MITx: Statistics, Computation & Applications

Criminal Networks Module

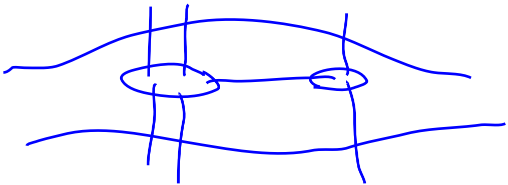
Lecture 1: Introduction to Networks

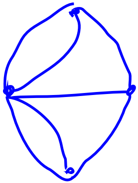
Network

A **network** (or **graph**) G is a collection of **nodes** (or **vertices**) V connected by **links** (or **edges**) E. The network is denoted by $G = (V, E)$.

Network research:

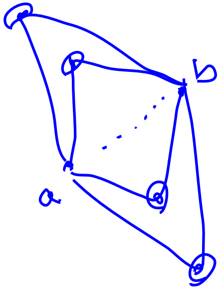
- Grew out of graph theory
  - e.g. Euler's celebrated 1735 solution of the Königsberg bridge problem





$$A \quad \begin{array}{c|cccc} & a & b & c & \cdots \\ \hline a & & 1 & 0 & \\ b & 1 & & & \\ c & 0 & & & \\ \vdots & & & & \end{array}$$

$E = \{\{a, b\}\}$



$$(A^m)_{ii} = 0$$

$$\hookrightarrow tr(A^m) = 0$$

$$(A^2)_{ij} = \sum_k \underline{A_{ik}}\, \underline{A_{kj}}$$

Representation of a network

Two common representations of a network $G = (V, E)$:

- **adjacency list**
  - undirected graph $1 - 2 - 3$: $E = \{\{1, 2\}, \{2, 3\}\}$
  - directed graph $1 \rightarrow 2 \leftarrow 3$: $E = \{(1, 2), (3, 2)\}$

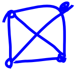
- **adjacency matrix** of size $n \times n$ (where $n = |V|$) with

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

  - For weighted graph, A_{ij} can be non-binary

How does the adjacency matrix of a simple graph look like? How to suggest new friends in a social network? And what about an acyclic graph?

$$\frac{\#\text{triangles}}{\binom{n}{3}} \qquad \frac{3 \cdot \#\text{triangles}}{\text{connected triples}} \qquad \frac{\#\text{edges}}{\binom{n}{2}} \begin{cases} \to 0 \\ \to c \end{cases}$$



MITx: Statistics, Computation & Applications

Criminal Networks Module

Lecture 1: Introduction to Networks

Network

A **network** (or **graph**) G is a collection of **nodes** (or **vertices**) V connected by **links** (or **edges**) E. The network is denoted by $G = (V, E)$.

Network research:

- Grew out of graph theory
  - e.g. Euler's celebrated 1735 solution of the Königsberg bridge problem
- In recent years network research witnessed a big change:
  - From study of a single graph on 10-100 nodes to the statistical properties of large networks on millions of nodes
  - Characterize the structure of networks
  - Identify important nodes / edges in a network
  - Develop network models
  - Predict behavior of network processes based on the network structure

Examples of networks

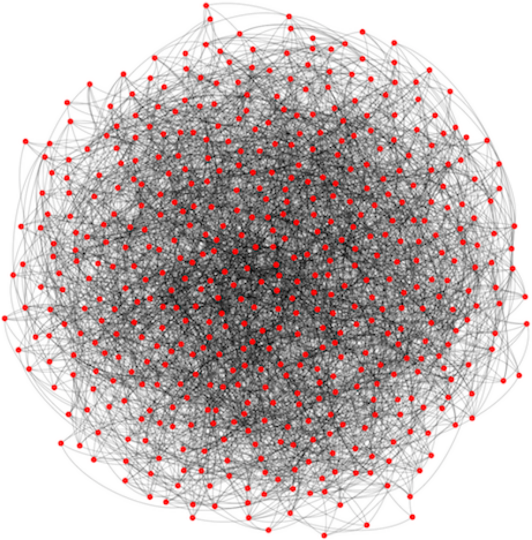
Network	Vertex	Edge
World Wide Web	web page	hyperlink
Internet	computer	network protocol interaction
power grid	generating station / substation	transmission line
friendship network	person	friendship
gene regulatory network	gene	regulatory effect
neural network	neuron	synapse
food web	species	who-eats-who
phylogenetic tree	species	evolution
Netflix	person / movie	rating

