ComplexNetworks CS5656

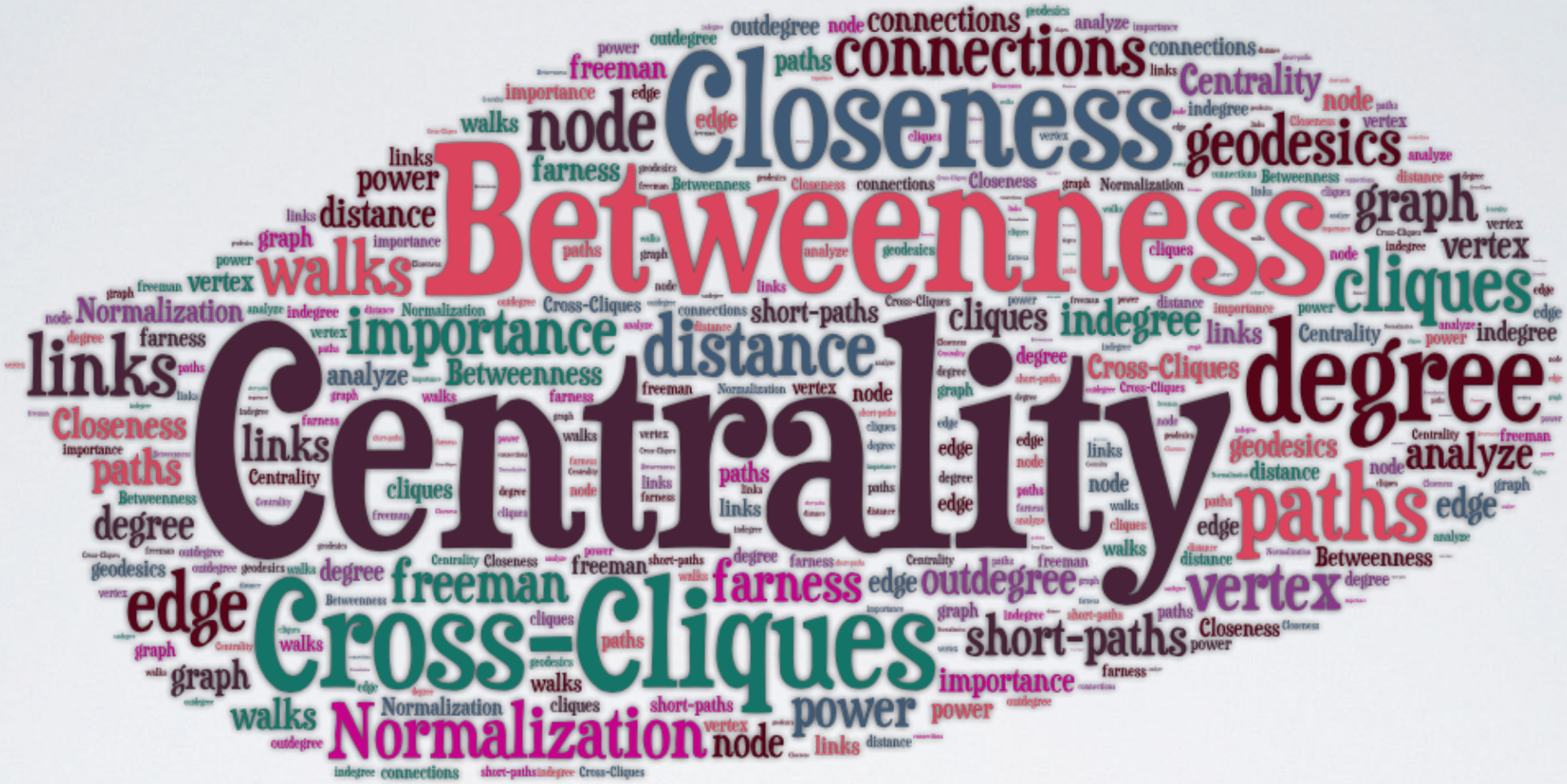
CENTRALITY

Degree, Betweenness, Closeness, Cross-Clique

and their applications

Abduljaleel Al Rubaye

September 2014

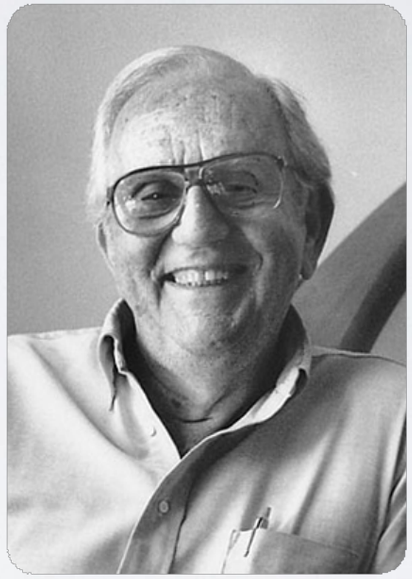


Which node is the most *important* one within a network ?

The History of Centrality

- *Alex Bavelas* (1948) and *Harold Leavitt* (1951) were the first scientists (psychologists) that used **centrality** in social network analyzing

Alex Bavelas
no picture



Harold Leavitt
src : google.com

- They wanted to explain the performance of communication networks
- That was the first glance that led to other experimental researches on network structure

- Centrality has been used in other researches in various areas:
  - Measuring the influence of a person in a social network
  - Investigating the competences in formal organizations
  - Employee opportunities
