Introduction to Network

Observational Data Reading Group on Network Sampling

Ran Wei

Department of Statistics
The Ohio State University

September 11th, 2014

- Networks are ubiquitous in science and have become a focal point for discussion in everyday life.
- Formal statistical models for the analysis of network data have emerged as a major topic of interest in diverse areas of study.
- With the popularity of online social networks, the scale of network data has become enormous.

Outline

Outline

Why Networks

- Network data is ubiquitous.
- We live in a connected world: we are each separated from any other person on the planet by at most six other people (i.e., 'six degrees')
- Two elements: Nodes and Edges; Network data describe different types of connections and relationships between nodes.

Why Networks

- Network data is ubiquitous.
- We live in a connected world: we are each separated from any other person on the planet by at most six other people (i.e., 'six degrees')
- Two elements: Nodes and Edges; Network data describe different types of connections and relationships between nodes.

Why Networks

- Network data is ubiquitous.
- We live in a connected world: we are each separated from any other person on the planet by at most six other people (i.e., 'six degrees')
- Two elements: Nodes and Edges; Network data describe different types of connections and relationships between nodes.

Network Data

Connections and Relationships

- Interpersonal social or professional relationships: Facebook, Google+, Twitter, Weibo, LinkedIn.
- Academic paper co-authorships and citation relationships: DBLP, Cora and PubMed.
- Protein-protein interactions: Biology network.
- Sexual relationships: HIV patients network.
- Purchase and co-purchase relationships: Online auction network and shopping network.

Network Data

Connections and Relationships

- Interpersonal social or professional relationships: Facebook, Google+, Twitter, Weibo, LinkedIn.
- Academic paper co-authorships and citation relationships: DBLP, Cora and PubMed.
- Protein-protein interactions: Biology network.
- Sexual relationships: HIV patients network.
- Purchase and co-purchase relationships: Online auction network and shopping network.

Network Data

Connections and Relationships

- Interpersonal social or professional relationships: Facebook, Google+, Twitter, Weibo, LinkedIn.
- Academic paper co-authorships and citation relationships: DBLP, Cora and PubMed.
- Protein-protein interactions: Biology network.
- Sexual relationships: HIV patients network.
- Purchase and co-purchase relationships: Online auction network and shopping network.

Network Data

Connections and Relationships

- Interpersonal social or professional relationships: Facebook, Google+, Twitter, Weibo, LinkedIn.
- Academic paper co-authorships and citation relationships: DBLP, Cora and PubMed.
- Protein-protein interactions: Biology network.
- Sexual relationships: HIV patients network.
- Purchase and co-purchase relationships: Online auction network and shopping network.

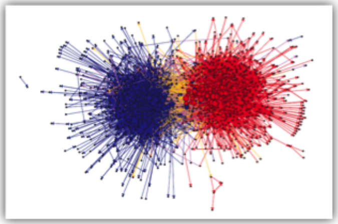
Network Data

Connections and Relationships

- Interpersonal social or professional relationships: Facebook, Google+, Twitter, Weibo, LinkedIn.
- Academic paper co-authorships and citation relationships: DBLP, Cora and PubMed.
- Protein-protein interactions: Biology network.
- Sexual relationships: HIV patients network.
- Purchase and co-purchase relationships: Online auction network and shopping network.

Outline

A visualization of US bloggers shows clearly how they tend to link predominantly to blogs supporting the same party, forming two distinct clusters (Adamic and Glance, 2005)



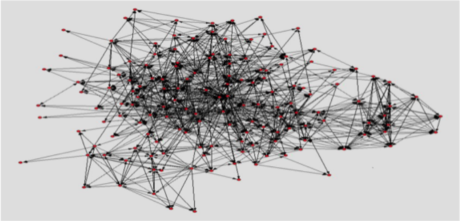


Figure 2.2: E-mail exchange data among 151 Enron executives, using a threshold of a minimum of 5 messages for each link. Source: [153].

