GL Applied Data Science Program

Network Analysis

August 25, 2021

Overview

**Overview of this week / module:**

- Data collection and visualization for exploratory data analysis

- Network analysis

- Unsupervised learning - clustering

Overview of this lecture:

- Examples of networks and representing networks

- Summary statistics of a network

- Centrality measures - finding important nodes in a network

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 $G = (V, E)$.
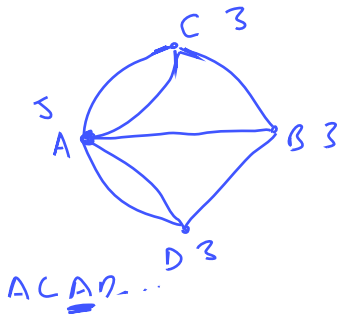

Movies

People
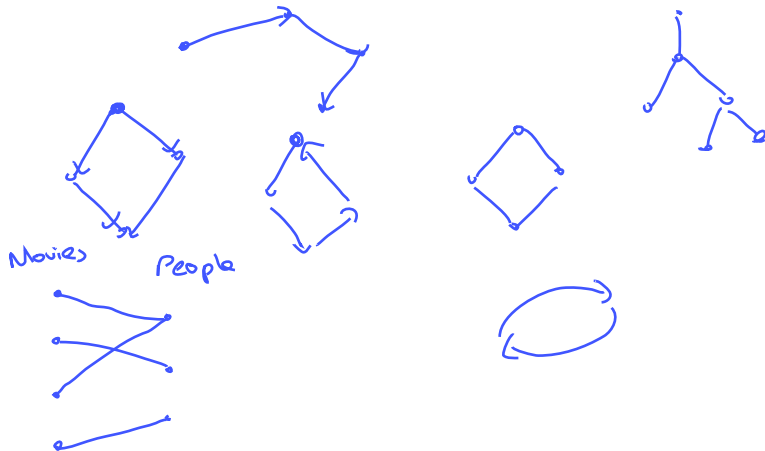
Network

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
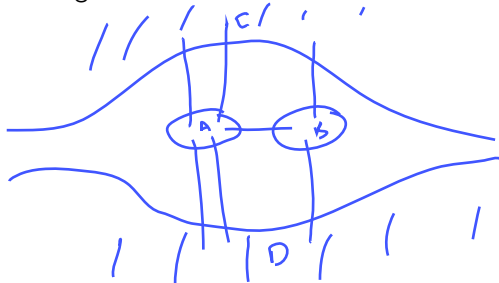 - e.g. Euler's celebrated 1735 solution of the Königsberg bridge problem

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
   - Identify missing links in a network

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
- **acyclic network**: graph with no directed cycles
  - 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
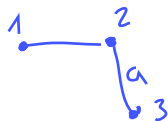
Large networks look like hairballs
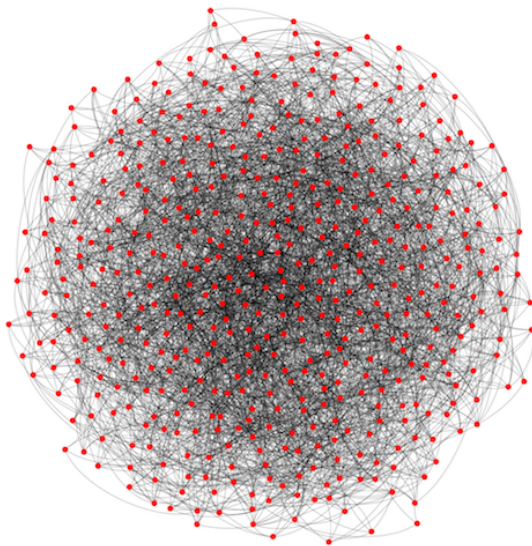
Representation of a network

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

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

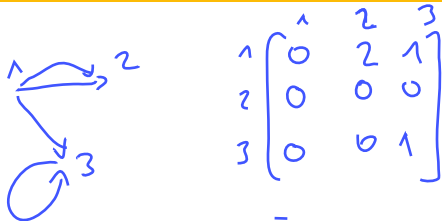- **adjacency matrix** of size $n \times n$ (where $n = |V|$) with

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

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

How does the adjacency matrix of an undirected graph look like? How to count the number of friends or suggest new friends in a social network?