  - Measuring the power in an organization

- Scientists proposed different interpretations for centrality measures:
  - power
  - risk
  - influence
  - independence
  - control
  - exposure
  - belongingness
  - …

- Despite the differences in interpretations, all were agree that centrality is a ***node-level*** construct

- At that time the important question was :

  - What do all measures have in common?

- ***Linton Freeman*** (1979) reviewed some of the previous published measures. He reduced them to 3 basic concepts:

Degree ***Betweenness*** ***Closeness***

We will discuss these measures in addition to the measure:
Cross-Clique Centrality

Linton Freeman
src: google.com

Finding

“the *most important one*”

is depends on how we define the centrality

Centralities can be categorized in 2 groups

Radial

Volume:

Counting the links of a node

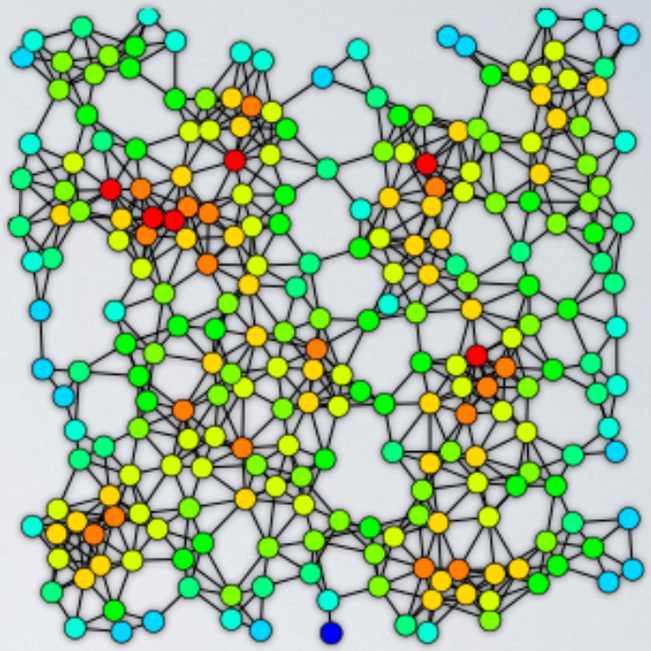
Length:

