Social Network Analysis

Lectured by: Dr. Xinzhi ZHANG

Research Assistant Professor, Department of Journalism

Hong Kong Baptist University

13 June, 2018

@CUCN Data-driven Journalism Workshop

Agenda

- Why social network analysis (SNA) is important
- Basic concepts of SNA
    - For nodes
    - For edges
    - For the network
- Centrality
- Network visualization
    - Layout
    - Community detection
- Network data structure
- Gephi tutorial

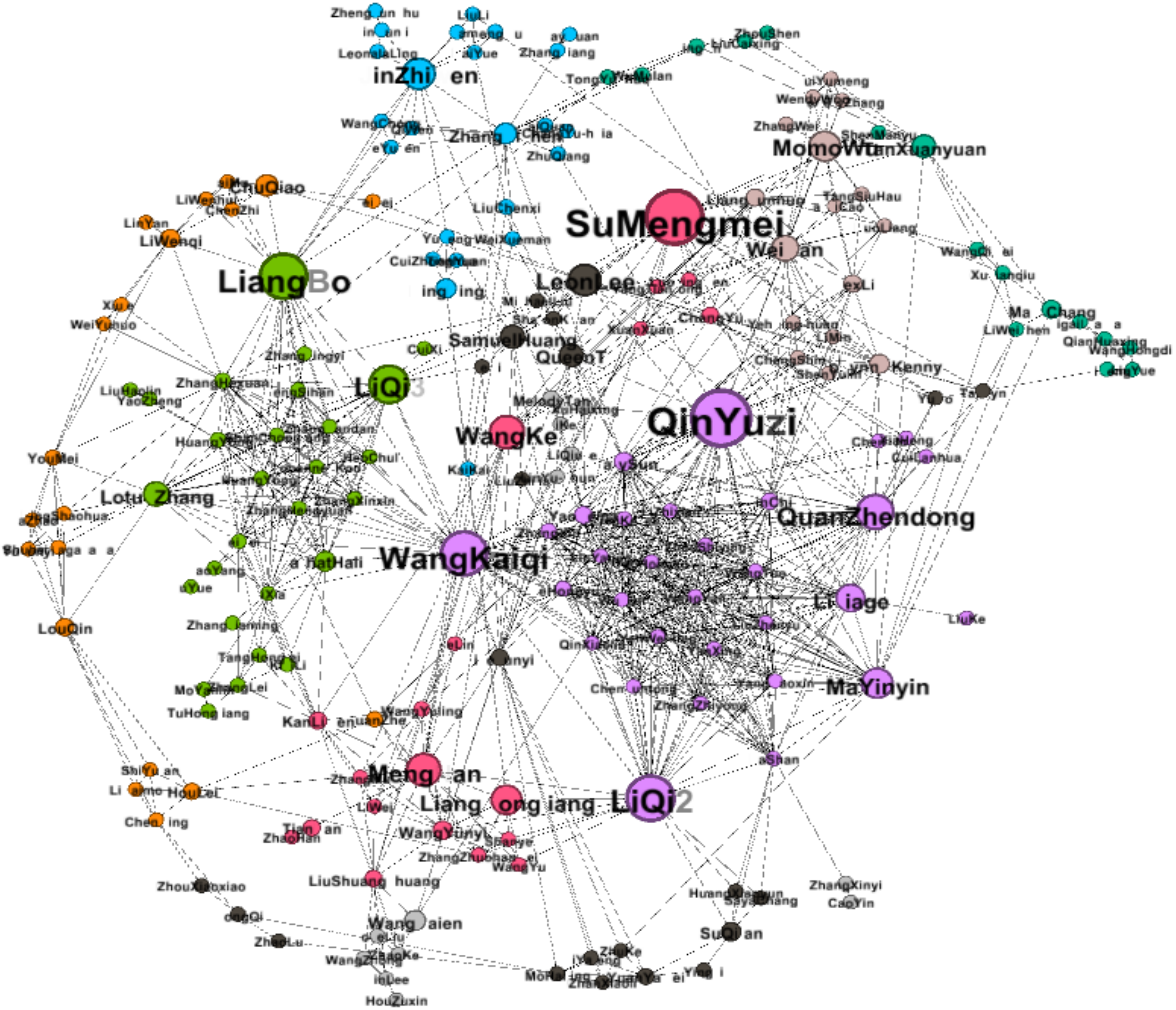


Figure 2. A snapshot of the contestant network of the Voice of China, illustrating the co-cover process (note: only the connected section of the graph is illustrated. The figure was plotted by Gephi).