Representation of a network

Quantitative measures of networks

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

- connected components

- degree distribution

- diameter and average path length

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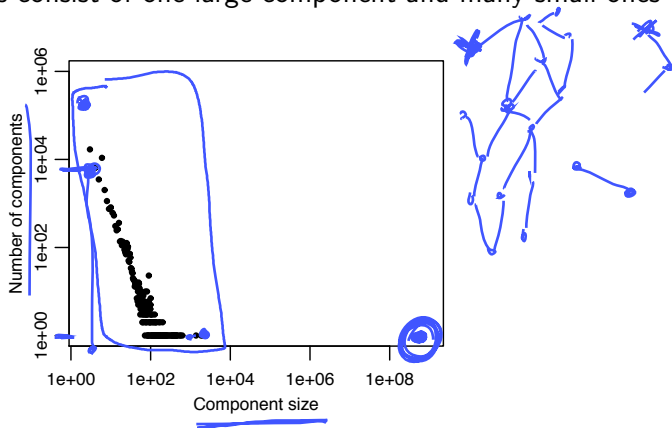
Component size distribution in the 2011 Facebook network on a log-log scale. Most vertices (99.91%) are in the largest component.

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:

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

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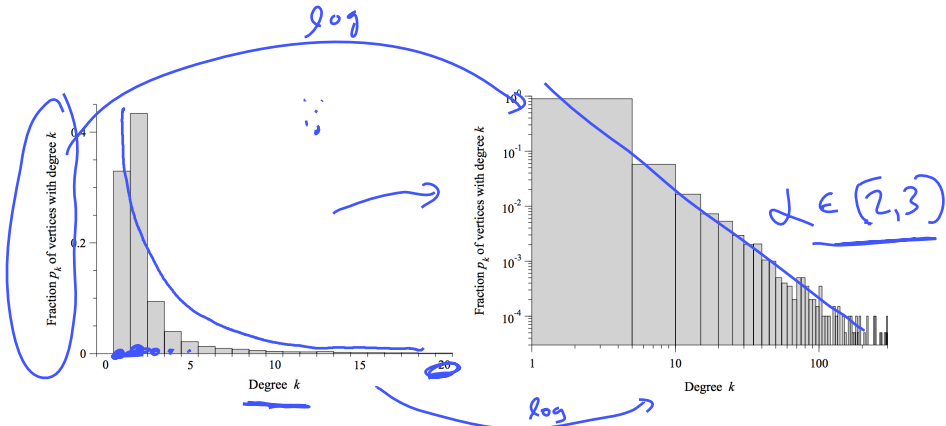
Figures from Chapter 8 in "Networks: An Introduction" by
M.E.J. Newman (2010)

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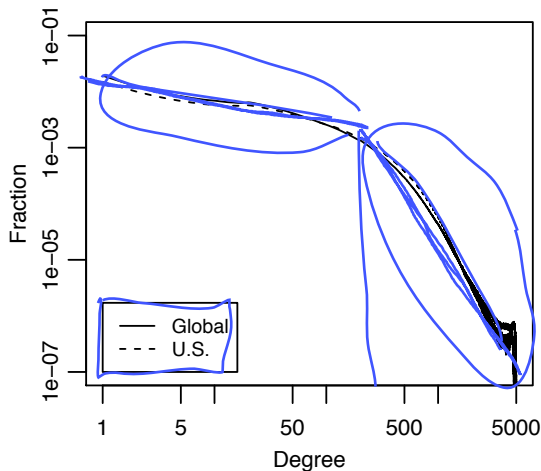
From "The Anatomy of the Facebook Social Graph" by Ugander et al. (2011)

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
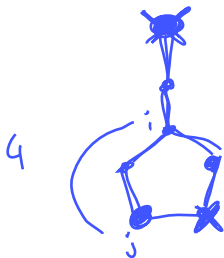
- The diameter of a network is the largest distance between any two nodes in the network:

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

- 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.

