GL Applied Data Science Program

Network Analysis

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 $G = (V, E)$.

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:

- 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

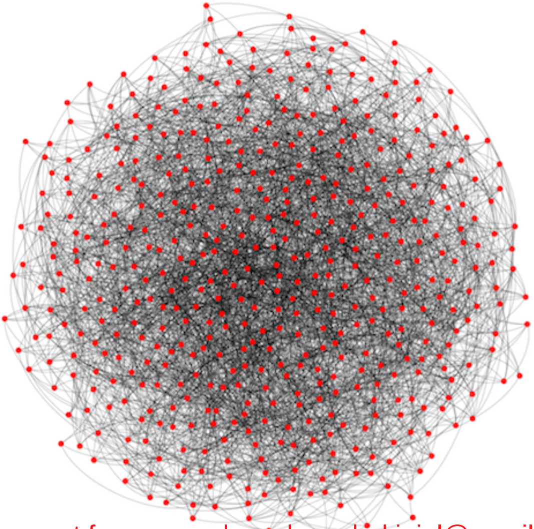
Examples of networks

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
transportation	airport	direct flight
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: $i \rightarrow j$ does not imply $j \rightarrow i$
  - e.g. World Wide Web, food web, citation network
- weighted network: with edge weights or vertex attributes
  - e.g. transportation networks
- bipartite network: edges between but not within classes
  - e.g. recommender systems such as 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 $1 \rightarrow 2 \leftarrow 3$: $E = \{(1,2),(3,2)\}$

- **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 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

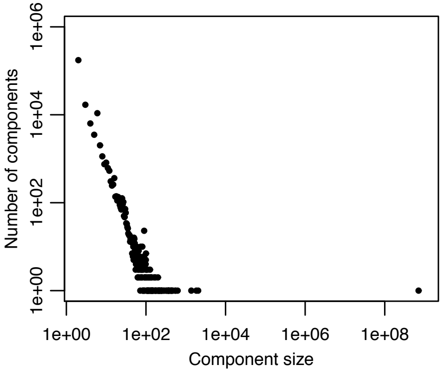
Connected Components

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

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.

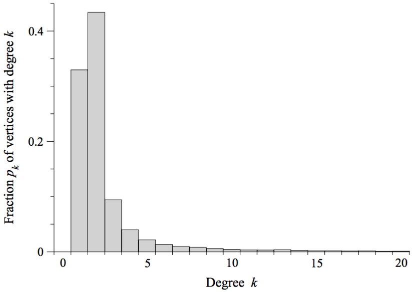
Degree distribution of the Internet

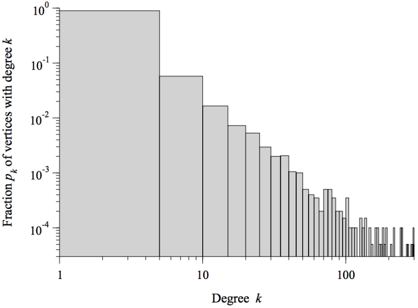
Degree of a node: number of edges connected to a node

Degree distribution of the Internet

Degree of a node: number of edges connected to a node

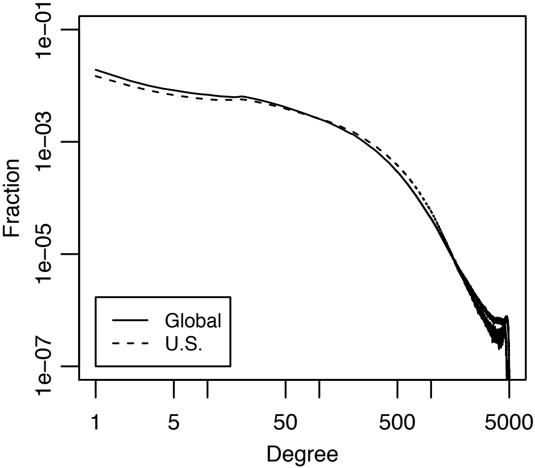
- Many networks show a power-law degree distribution (i.e., distribution that is linear in log-log plot)





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 of a graph

- 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:

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

Diameter of a graph

- 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:

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

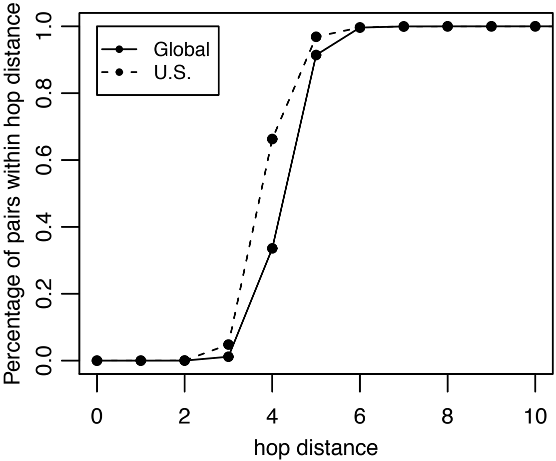
- If network is not connected, one often computes the diameter in the largest component.

- 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, suggesting that the diameter of the social network in the US is 6
- Any reasons why we should take the conclusion of 6 degrees of separation with a grain of salt?