Why Network Analysis?

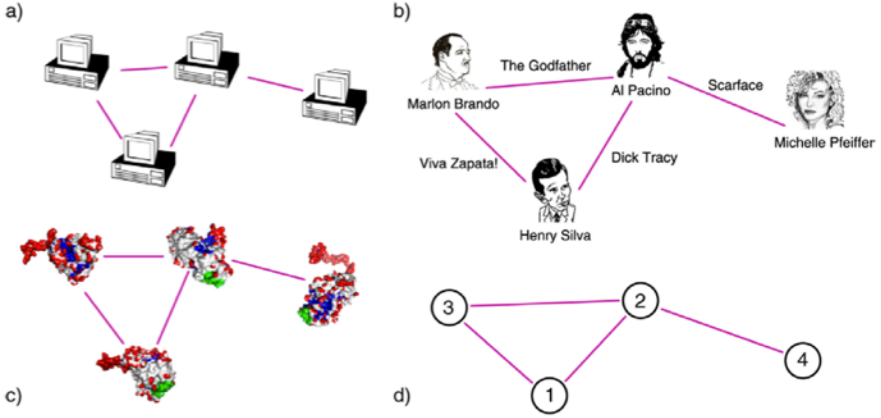
- "Behind each complex system, there is an intricate network that encodes the interactions between the system's components" (Barabási, 2012, p. 6).
- Deriving from the discrete mathematics and statistical physics, social network analysis gains its popularity in social sciences/humanities in 2000s;
- It becomes one of the most important techniques of data-driven storytelling, pattern recognition, digital humanities, computation social sciences, infographics, and a number of others.
- It addresses a "network analytics approach."

Why Network Analysis?

- Homophily (Similarity) 物以类聚，人以群分
- Social relationships 社会关系
- Social exchanges, reciprocity 礼尚往来
- Diffusion and social influence 不胫而走？
- Co-occurrence 为了相同的梦想走到一起
- …

SNA: Basic Concepts

- A network is a catalog of a system’s components often called nodes or vertices and the direct interactions between them, called links or edges (Barabási, 2012)
- There are directed networks and undirected networks.



Directed | Undirected

SNA: Basic Concepts

Discipline	“Notes”	“Link”	Network Name	Directed/ Undirected
Mathematics/Geometry/Graph Theory	Vertex	Edge	Graph	Either
Physics/Network Science	Node	Link	Network	Either
Sociology	Agency	Social Relations	Structure	Either
Citation Network	Scholar	Citation	Citation Network	Directed
Global Communication	Counties	Co-current in the media coverage	Media Coverage of International Relations	Undirected
Cooking Culture	Recipient	Cooking Methods	Cuisine	Undirected

SNA: Basic Concepts: for Nodes

- **Degree**: k(i): for each node, the degree represents the total number of its links/edges.
    - For directed networks, nodes could have incoming links and outgoing links (followers vs followings).

SNA: Basic Concepts: for Edges

- Edges can be weighted, or unweighted.
- For example, “in mobile call networks the weight can represent the total number of minutes two mobile phone users talk with each other on the phone; on the power grid the weight is the amount of current flowing through a transmission line” (Barabási, 2012).