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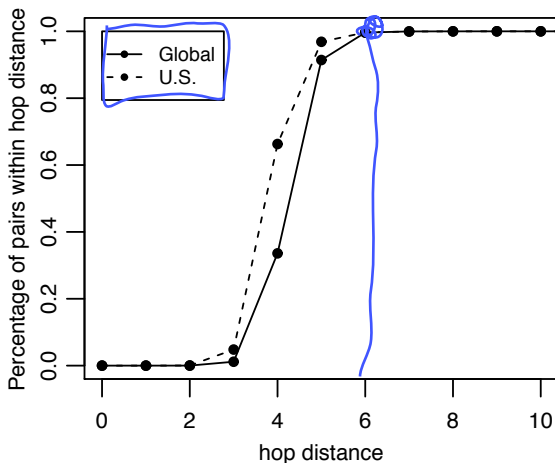
From "The Anatomy of the Facebook Social Graph" by Ugander et al. (2011)

Homophily

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)

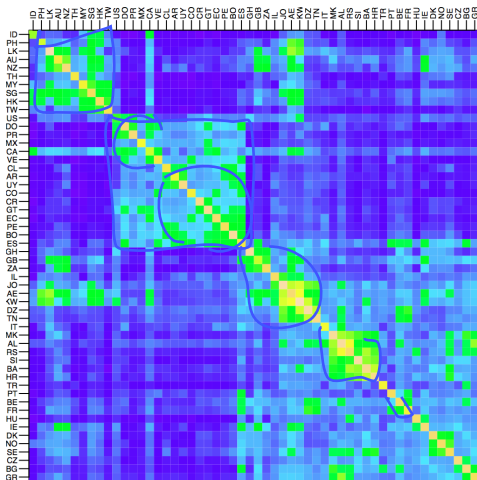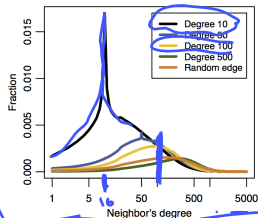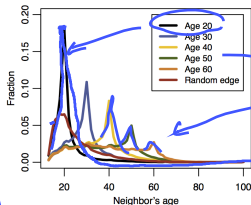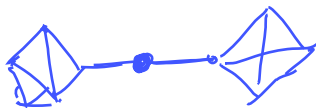|	Network	Type	n	m	c	S	ℓ	α
Social	Film actors	Undirected	449 913	25 516 482	113.43	0.980	3.48	2.3
	Company directors	Undirected	7 673	55 392	14.44	0.876	4.60	–
	Math coauthorship	Undirected	253 339	496 489	3.92	0.822	7.57	–
	Physics coauthorship	Undirected	52 909	245 300	9.27	0.838	6.19	–
	Biology coauthorship	Undirected	1 520 251	11 803 064	15.53	0.918	4.92	–
	Telephone call graph	Undirected	47 000 000	80 000 000	3.16			2.1
	Email messages	Directed	59 812	86 300	1.44	0.952	4.95	1.5/2.0
	Email address books	Directed	16 881	57 029	3.38	0.590	5.22	–
	Student dating	Undirected	573	477	1.66	0.503	16.01	–
	Sexual contacts	Undirected	2 810					3.2
Information	WWW nd.edu	Directed	269 504	1 497 135	5.55	1.000	11.27	2.1/2.4
	WWW AltaVista	Directed	203 549 046	1 466 000 000	7.20	0.914	16.18	2.1/2.7
	Citation network	Directed	783 339	6 716 198	8.57			3.0/–
	Roget's Thesaurus	Directed	1 022	5 103	4.99	0.977	4.87	–
	Word co-occurrence	Undirected	460 902	16 100 000	66.96	1.000		2.7
Technological	Internet	Undirected	10 697	31 992	5.98	1.000	3.31	2.5
	Power grid	Undirected	4 941	6 594	2.67	1.000	18.99	–
	Train routes	Undirected	587	19 603	66.79	1.000	2.16	–
	Software packages	Directed	1 439	1 723	1.20	0.998	2.42	1.6/1.4
	Software classes	Directed	1 376	2 213	1.61	1.000	5.40	–
	Electronic circuits	Undirected	24 097	53 248	4.34	1.000	11.05	3.0
	Peer-to-peer network	Undirected	880	1 296	1.47	0.805	4.28	2.1
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4
	Marine food web	Undirected	134	598	4.46	1.000	2.05	–
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–

Additional α column values (rightmost): Film actors 0.20, Company directors 0.59, Math coauthorship 0.15, Physics coauthorship 0.45, Biology coauthorship 0.088, Email address books 0.17, Student dating 0.005, WWW nd.edu 0.11, Roget's Thesaurus 0.13, Internet 0.035, Power grid 0.10, Software packages 0.070, Software classes 0.033, Electronic circuits 0.010, Peer-to-peer network 0.012, Metabolic network 0.090, Protein interactions 0.072, Marine food web 0.16, Freshwater food web 0.20, Neural network 0.18

$n = |\text{nodes}|$, $m = |\text{edges}|$, c: mean degree, S: prop. largest component, ℓ: mean geodesic, α: exp. power-law degree distribution, C: clustering coeff.