Different kinds of networks

- simple network: undirected network with at most one edge between any pair of vertices and no self-loops
  - e.g. Internet, power grid, telephone network
- multigraph: self-loops and multiple links between vertices possible
  - e.g. neural network, road network
- directed network: $(i, j) \in E$ does not imply $(j, i) \in E$
  - e.g. World Wide Web, food web, citation network
- weighted network: with edge weights or vertex attributes
- tree: graph with no cycles
  - e.g. phylogenetic tree (how to check that network is a tree?)
- acyclic network: graph with no directed cycles (how to check that network is acyclic?)
  - e.g. food web, citation network

- bipartite network: edges between but not within classes
  - e.g. recommender systems, Netflix
- hypergraph: generalized 'edges' for interaction between > 2 nodes
  - e.g. protein-protein interaction network

Large networks look like hairballs



Two common representations of a network $G = (V, E)$:

- **adjacency list**
  - undirected graph $1 - 2 - 3$: $E = \{\{1, 2\}, \{2, 3\}\}$
  - directed graph $1 \to 2 \leftarrow 3$: $E = \{(1, 2), (3, 2)\}$

- **adjacency matrix** of size $n \times n$ (where $n = |V|$) with

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{array}{cc} i & \multirow{2}{*}{$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$} \\ j & \end{array}$$

  - For weighted graph, A_{ij} can be non-binary

How does the adjacency matrix of a simple graph look like? How to suggest new friends in a social network? And what about an acyclic graph?

Quantitative measures of networks

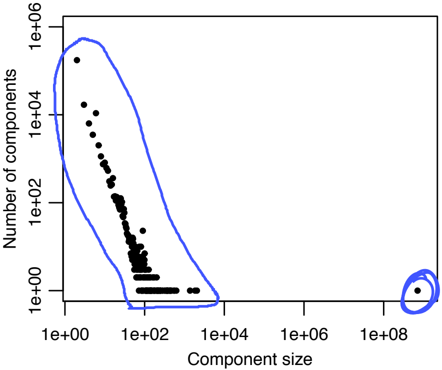
Some quantitative measures of networks to describe structural patterns of a network and to compare networks:

- connected components
- edge density
- degree distribution
- diameter and average path length
- clustering
- homophily or assortative mixing

Connected Components

Connected component: set of nodes that are reachable from one another

- Many networks consist of one large component and many small ones



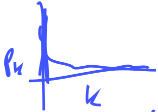
Component size distribution in the 2011 Facebook network on a log-log scale. Most vertices (99.91%) are in the largest component.

Degree distribution

- Degree of node i: k_i
- Average degree: $\frac{1}{n}\sum_i k_i = \frac{\sum_{i,j} A_{ij}}{n} = \frac{2m}{n}$, where $|V| = n$, $|E| = m$
- More information captured by degree distribution
  - histogram of fraction of nodes with degree k.

Degree distribution

- Degree of node i: k_i
- Average degree: $\frac{1}{n}\sum_i k_i = \frac{\sum_{i,j} A_{ij}}{n} = \frac{2m}{n}$, where $|V| = n$, $|E| = m$
- More information captured by degree distribution
  - histogram of fraction of nodes with degree k.



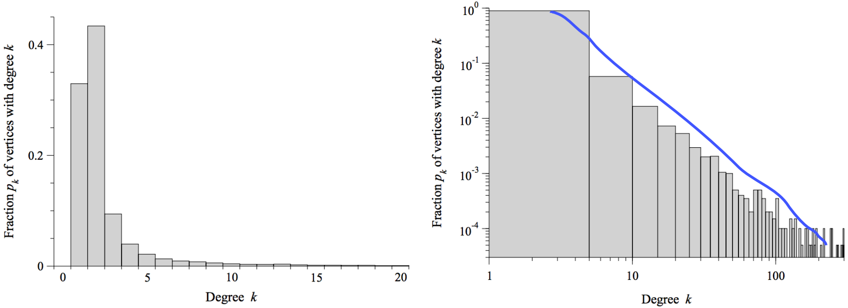
- Special type of degree distribution: power-law distribution:

