Introduction of Network Analysis for Digital Trace Data

Taehee Kim

December 8, 2020

Social network analysis

Relations between individuals matter for individuals!

- Does obesity spread through social networks like a virus?
- Do disconnected networks cause loneliness and depression?
- Do students who are well integrated perform better?

Digital trace data and network analysis are compatible

In many cases, we observe relations in digital trace data which are quite informative

What is network data?

Network Data

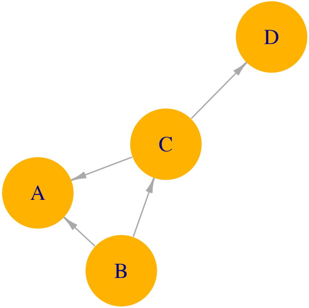
Dyads

Pairs of objects forming a compound of two objects

Network data

- We say that two dyads overlap, if they share a member. Network data are..
- the units of observation are dyads
- these dyads are overlapping

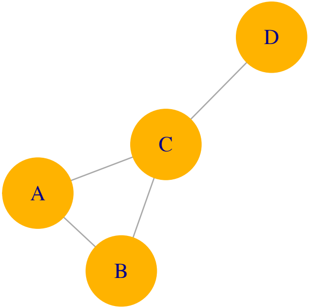
Basic terminology



- Node or Vertex
- Edge or Tie
- Degree (in, out)

Network representation

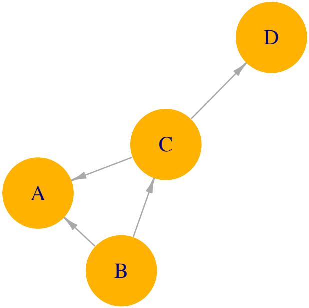
Undirected



$$G_U = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Row sums? $(2, 2, 3, 1) \rightarrow$ degree sequence

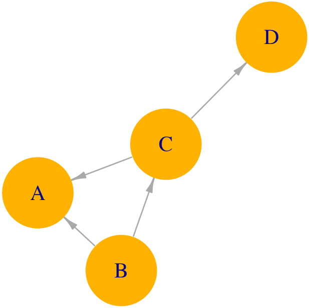
Directed



$$G_U = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Row sums? out-degree Column sums? in-degree

Directed



D
(b, a)
(b, c)
(c, a)
(c, d)

Centurality

Concept

Centrality

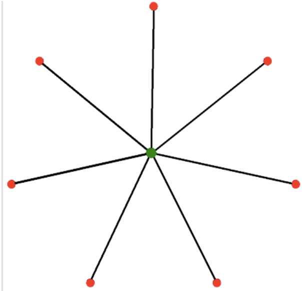
- One of the most important concepts in network analysis is the idea that structural position – where a node is in a network – limits or enhances access to information or other resources (Bernard 2013, 454).
- Centrality is a property of a node's position in a network (Borgatti, Evertt & Johnson 2013, 164).

The key questions

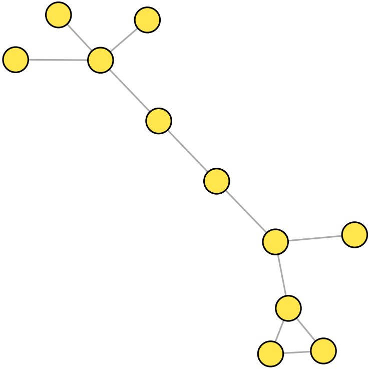
- How do we specify which indices should be considered cenralities?
- Which index should be used in a particular application context?

The agreement so far..

- the center of a star graph should be the most central vertex for any such index (Freeman 1979)



Which node is the center of the graph?