- **Centrality measure**: A measure that captures importance of a node's position in the network
- There are many different centrality measures

Find important nodes in a network

- **Centrality measure**: A measure that captures importance of a node's position in the network

- There are many different centrality measures

    - degree centrality (indegree / outdegree)
    - "propagated" degree centrality (score that is proportional to the sum of the score of all neighbors)
    - closeness centrality
    - betweenness centrality

Degree centrality

- For undirected graphs the degree k_i of node i is the number of edges connected to i, i.e. $k_i = \sum_j A_{ij}$

- For directed graphs the indegree of node i is $k_i^{\text{in}} = \sum_j A_{ji}$ and the outdegree is $k_i^{\text{out}} = \sum_j A_{ij}$

Degree centrality

- For undirected graphs the degree k_i of node i is the number of edges connected to i, i.e. $k_i = \sum_j A_{ij}$

- For directed graphs the indegree of node i is $k_i^{\text{in}} = \sum_j A_{ji}$ and the outdegree is $k_i^{\text{out}} = \sum_j A_{ij}$

- Simple, but intuitive: individuals with more connections have more influence and more access to information.

- Does not capture "cascade of effects": importance better captured by having connections to important nodes

Eigenvector centrality

- gives each node a score that is proportional to the sum of the scores of all its neighbors

$1 \to 2 \to 5$ $1 \to 2 \to 4$

$1 \to 3$
$\downarrow$
5

$1 \to 2 \to 3$

$1 \to 1 \to 3$
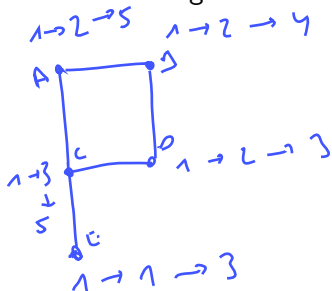
$$\begin{pmatrix} A \\ B \\ C \\ D \\ E \end{pmatrix} = \begin{pmatrix} A' \\ B' \\ C' \\ D' \\ E' \end{pmatrix} + \begin{pmatrix} A'' \\ B'' \\ C'' \\ D'' \\ E'' \end{pmatrix} \cdots$$

$$\cdots \quad \lambda v \quad v = \begin{pmatrix} A \\ B \\ C \\ D \\ E \end{pmatrix}$$

Eigenvector centrality

- gives each node a score that is proportional to the sum of the scores of all its neighbors

- need to know scores of all neighbors, which we don't know

- start with equal centrality: $x_i^{(0)} = 1$ for all nodes $i = 1, \ldots, n$

- update each centrality by the centrality of the neighbors:

$$x_i^{(1)} = \sum_{j=1}^{n} A_{ij} x_j^{(0)}$$

Eigenvector centrality

- gives each node a score that is proportional to the sum of the scores of all its neighbors

- need to know scores of all neighbors, which we don't know

- start with equal centrality: $x_i^{(0)} = 1$ for all nodes $i = 1, \ldots, n$

- update each centrality by the centrality of the neighbors:

$$x_i^{(1)} = \sum_{j=1}^{n} A_{ij} x_j^{(0)}$$

- iterate this process: $\quad x^{(k)} = A^k x^{(0)}$

Eigenvector centrality

- gives each node a score that is proportional to the sum of the scores of all its neighbors

- need to know scores of all neighbors, which we don't know

- start with equal centrality: $x_i^{(0)} = 1$ for all nodes $i = 1, \ldots, n$

- update each centrality by the centrality of the neighbors:

$$x_i^{(1)} = \sum_{j=1}^{n} A_{ij} x_j^{(0)}$$

- iterate this process: $\quad x^{(k)} = A^k x^{(0)}$

- if there exists $m > 0$ such that $A^m > 0$, then one can show that

$$x^{(k)} \xrightarrow{k \to \infty} \alpha \lambda_{max}^k v,$$

where λ_{max} is the largest eigenvalue and $v \geq 0$ the corresponding eigenvector; α depends on choice of $x^{(0)}$ (Perron-Frobenius theorem)

Eigenvector centrality

Interpretation: $v_i = \frac{1}{\lambda_{\max}} \sum_{j=1}^{n} A_{ij} v_j$

- node is important if it has important neighbors
- node is important if it has many neighbors
- eigenvector corresponding to largest eigenvalue of A provides a ranking of all nodes

Eigenvector centrality

Interpretation: $v_i = \frac{1}{\lambda_{\max}} \sum_{j=1}^{n} A_{ij} v_j$

- node is important if it has important neighbors
- node is important if it has many neighbors
- eigenvector corresponding to largest eigenvalue of A provides a ranking of all nodes

What happens when G is directed?