$$\log p_k = -\alpha \log k + c \quad \text{for some } \alpha, c > 0$$

$p_k = k^{-\alpha} \cdot c$

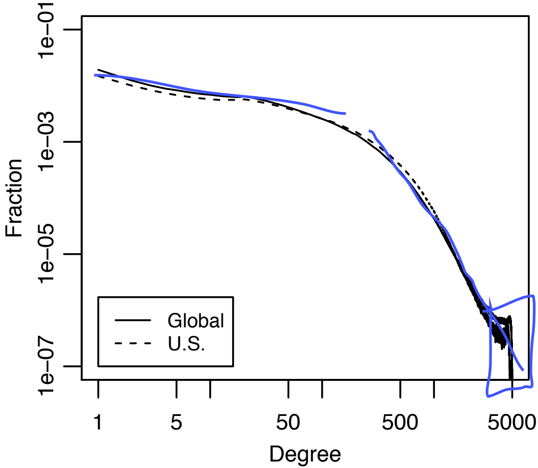
  - tail of distribution is fat, i.e., there are many nodes with high degrees
  - appears linear on a log-log plot
  - appear in wide variety of settings including WWW, Internet

Degree distribution of the Internet



Figures from Chapter 8 in "Networks: An Introduction" by M.E.J. Newman (2010)

Degree distribution of Facebook network



From "The Anatomy of the Facebook Social Graph" by Ugander et al. (2011)

Edge density

The edge density or connectence is defined as

$$\rho = \frac{m}{\binom{n}{2}} = \frac{\sum_{i,j} A_{ij}}{n(n-1)}, \quad \text{where } |V| = n,\ |E| = m$$

- Different kinds of networks show very different edge densities

Edge density

The edge density or connectence is defined as

$$\rho = \frac{m}{\binom{n}{2}} = \frac{\sum_{i,j} A_{ij}}{n(n-1)}, \quad \text{where } |V| = n,\ |E| = m$$

- Different kinds of networks show very different edge densities
- Most networks are sparse, i.e., $\rho \overset{n\to\infty}{\longrightarrow} 0$ (the number of edges does not grow proportionally with the number of nodes)
  - E.g., friendship network: if each person has a constant number of friends c, then $\rho = \frac{cn}{\binom{n}{2}} \overset{n\to\infty}{\longrightarrow} 0$
- Some networks are dense, i.e., $\rho \overset{n\to\infty}{\longrightarrow} \text{const.}$
  - E.g., food web, when comparing ecosystems of different sizes

Diameter and average distance

- Let d_{ij} denote the length of the geodesic path (or shortest path) between node i and j
- The diameter of a network is the largest distance between any two nodes in the network:

$$\text{diameter} = \max_{i,j \in V} d_{ij}$$

- The average path length is the average distance between any two nodes in the network:

$$\text{average path length} = \frac{1}{\binom{n}{2}} \sum_{i \leq j} d_{ij}$$

Diameter and average distance

- Let d_{ij} denote the length of the geodesic path (or shortest path) between node i and j
- The diameter of a network is the largest distance between any two nodes in the network:

$$\text{diameter} = \max_{i,j \in V} d_{ij}$$

- The average path length is the average distance between any two nodes in the network:

$$\text{average path length} = \frac{1}{\binom{n}{2}} \sum_{i \leq j} d_{ij}$$

- If network is not connected, one often computes the diameter and the average path length in the largest component.

Diameter and average distance

- Let d_{ij} denote the length of the geodesic path (or shortest path) between node i and j
- The diameter of a network is the largest distance between any two nodes in the network:

$$\text{diameter} = \max_{i,j \in V} d_{ij}$$

- The average path length is the average distance between any two nodes in the network:

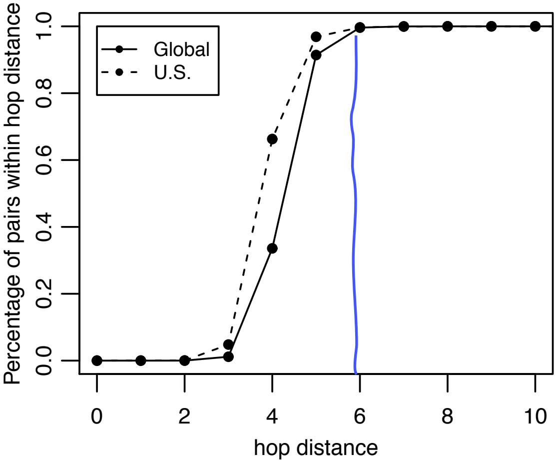
$$\text{average path length} = \frac{1}{\binom{n}{2}} \sum_{i \leq j} d_{ij}$$

- If network is not connected, one often computes the diameter and the average path length in the largest component.
- Algorithms for finding shortest paths: breadth-first search for unweighted graph, Dijkstra's algorithm for weighted graphs

Small-world and 6 degrees of separation

- Concept of 6 degrees of separation was made famous by sociologist Stanley Milgram and his study "The Small World Problem" (1967)
- In his experiment participants from a particular town were asked to get a letter to a particular person in a different town by passing it from acquaintance to acquaintance.
- 18 out of 96 letters made it in an average of 5.9 steps
- Any reasons why we should take the conclusion of 6 degrees of separation with a grain of salt?