Large-Scale Online Social Networks

Scale of Online Social Networks

Number of monthly active users:

- Facebook: 1.23 billion (December, 2013)
- Google+: 540 million (October, 2013)
- LinkedIn: 259 million (June, 2013)
- Twitter: 200 million (February, 2013)

Large-Scale Online Social Networks

Studies and Applications

- User behavior analysis
  - Influence and passivity of users: Does high popularity imply high influence and vice-versa?
- Community detection
  - Identify clusters of customers with similar interests in the network of purchase relationships.
- Link and attribute prediction
  - 'Friend you may know', 'Who to add to circle', 'Who to connect' and 'Who to follow'.
- Make predictions on real-time social events
  - Flu trends, box-office revenues for movies, the stock market, and earthquakes.

Large-Scale Online Social Networks

Studies and Applications

- User behavior analysis
  - Influence and passivity of users: Does high popularity imply high influence and vice-versa?
- Community detection
  - Identify clusters of customers with similar interests in the network of purchase relationships.
- Link and attribute prediction
  - 'Friend you may know', 'Who to add to circle', 'Who to connect' and 'Who to follow'.
- Make predictions on real-time social events
  - Flu trends, box-office revenues for movies, the stock market, and earthquakes.

Large-Scale Online Social Networks

Studies and Applications

- User behavior analysis
  - Influence and passivity of users: Does high popularity imply high influence and vice-versa?
- Community detection
  - Identify clusters of customers with similar interests in the network of purchase relationships.
- Link and attribute prediction
  - 'Friend you may know', 'Who to add to circle', 'Who to connect' and 'Who to follow'.
- Make predictions on real-time social events
  - Flu trends, box-office revenues for movies, the stock market, and earthquakes.

Large-Scale Online Social Networks

Studies and Applications

- User behavior analysis
  - Influence and passivity of users: Does high popularity imply high influence and vice-versa?
- Community detection
  - Identify clusters of customers with similar interests in the network of purchase relationships.
- Link and attribute prediction
  - ‘Friend you may know’, ‘Who to add to circle’, ‘Who to connect’ and ‘Who to follow’.
- Make predictions on real-time social events
  - Flu trends, box-office revenues for movies, the stock market, and earthquakes.

Outline

Graph Representation of a Network

- $G = \{V, E\}$
- $V = \{v_1, v_2, ..., v_{N_v}\}$ and $E = \{e_1, e_2, ..., e_{N_e}\}$
- $W = \{w_1, w_2, ..., w_{N_e}\}$.
- Attributes of nodes: $Y_{N_v \times p}$.
  The p-th attribute vector: $Y_p = (y_{p1}, y_{p2}, ..., y_{pN_v})^T$.
- Adjacency matrix $A = (a_{ij})_{N \times N}$:
  $a_{ij} = 1$ if there's a link from node v_i to v_j;
  $a_{ij} = 0$ if there's no link from v_i to v_j $(v_i, v_j \in V)$.

Graph Representation of a Network

- $G = \{V, E\}$
- $V = \{v_1, v_2, ..., v_{N_v}\}$ and $E = \{e_1, e_2, ..., e_{N_e}\}$
- $W = \{w_1, w_2, ..., w_{N_e}\}$.
- Attributes of nodes: $Y_{N_v \times p}$.
  The p-th attribute vector: $Y_p = (y_{p1}, y_{p2}, ..., y_{pN_v})^T$.
- Adjacency matrix $A = (a_{ij})_{N \times N}$:
  $a_{ij} = 1$ if there's a link from node v_i to v_j;
  $a_{ij} = 0$ if there's no link from v_i to v_j ($v_i, v_j \in V$).

Graph Representation of a Network

- $G = \{V, E\}$
- $V = \{v_1, v_2, ..., v_{N_v}\}$ and $E = \{e_1, e_2, ..., e_{N_e}\}$
- $W = \{w_1, w_2, ..., w_{N_e}\}$.
- Attributes of nodes: $Y_{N_v \times p}$.
  The p-th attribute vector: $Y_p = (y_{p1}, y_{p2}, ..., y_{pN_v})^T$.
- Adjacency matrix $A = (a_{ij})_{N \times N}$:
  $a_{ij} = 1$ if there's a link from node v_i to v_j;
  $a_{ij} = 0$ if there's no link from v_i to v_j $(v_i, v_j \in V)$.

