional - they are not removed from the system
  - Think of percolation as a process that switches nodes or edges either on or off

Percolation

- Percolation can be used to study the failure of routers on the Internet.
  - At any one time about 3% of routers (nodes) on the Internet are non-functional for some reason.
  - One can use percolation to study the impact of these types of failures on system performance.
- Percolation is also relevant for considering vaccination or immunization of individuals.
  - In a contact network individuals are represented by nodes and edges are potential conduits for pathogens.
  - Vaccination can be represented by removing vertices, in some cases leading to **herd immunity**.

- Here we focus on site percolation (node removal).
- Percolation process is parameterized by occupation probability ϕ.
- This is the probability that a vertex is present or functioning in the network (**occupied** in the terminology of percolation theory).
- If $\phi = 1$, all vertices in the network are occupied (functional).
- If $\phi = 0$, no vertices are occupied (all have been removed).

: Site percolation with occupation probability $\phi = 1$ (top left), $\phi = 2/3$ (top right), $\phi = 1/3$
m left), and $\phi = 0$ (bottom right).

Did not discuss today

- Generate a random network:
  1. Random graph models
  2. Erdos-Renyi (E-R) model, or E-R random graph named after Hungarian mathematicians; Also known as Poisson random graph (degree distribution of the model follows a Poisson)
  3. Barabasi-Albert model (preferential attachment)
  4. Small-world model/Watts-Strogatz model (high transitiity; small-world property)
  5. Exponential Random Graph Models (ERGM)
  6. Stochastic block models (community structure)

- Network Fundamentals
  1. Basics: Chapter 6; Descriptors: Chapter 7-8; Models: Chapter 12-15, Newman (2010). [Networks: An Introduction. Oxford University Press.]
- Social Networks:
  1. Chapter 3, Newman book.
  2. Hoff, Raftery and Handcock (2002). Latent Space Approaches to Social Network Analysis. *JASA*.
- Social Influence (Peer-Effects; Contagion):
  1. Christakis and Fowler (2007). The Spread of Obesity in a Large Social Network over 32 Years. NEJM.
  2. Responses to CF2007: Cohen-Cole and Fletcher (2008); Lyons (2011); Shalizi and Thomas (2011); and More
  3. O'Malley et al. (2014). Estimating Peer Effects in Longitudinal Dyadic Data Using Instrumental Variables. *Biometrics*.
- Infectious Disease Dynamics
  1. Chapter 21, Easley and Kleinberg (2010). [Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.]

- Notes partially sourced from Betsy Ogburn and JP Onella