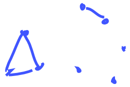
Diameter of Facebook (2011)



From "The Anatomy of the Facebook Social Graph" by Ugander et al. (2011)

Clustering

- In social networks: It is often the case that two nodes who share a common friend are also friends
- Triangle density: $\frac{\text{number of triangles}}{\binom{n}{3}}$



- Triangle density does not necessarily characterize clustering (why?)

Clustering

- In social networks: It is often the case that two nodes who share a common friend are also friends
- Triangle density: $\frac{\text{number of triangles}}{\binom{n}{3}}$
- Triangle density does not necessarily characterize clustering (why?)
- **Remedy:** clustering coefficient (or network transitivity)

$$C = \frac{3 \cdot \text{number of triangles in network}}{\text{number of connected triples}} \quad \in [0,1]$$

Clustering

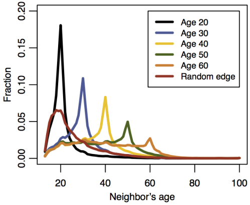
- In social networks: It is often the case that two nodes who share a common friend are also friends
- Triangle density: $\frac{\text{number of triangles}}{\binom{n}{3}}$
- Triangle density does not necessarily characterize clustering (why?)
- **Remedy:** clustering coefficient (or network transitivity)

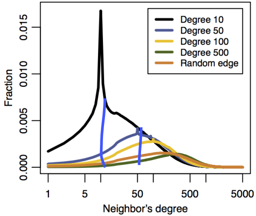
$$C = \frac{3 \cdot \text{number of triangles in network}}{\text{number of connected triples}} \quad \in [0, 1]$$

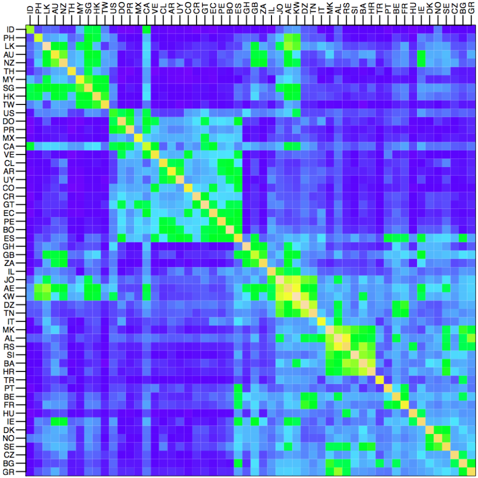
- Can also be defined node-wise: local clustering coefficient:

$$C_i = \frac{\text{number of triangles connected to node } i}{\text{number of triples centered at } i}$$

Homophily (or assortative mixing): tendency of people to associate with others that are similar







From "The Anatomy of the Facebook Social Graph" by Ugander et al. (2011)

Homophily

- How can we quantify homophily?

Homophily

- How can we quantify homophily?
- Measure the fraction of edges in the network that run between nodes of the same type
  - But this measure is 1 in a network where all nodes are of same type
  - Would like homophily measure to be 1 only in non-trivial setting

Homophily

- How can we quantify homophily?
- Measure the fraction of edges in the network that run between nodes of the same type
  - But this measure is 1 in a network where all nodes are of same type
  - Would like homophily measure to be 1 only in non-trivial setting
- **Remedy:** Fraction of edges that run between same type of nodes minus fraction of such edges if edges were placed at random
  - $\#$ edges of same type $= \sum_{(i,j)\in E} \delta(t_i, t_j) = \frac{1}{2}\sum_{i,j} A_{ij}\delta(t_i, t_j)$, where t_i is type of node i and $\delta(a, b) = 1$ if $a = b$ and 0 otherwise
  - expected $\#$ edges of same type $= \frac{1}{2}\sum_{i,j} \frac{k_i k_j}{2m}\delta(t_i, t_j)$
- Modularity: $\frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(t_i, t_j) \quad \in [-1, 1]$

References

- Chapters 6 - 10 (but mostly chapters 6 and 8) in

  M. E. J. Newman. *Networks: An Introduction*. 2010.

- For an analysis of the Facebook network:

  J. Ugander, B. Karrer, L. Backstrom and C. Marlow. *The Anatomy of the Facebook Social Graph*. 2011.