Centrality measures

Degree centrality

- A traditional indicator
- How many edges does a node have?
- A node is central if it has a high degree
- Out-degree, in-degree or the sum of both

For a simple undirected graph $G = (V, E)$,

$$c_D(i) = deg(i)$$

Closeness

- How short are the distances to all other nodes?
- Radial centrality
- Inverse of the sum over all dyadic distances ($d(i,j)$)

$$c_C(i) = \frac{1}{\sum_{j \in V} d(i,j)}$$

Betweenness

- How many shortest paths go through a certain node?
- Medial centrality

$$c_B(i) = \sum_{s \neq t \neq i \in V} \delta(s, t|i)$$

$$\delta(s, t|i) = \frac{\sigma(s, t|i)}{\sigma(s, t)} = \frac{\text{number of shortest st-paths via i}}{\text{number of shortest st-paths}}$$

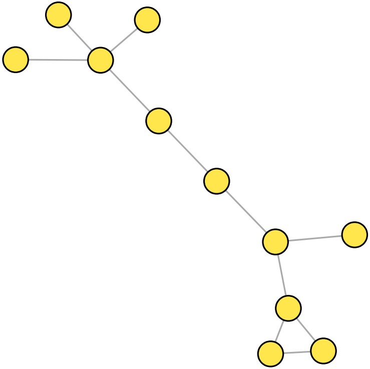
Feedback Centralities (Eigenvector)

- A central node has central neighbors
- An actor's centrality is proportional to the total centrality of tis neighbors.
- Feedback centrality

$$c_E(i) = \alpha \sum_{j \in N^-(i)} c_E(j)$$

- In undirected networks, $\alpha = 1$ and is equivalent to degree.

Which node is the center of the graph?



Which node is the center of the graph?

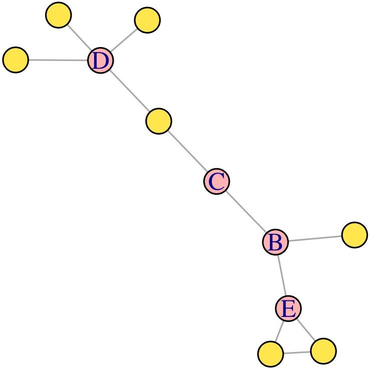


Figure 1: D: Degree, C: Closeness, B: Betweenness, E: Eigenvector

Clustering

Communities in Network

Why it is important?

- Community structure can affect individuals, group, and network

Communities

- Communities are densely connected internally and separated externally

Clustering

- Clusters: Groups that are internally more cohesive than externally
- Clustering: Particion the graph into several clusters

Definition

Let $G = (V, E)$ be graph. A partition $C = \{C_1, ..., C_k\}$ with $V = \biguplus_{i=1}^{k} C_i$ and $C_i \neq \emptyset$ is called a clustering of G.

Modularity

Definition

For a simple undirected graph $G = (V, E)$ and clustering $C = \{C_1, ..., C_k\}$,

$$Q(C) = \left[\sum_{i=1}^{k} \frac{|E(C_i)|}{m} - \left(\frac{\sum_{v \in C_i} deg(v)}{2m} \right)^2 \right]$$

$$= \left[\sum_{i=1}^{k} \frac{|E(C_i)|}{m} - \left(\frac{|E(C_i)| + \sum_{j=1}^{k} |E(C_i, C_j)|}{2m} \right)^2 \right]$$

$Q(C)$ is called the modularity of C.

- $E(C)$ be the set of edges in the induced graph $G[C]$.
- $E(C_1, C_2) = \{\{v, w\} : v \in C_1, w \in C_2\}$, the edges connecting $C_1, C_2 \subset V$

- Modularity ranges between -1 and 1 (undirected: between $-\frac{1}{2}, 1$).
- Modularity is used to evaluate the communityness of a network, and to compare alternative classifications
- But how to find clustering pattern which maximize modularity?
- Modularity maximization turns out to be *NP*-hard problem (Brandes et al., 2008) $\rightarrow$ heuristics are used to find clusterings with good modularity

Clustering Algorithms

- Fast-Greedy Algorithm
- Betweenness Algorithm

Fast-Greedy Algorithm

- Start with an empty graph in which each node is its own community.
- Evaluate which merger of communities would lead to the maximum increase in modularity.
- Merge these communities.
- Return to step 2 until: Merging of communities results in a decrease in modularity.

  Pro's : Quick and possible to use on larger graphs. Gives a dendrogram of the hierarchical clustering.