- right eigenvector: $v_i = \frac{1}{\lambda_{\max}} \sum_{j=1}^{n} A_{ij} v_j$
  - importance comes from nodes i points to
  - Example: determining malfunctioning genes

- left eigenvector: $w_i = \frac{1}{\lambda_{\max}} \sum_{j=1}^{n} w_j A_{ji}$
  - importance comes from nodes pointing to i
  - Example: ranking websites
  - Is the foundation for Google's PageRank algorithm

Other centrality measures

- **Closeness centrality:** Tracks how close a node is to any other node:

$$C_i = \left(\frac{1}{n-1} \sum_{j \neq i} d_{ij} \right)^{-1},$$

where d_{ij} is the distance between nodes i and j

Other centrality measures

- **Closeness centrality:** Tracks how close a node is to any other node:

$$C_i = \left(\frac{1}{n-1} \sum_{j \neq i} d_{ij} \right)^{-1},$$

where d_{ij} is the distance between nodes i and j

  - In disconnected networks: average over nodes in same component as i or use **harmonic centrality**: $H_i = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$

Other centrality measures

- **Closeness centrality:** Tracks how close a node is to any other node:

$$C_i = \left(\frac{1}{n-1} \sum_{j \neq i} d_{ij} \right)^{-1},$$

where d_{ij} is the distance between nodes i and j

  - In disconnected networks: average over nodes in same component as i or use **harmonic centrality**: $H_i = \frac{1}{n-1} \sum_{j \neq i} \frac{1}{d_{ij}}$

- **Betweenness centrality:** Measures the extent to which a node lies on paths between other nodes:

$$B_i = \frac{1}{n^2} \sum_{s,t} \frac{n_{st}^i}{g_{st}},$$

where n_{st}^i is number of shortest paths between s and t that pass through i, and g_{st} is total number of shortest paths between s and t

Which centrality measure to use

Choice of centrality measure depends on application!

Which centrality measure to use

Choice of centrality measure depends on application!

In a friendship network:

- high degree centrality: most popular person

- high eigenvector centrality: most popular person that is friends with popular people

- high closeness centrality: person that could best inform the group

- high betweenness centrality: person whose removal could best break the network apart
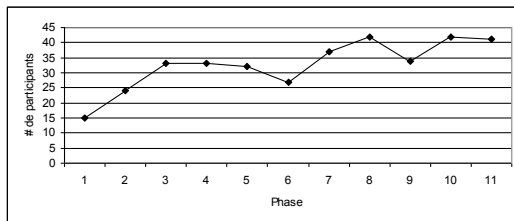
Case study: CAVIAR (criminal network in Montreal)

- Data based on 11 <u>wiretap</u> warrants from 1994-1996 → 11 periods

- Mandate of <u>CAVIAR project</u>: Seize drugs, arrests only in period 11

- 11 seizures total with monetary losses for traffickers of $32 mio
  - phase 4: 1 seizure $ 2.5mio, 300kg of <u>marijuana</u>
  - phase 6: 3 seizures $ 1.3mio, 2 x 15kg of marijuana, 1 x 2 kg of cocaine
  - phase 7: 1 <u>seizure</u> $ 3.5mio, 401kg of marijuana
  - phase 8: 1 seizure $ 0.4mio, 9kg of cocaine
  - phase 9: 2 seizures $ 4.3mio, 2kg of cocaine + 1 x 500kg marijuana
  - phase 10: 1 seizure $ 18.7mio, 2200kg of marijuana
  - phase 11: 2 seizures $ 1.3mio, 12kg of <u>cocaine</u> + 11kg of cocaine

Unique opportunity to study changes in the structure of a criminal network in upheaval by police forces
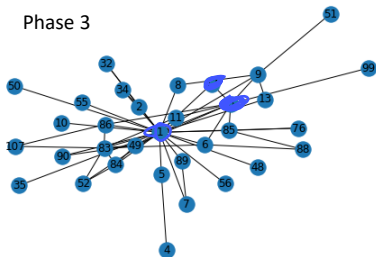
Case study: CAVIAR (criminal network in Montreal)

- network consists of 110 (numbered) players: 1-82 are traffickers, 83-110 are non-traffickers (financial investors, accountants, owners of various importation businesses, etc.)