SNA: Basic Concepts: for Networks

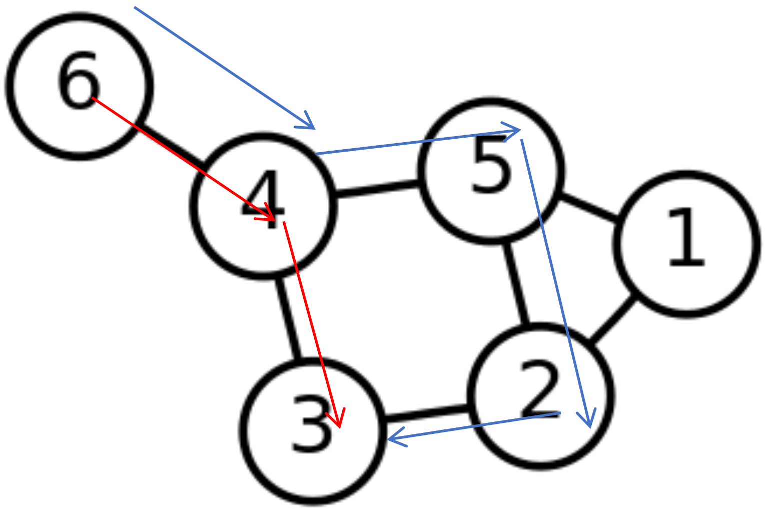
- N: total number of nodes (size of the network)
- L: total number of links
- <k>: Average Degree
- Degree Distribution
- Average (global) Clustering Coefficient
- Average Path Length

SNA: Basic Concepts: for Networks

- The degree distribution provides the probability that a randomly selected node in the network has degree k.
- Normally, for social networks, many low degree nodes and fewer high degree nodes (power-law distribution)

SNA: Basic Concepts: for Networks

- A Path is a route that runs along the links of the network, its length representing the number of links the path contains.
- Shortest Path between nodes i and j is the path with fewest number of links.
- Average path length is the average number of steps along the shortest paths for all possible pairs of network nodes.
- It is a measure of the efficiency of information or mass transport on a network.



SNA: Basic Concepts: for Networks

- Network diameter: the diameter of a network, denoted by d_{max}, is the maximal shortest path in the network.
- In other words, it is the largest distance recorded between any pair of nodes.

SNA: Basic Concepts: for Networks

- Clustering Coefficient (CC): “local CC & global CC”
- The global CC was designed to give an overall indication of the clustering in the network;
- The local gives an indication of the embeddedness of one single nodes.

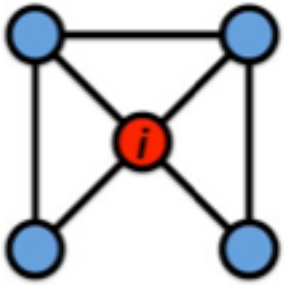
SNA: Basic Concepts: for Networks

- Local CC: the degree to which the neighbors of a given node link to each other. For a node i with degree ki the local clustering coefficient is defined as:

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

- Li represents the number of links between the ki neighbors of node i.
- Ci is between 0 and 1.

- ***"How many of your friends are also friends themselves?"***



$C_i = 1/2$

SNA: Basic Concepts: for Networks

- Average CC (averaging the local CC or all the nodes): The degree of clustering of a whole network is captured by the average clustering coefficient, <C>, representing the average of Ci over all nodes i = 1, ..., N.

SNA: Basic Concepts: for Networks

- Global CC: the total number of closed triangles in a network. [Transitivity]
- The degree of a network's global clustering is captured by the global clustering coefficient, defined as

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triples of vertices}}.$$

SNA: Centrality [Important]

- Centrality measures the “importance” or the “power of influence” of the nodes.
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality

SNA: Degree Centrality

- Degree Centrality: the number of links incident upon a node (i.e., the number of ties that a node has).
- Sometimes Degree Centrality may not fully reveal the importance of one node in the network, however.
- Degree centrality only takes into account the immediate ties that an actor has, or the ties of the actor's neighbors, rather than indirect ties to all others.
- One actor might be tied to a large number of others, but those others might be rather disconnected from the network as a whole. In a case like this, the actor could be quite central, but only in a local neighborhood.
- [URL: http://www.activatenetworks.net/blog/who-is-central-to-a-social-network-it-depends-on-your-centrality-measure]