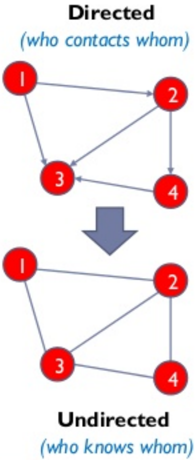
Graph Representation of a Network

- $G = \{V, E\}$
- $V = \{v_1, v_2, ..., v_{N_v}\}$ and $E = \{e_1, e_2, ..., e_{N_e}\}$
- $W = \{w_1, w_2, ..., w_{N_e}\}$.
- Attributes of nodes: $Y_{N_v \times p}$.
  The p-th attribute vector: $Y_p = (y_{p1}, y_{p2}, ..., y_{pN_v})^T$.
- Adjacency matrix $A = (a_{ij})_{N \times N}$:
  $a_{ij} = 1$ if there's a link from node v_i to v_j;
  $a_{ij} = 0$ if there's no link from v_i to v_j $(v_i, v_j \in V)$.

Graph Representation of a Network

- $G = \{V, E\}$
- $V = \{v_1, v_2, ..., v_{N_v}\}$ and $E = \{e_1, e_2, ..., e_{N_e}\}$
- $W = \{w_1, w_2, ..., w_{N_e}\}$.
- Attributes of nodes: $Y_{N_v \times p}$.
  The p-th attribute vector: $Y_p = (y_{p1}, y_{p2}, ..., y_{pN_v})^T$.
- Adjacency matrix $A = (a_{ij})_{N \times N}$:
  $a_{ij} = 1$ if there's a link from node v_i to v_j;
  $a_{ij} = 0$ if there's no link from v_i to v_j $(v_i, v_j \in V)$.

Edge List and Adjacency Matrix



Edge list remains the same

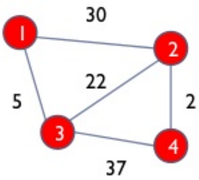
Vertex	Vertex
1	2
1	3
2	3
2	4
3	4

But interpretation is different now

Adjacency matrix becomes symmetric

Vertex	1	2	3	4
1	-	1	1	0
2	1	-	1	1
3	1	1	-	1
4	0	1	1	-

Weight of Edges



Weights could be:

- *Frequency of interaction in period of observation*
- *Number of items exchanged in period*
- *Individual perceptions of strength of relationship*
- *Costs in communication or exchange, e.g. distance*
- *Combinations of these*

Edge list: add column of weights

Vertex	Vertex	Weight
1	2	30
1	3	5
2	3	22
2	4	2
3	4	37

Adjacency matrix: add weights instead of 1

Vertex	1	2	3	4
1	-	30	5	0
2	30	-	22	2
3	5	22	-	37
4	0	2	37	-

Outline

Metrics in Network

- Centrality Measures
  - Degree
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality
- Transitivity
  - Clustering Coefficient
- Community Structure
  - Clustering Algorithm

Metrics in Network

- Centrality Measures
  - Degree
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality
- Transitivity
  - Clustering Coefficient
- Community Structure
  - Clustering Algorithm

Metrics in Network

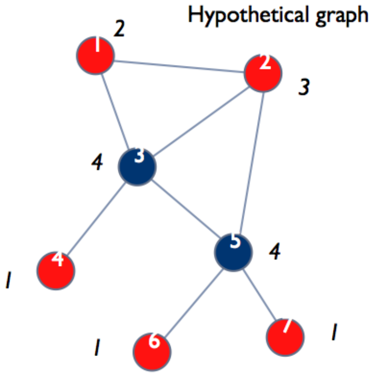
- Centrality Measures
  - Degree
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality
- Transitivity
  - Clustering Coefficient
- Community Structure
  - Clustering Algorithm

Centrality Metrics

Degree

- A node's (in-) or (out-)degree is the number of links that lead into or out of the node
- In an undirected graph they are of course identical
- Often used as measure of a node's degree of connectedness and hence also influence and/or popularity
- Useful in assessing which nodes are central with respect to spreading information and influencing others in their immediate 'neighborhood'