- initially, investigation targeted Daniel Serero alleged mastermind of drug network in downtown Montreal

- initially marijuana was imported to Canada from Morocco

- after first seizure in phase 4, traffickers reoriented to cocaine import from Colombia, transiting through the United States
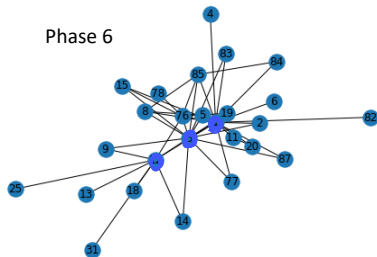
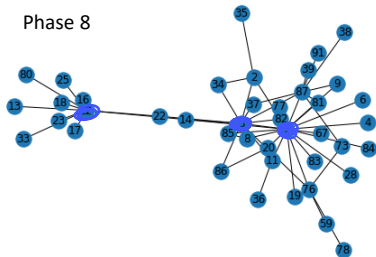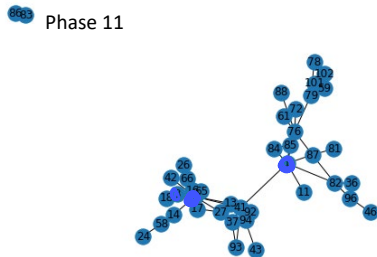# Case study: CAVIAR (criminal network in Montreal)

Case study: CAVIAR (criminal network in Montreal)

Role of the different actors:

- Daniel Serero (node 1): mastermind of the network

- Pierre Perlini (node 3): principal lieutenant of Serero (executes his instructions)

- Ernesto Morales (node 12): principal organizer of the cocaine import, intermediary between the Colombians and the Serero organization
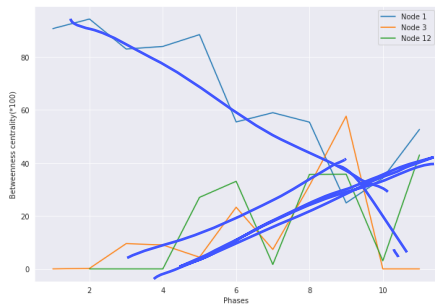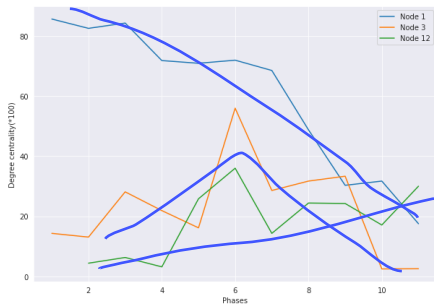
Additional thoughts: Criminal networks

- Given a social network and k criminal suspects, how to determine other suspects?

- Same question is extremely important in biology: given certain genes that are known to cause a certain disease, determine other candidate genes (e.g. based on protein-protein interaction network for determining autism genes: http://dx.doi.org/10.1101/057828)

- How do we identify nodes that are "between" a given set of seed nodes?

Steiner trees

Determine a small subnetwork that contains the given suspects / genes and connects these nodes

Steiner trees

Determine a small subnetwork that contains the given suspects / genes and connects these nodes

Steiner tree:

- shortest subnetwork that contains a given set of nodes
- NP-complete problem
- there exist polynomial time approximations

Steiner trees

Determine a small subnetwork that contains the given suspects / genes and connects these nodes

Steiner tree:

- shortest subnetwork that contains a given set of nodes
- NP-complete problem
- there exist polynomial time approximations

$\Rightarrow$ use collection of approximate Steiner trees for further analysis: autism interactome / criminal interactome

For genomics applications, see:

http://fraenkel-nsf.csbi.mit.edu/steinernet/tutorial.html

$\Rightarrow$ compute nodes with high betweenness centrality in interactome to obtain candidate genes / suspects

References

- Chapters 1 - 10 (but mostly chapters 6 - 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.

- For more information on the CAVIAR network:

  C. Morselli. *Inside Criminal Networks* (Springer, New York). Chapter 6: Law-enforcement disruption of a drug-importation network. 2009.