SNA: Closeness Centrality

- Closeness Centrality: the reverse of average distance (=the length of the shortest paths) from each individual to every other individual in the network.
- Closeness can be regarded as a measure of how long it will take to spread information from the node to all other nodes sequentially.
  - Example: Duncan Watts: Small Worlds: The Dynamics of Networks between Order and Randomness: “Kelvin Bacon Index”

SNA: Closeness Centrality

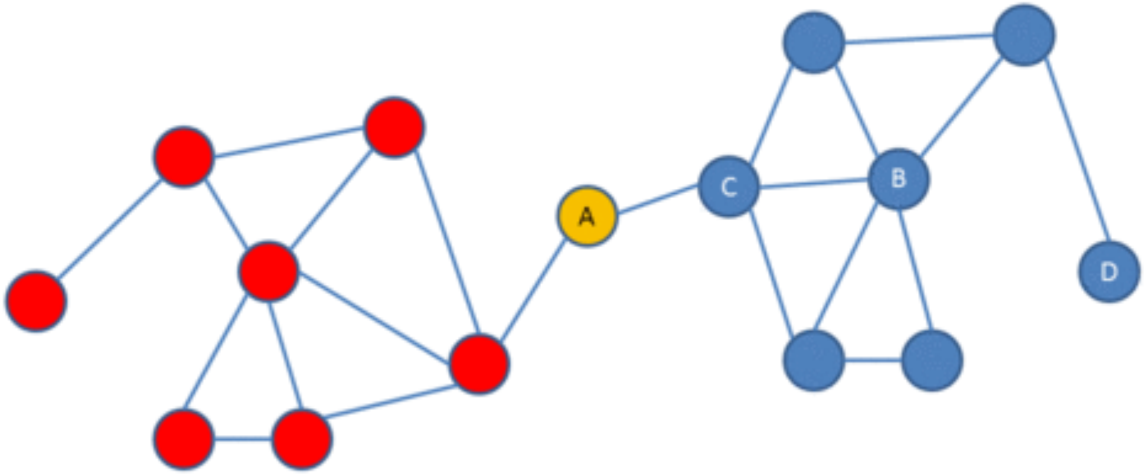
- Six Degrees of Kevin Bacon is a parlor game based on the "six degrees of separation" concept, which posits that any two people on Earth are six or fewer acquaintance links apart.
    - Kevin Bacon has a Bacon number of 0.
    - Those actors who have worked directly with Kevin Bacon have a Bacon number of 1.
    - If the lowest Bacon number of any actor with whom X has appeared in any movie is N, X's Bacon number is N+1.
    - (Duncan Watts: Small Worlds: The Dynamics of Networks between Order and Randomness)

SNA: Closeness Centrality

- Application of Closeness Centrality [URL: http://www.activatenetworks.net/blog/who-is-central-to-a-social-network-it-depends-on-your-centrality-measure]
- High closeness centrality individuals tend to be important influencers within their local network community.
- They may NOT be public figures to the entire network of a corporation or profession, but they are often respected locally and they occupy short paths for information spread within their network community.

SNA: Betweenness Centrality

- Betweenness Centrality: the number of shortest paths from all vertices to all others that pass through that node.
- Example: Which node is more “important” in the entire network? Node A, or Node B?



SNA: Betweenness Centrality

- “一夫當關，萬夫莫開”
- High betweenness individuals are often critical to collaboration across departments and to maintaining the spread of a new product through an entire network. Because of their locations between network communities, they are natural brokers of information and collaboration.
- High betweenness individuals often are overlooked. Because they are NOT central to any single social clique, and instead reside on the periphery of several such cliques each of which all engender more trust and admiration within rather than outside of the clique.
- [URL: http://www.activatenetworks.net/blog/who-is-central-to-a-social-network-it-depends-on-your-centrality-measure]

SNA: Eigenvector Centrality

- Eigenvector Centrality: how well connected an individual is to the parts of the network with the greatest connectivity.
- Individuals with high eigenvector scores have many connections, and their connections have many connections, and their connections have many connections … out to the end of the network.
- [URL: http://www.activatenetworks.net/blog/who-is-central-to-a-social-network-it-depends-on-your-centrality-measure]