  Con's : Not good at detecting smaller communities, tends to quickly form larger communities.

Edge betweenness Algorithm

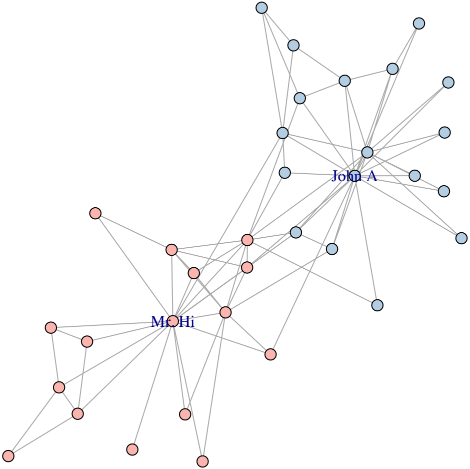
The idea of the edge betweenness based community structure detection is that it is likely that edges connecting separate modules have high edge betweenness as all the shortest paths from one module to another must traverse through them.

- Evaluate the edge-betweenness of each edge in the network.
- Find the edge with the highest edge-betweenness.
- Delete this edge.
- Repeat steps 1-3 until: No better quality function (i.e., modularity) value is obtained.

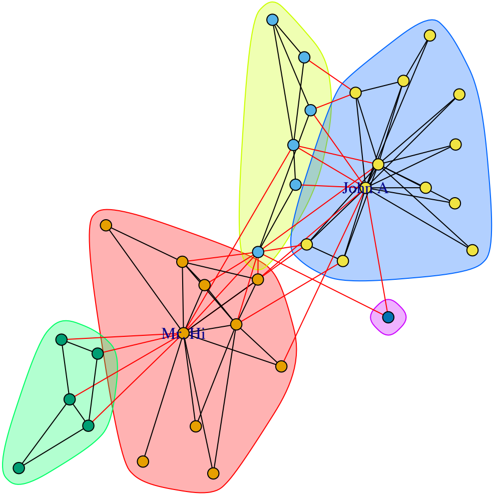
  Pro's : No random variation.

  Con's : Does not necessarily maximize modularity. Calculation of betweenness is slow.

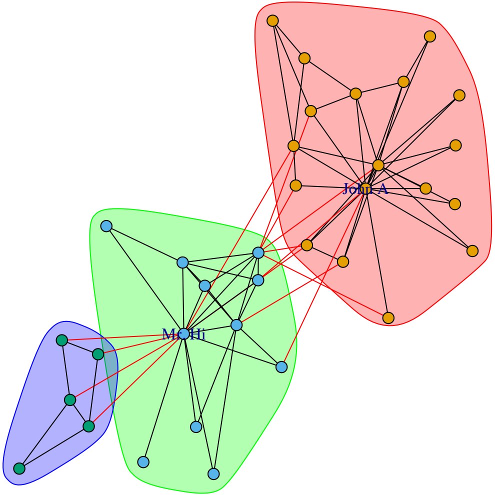
Zachary's Karate Club



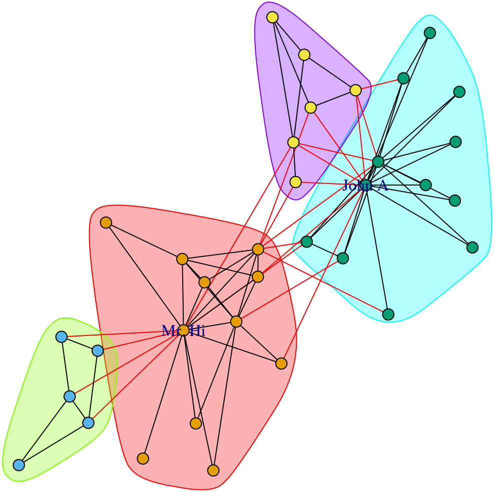
Edge betweenness Algorithm



Fast greedy Algorithm



Moduraity optimization



Detecting community is difficult!

- Many proposals, many sensible approaches, but hard to know what will be the best for your purpose.
- Main criticism: what is the meaning, relevance, function of the groups we find?