Nodes 3 and 5 have the highest degree (4)

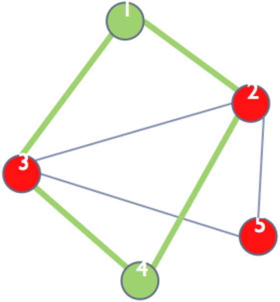


Centrality Metrics

Shortest Path

- A *path* between two nodes is any sequence of non-repeating nodes that connects the two nodes
- The *shortest path* between two nodes is the path that connects the two nodes with the shortest number of edges (also called the *distance* between the nodes)
- In the example to the right, between nodes 1 and 4 there are two shortest paths of length 2: {1,2,4} and {1,3,4}
- Other, longer paths between the two nodes are {1,2,3,4}, {1,3,2,4}, {1,2,5,3,4} and {1,3,5,2,4} (the longest paths)
- Shorter paths are desirable when speed of communication or exchange is desired (often the case in many studies, but sometimes not, e.g. in networks that spread disease)

Hypothetical graph

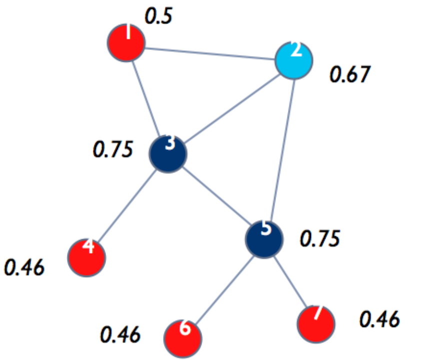


Centrality Metrics

Closeness Centrality

- Calculate the mean length of all shortest paths from a node to all other nodes in the network (i.e. how many hops on average it takes to reach every other node)
- Take the reciprocal of the above value so that higher values are 'better' (indicate higher closeness) like in other measures of centrality
- It is a measure of *reach*, i.e. the speed with which information can reach other nodes from a given starting node

Nodes 3 and 5 have the highest (i.e. best) closeness, while node 2 fares almost as well



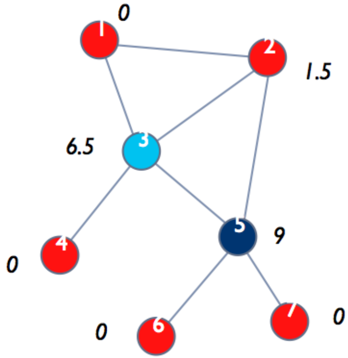
Note: Sometimes closeness is calculated without taking the reciprocal of the mean shortest path length. Then lower values are 'better'.

Centrality Metrics

Betweenness Centrality

- For a given node v, calculate the number of shortest paths between nodes i and j that pass through v, and divide by all shortest paths between nodes i and j
- Sum the above values for all node pairs i,j
- Sometimes normalized such that the highest value is 1or that the sum of all betweenness centralities in the network is 1
- Shows which nodes are more likely to be in communication paths between other nodes
- Also useful in determining points where the network would break apart (think who would be cut off if nodes 3 or 5 would disappear)

Node 5 has higher betweenness centrality than 3

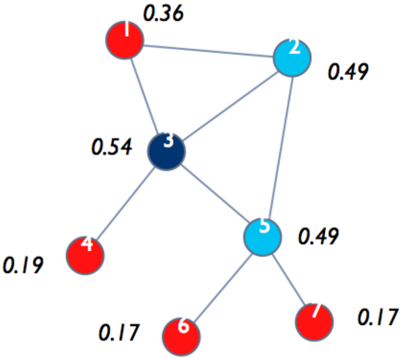


Centrality Metrics

Eigenvector Centrality

- A node's eigenvector centrality is proportional to the sum of the eigenvector centralities of all nodes directly connected to it
- In other words, a node with a high eigenvector centrality is connected to other nodes with high eigenvector centrality
- This is similar to how Google ranks web pages: links from highly linked-to pages count more
- Useful in determining who is connected to the most connected nodes

Node 3 has the highest eigenvector centrality, closely followed by 2 and 5



Note: The term 'eigenvector' comes from mathematics (matrix algebra), but it is not necessary for understanding how to interpret this measure

Centrality Metrics

Comparison

Centrality measure	Interpretation in social networks
Degree	How many people can this person reach directly?
Betweenness	How likely is this person to be the most direct route between two people in the network?
Closeness	How fast can this person reach everyone in the network?
Eigenvector	How well is this person connected to other well-connected people?