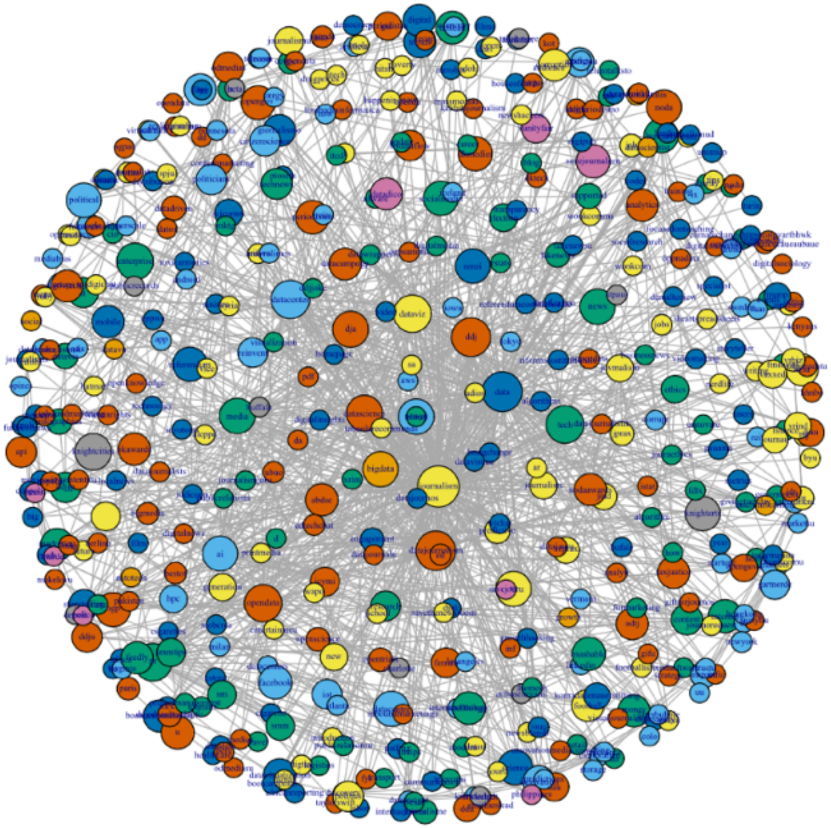
SNA: Eigenvector Centrality

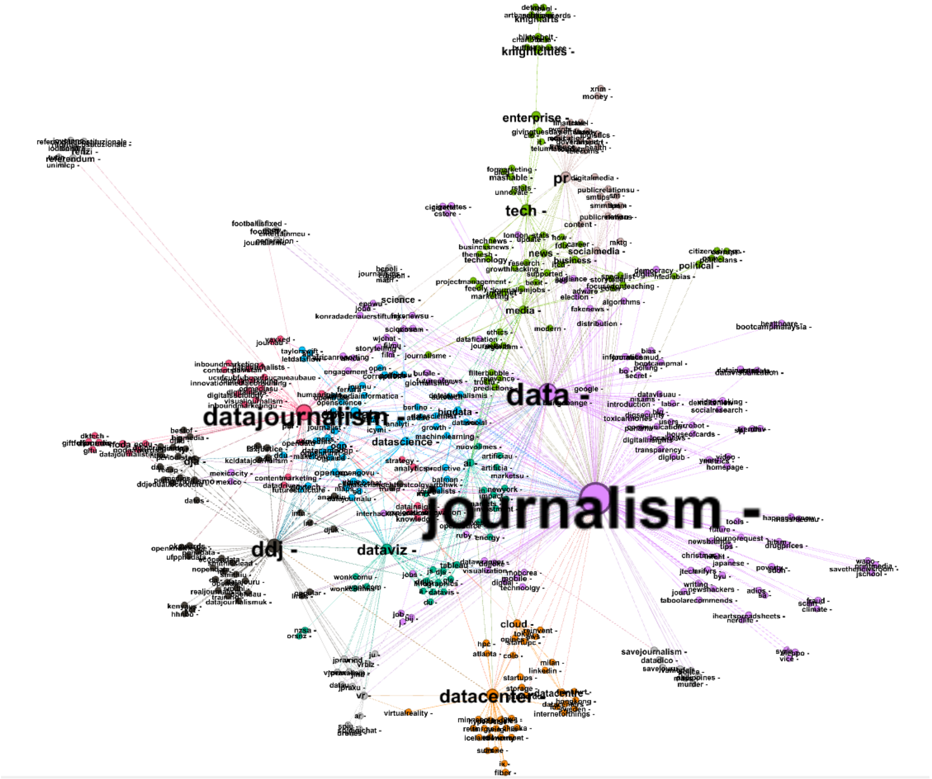
- High eigenvector centrality individuals are leaders of the network. They are often public figures with many connections to other high-profile individuals.
    - Private assist of the President.
    - Google's page rank.
- High eigenvector centrality individuals, however, cannot necessarily perform the roles of high closeness and betweenness.
- They do NOT always have the greatest local influence and may have limited brokering potential.
- [URL: http://www.activatenetworks.net/blog/who-is-central-to-a-social-network-it-depends-on-your-centrality-measure]