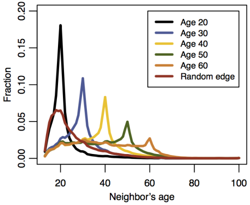
Diameter of Facebook (2011)

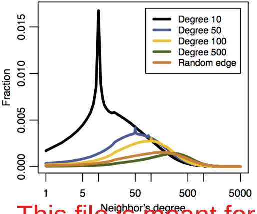


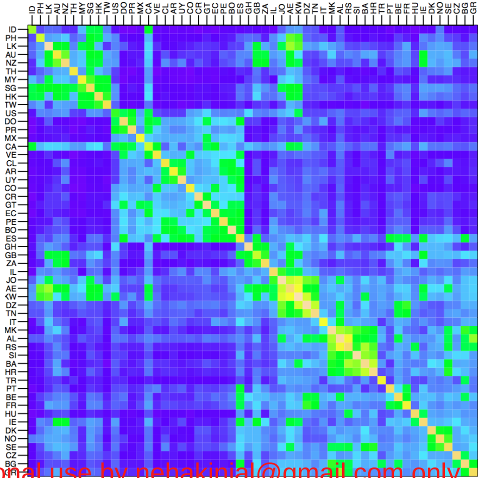
From "The Anatomy of the Facebook Social Graph" by Ugander et al. (2011)

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)

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:

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**
  - Simple and 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

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**
  - Simple and 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**
  - score that is proportional to the sum of the score of all neighbors is captured by largest eigenvector of adjacency matrix
  - builds the foundation for Google's PageRank algorithm

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**
  - Simple and 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**
  - score that is proportional to the sum of the score of all neighbors is captured by largest eigenvector of adjacency matrix
  - builds the foundation for Google's PageRank algorithm
- **closeness centrality**
  - tracks how close a node is to any other node

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**
  - Simple and 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**
  - score that is proportional to the sum of the score of all neighbors is captured by largest eigenvector of adjacency matrix
  - builds the foundation for Google's PageRank algorithm
- **closeness centrality**
  - tracks how close a node is to any other node
- **betweenness centrality**
  - measures the extent to which a node lies on paths between other nodes

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

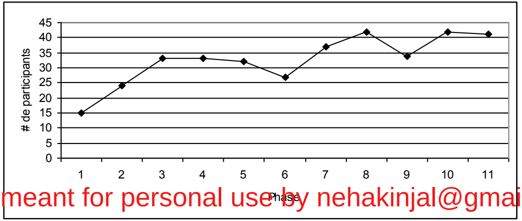
Case study: CAVIAR (criminal network in Montreal)

- Data based on 11 wiretap warrants from 1994-1996 $\rightarrow$ 11 periods
- Mandate of CAVIAR project: 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 marijuana
  - phase 6: 3 seizures $ 1.3mio, 2 x 15kg of marijuana, 1 x 2 kg of cocaine
  - phase 7: 1 seizure $ 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 cocaine + 11kg of cocaine

Unique opportunity to study changes in the structure of a criminal network in upheaval by police forces

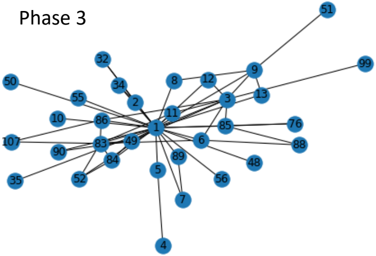
Case study: CAVIAR (criminal network in Montreal)

- 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

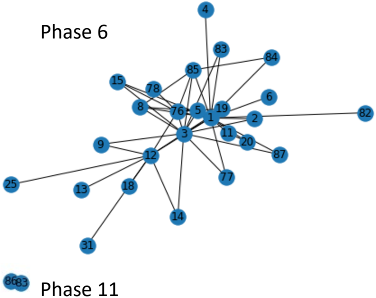


Case study: CAVIAR (criminal network in Montreal)

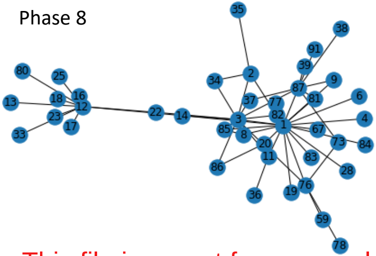
Phase 3



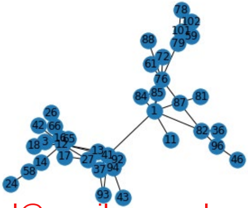
Phase 6



Phase 8



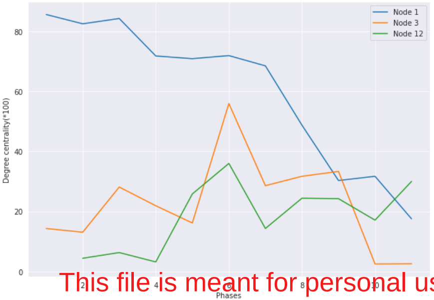
Phase 11

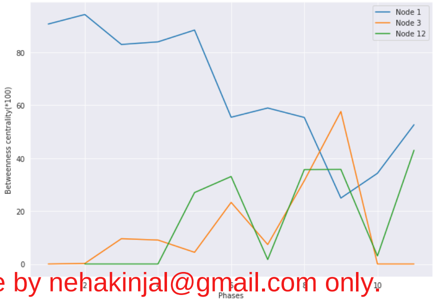


Case study: CAVIAR (criminal network in Montreal)

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





Optional: 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?

Optional: 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

⇒ 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

⇒ 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.