Centrality Metrics

Comparison – Example

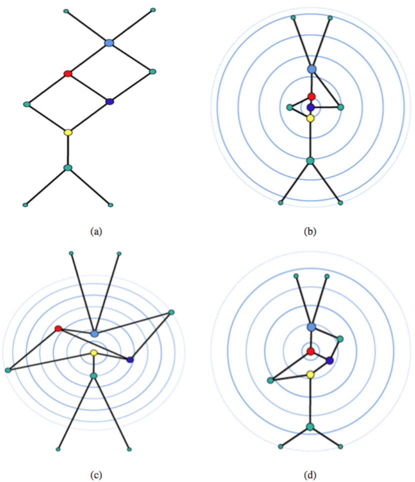
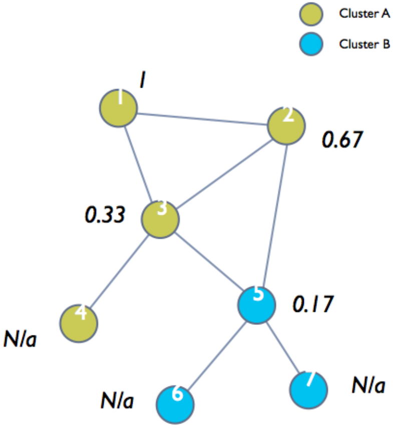


Fig. 4.4 Illustration of (b) closeness, (c) betweenness, and (d) eigenvector centrality measures on the graph in (a). Example and figures courtesy of Ulrik Brandes.

Clustering

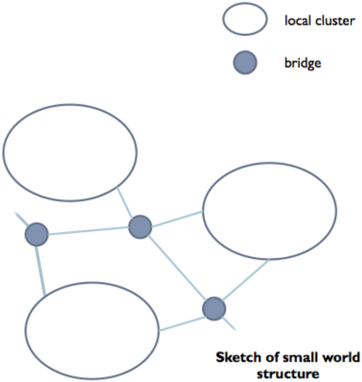
- A node's *clustering coefficient* is the number of closed triplets in the node's neighborhood over the total number of triplets in the neighborhood. It is also known as *transitivity*.
- E.g., node 1 to the right has a value of 1 because it is only connected to 2 and 3, and these nodes are also connected to one another (i.e. the only triplet in the neighborhood of 1 is closed). We say that nodes 1,2, and 3 form a *clique*.
- Clustering algorithms identify clusters or 'communities' within networks based on network structure and specific clustering criteria (example shown to the right with two clusters is based on edge betweenness, an equivalent for edges of the betweenness centrality presented earlier for nodes)



Network clustering coefficient = 0.375
(3 nodes in each triangle x 2 triangles = 6 closed triplets divided by 16 total)

Small World

- A small world is a network that looks almost random but exhibits a significantly *high clustering coefficient* (nodes tend to cluster locally) and a relatively *short average path length* (nodes can be reached in a few steps)
- It is a very common structure in social networks because of transitivity in strong social ties and the ability of weak ties to reach across clusters (see also next page...)
- Such a network will have many clusters but also many bridges between clusters that help shorten the average distance between nodes



Sketch of small world structure

You may have heard of the famous "6 degrees" of separation

Summary

- Motivation of studying network
- Network Data and Examples
- Graph Notification of Networks
- Metrics in Network

Summary

- Motivation of studying network
- Network Data and Examples
- Graph Notification of Networks
- Metrics in Network

Summary

- Motivation of studying network
- Network Data and Examples
- Graph Notification of Networks
- Metrics in Network

Summary

- Motivation of studying network
- Network Data and Examples
- Graph Notification of Networks
- Metrics in Network