SNA: Community Detection

- A community is “a group of people who think or act collectively.”

Network Visualization: Layout





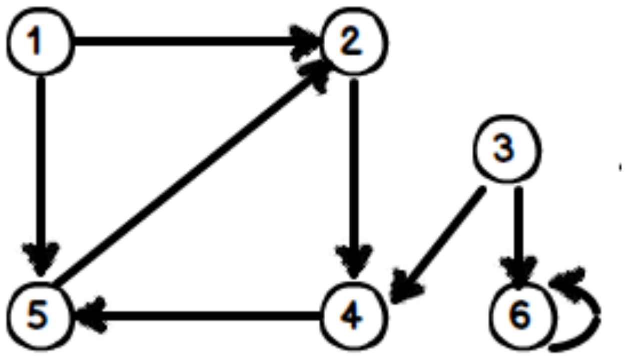
- Zhang (2017)
- Graph plotted by *igraph* with sphere layout,
- node sizes are proportioned to betweeness centrality,
- Node colors indicate the community (fast greedy community detection)

- Zhang (2017)
- Graph plotted by *Gephi* with Forced atlas 2 layout
- node sizes are proportioned to betweeness centrality,
- Node colors indicate the community (fast greedy community detection)

Data Structure for Network Data

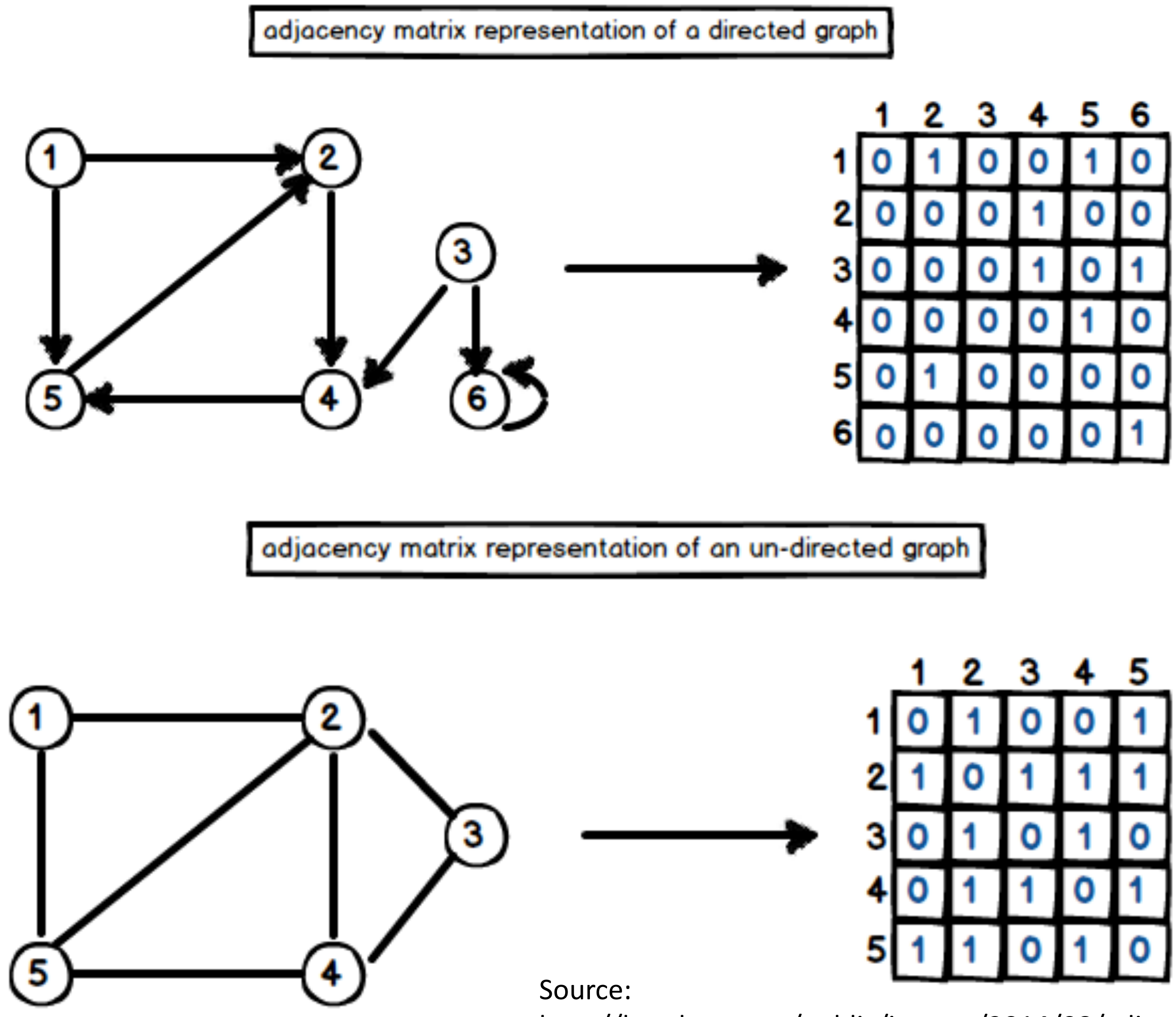
- Edge list (Gephi, R, igraph, Statnet)
- Adjacency matrix (R, igraph, Statnet)

Edge list



Source	Target	Type	Weight (n/a)
Node 1	Node 2	Directed	
Node 1	Node 5	Directed	
Node 5	Node 2	Directed	
Node 2	Node 4	Directed	
Node 4	Node 5	Directed	
Node 3	Node 4	Directed	
Node 3	Node 6	Directed	
Node 6	Node 6	Directed	

Adjacency matrix



adjacency matrix representation of a directed graph

	1	2	3	4	5	6
1	0	1	0	0	1	0
2	0	0	0	1	0	0
3	0	0	0	1	0	1
4	0	0	0	0	1	0
5	0	1	0	0	0	0
6	0	0	0	0	0	1

adjacency matrix representation of an un-directed graph

	1	2	3	4	5
1	0	1	0	0	1
2	1	0	1	1	1
3	0	1	0	1	0
4	0	1	1	0	1
5	1	1	0	1	0

Source:
http://buraktas.com/public/images/2014/08/adjacency_matrix_representation.png

Packages for network analysis

- Gephi
- NodeXL (Excel-based, for windows only)
- R (igraph, statnet)
- Python (NetworkX)

The Potentials of SNA

- A network perspective helps to examine the hidden forces of social changes
- How to construct the network requires theories in social sciences – and sometimes, imagination
- SNA detects the most important actors within a constructed network – journalists are watching those who have "power" in the society.

Further resources on SNA

- Open network data
  - SNAP [Stanford Large Network Dataset Collection]
  - UCI Prof Freeman [Datasets]
  - Network repository [Datasets]