Capturing the distance between nodes

Medial

Counting the number of paths that pass through a given node

Degree Centrality



src: google.com

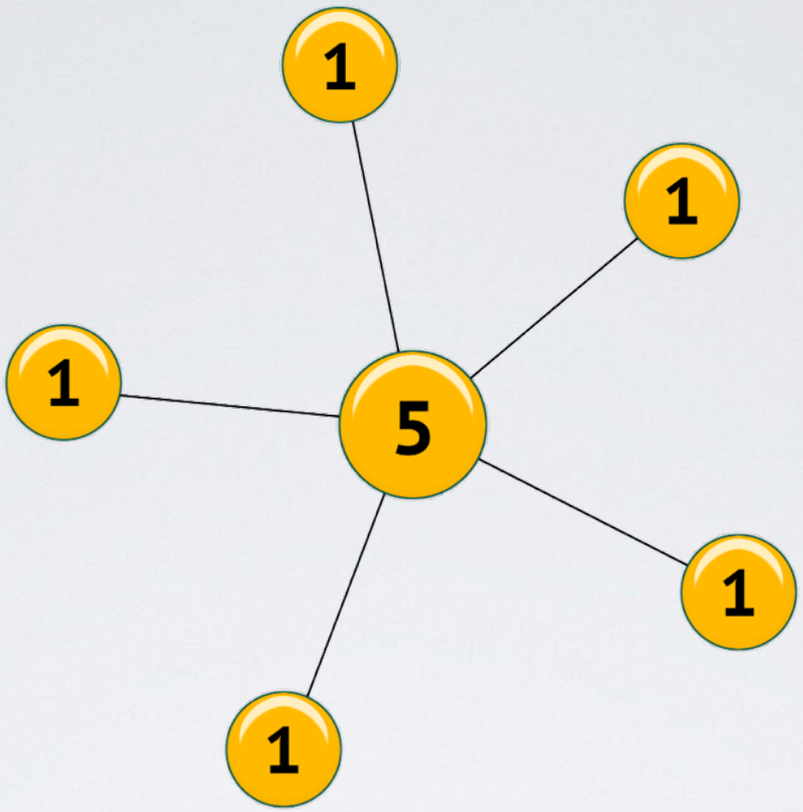
- The simplest and the first measure of centrality
- Degree centrality of a node is the number of connections of that node
- Defined as :

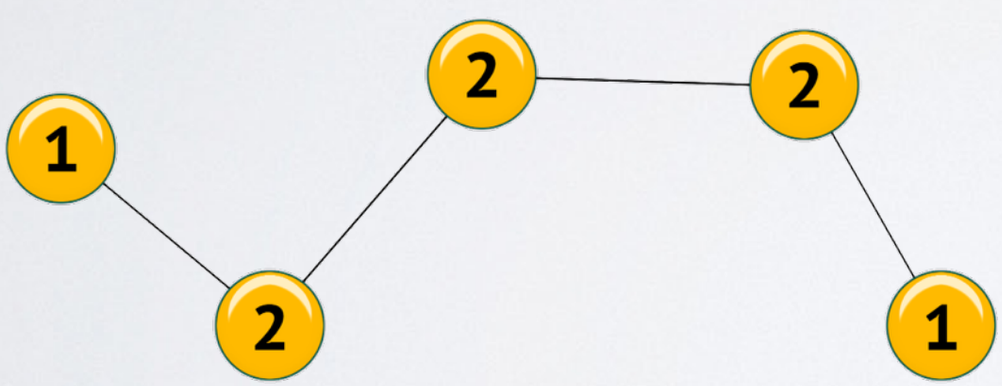
$$C_D(n_i) = d(n_i)$$

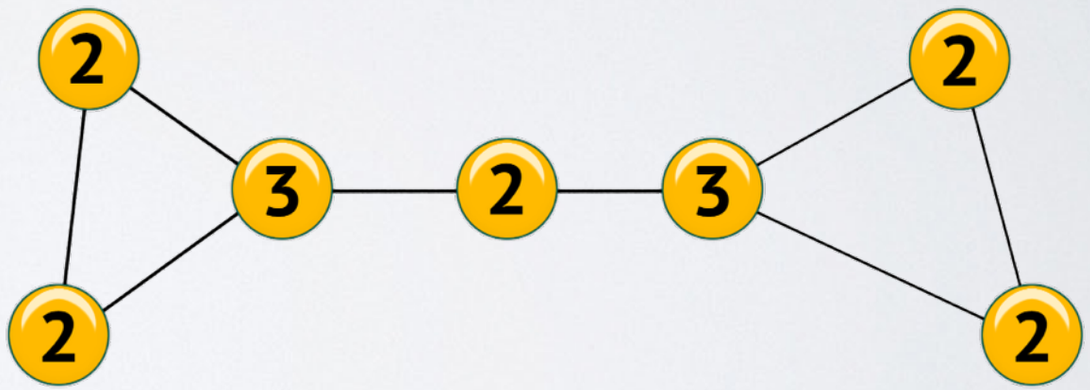
- For example:

  A person with many relatives and friends can be seen as an important person

Degree Centrality

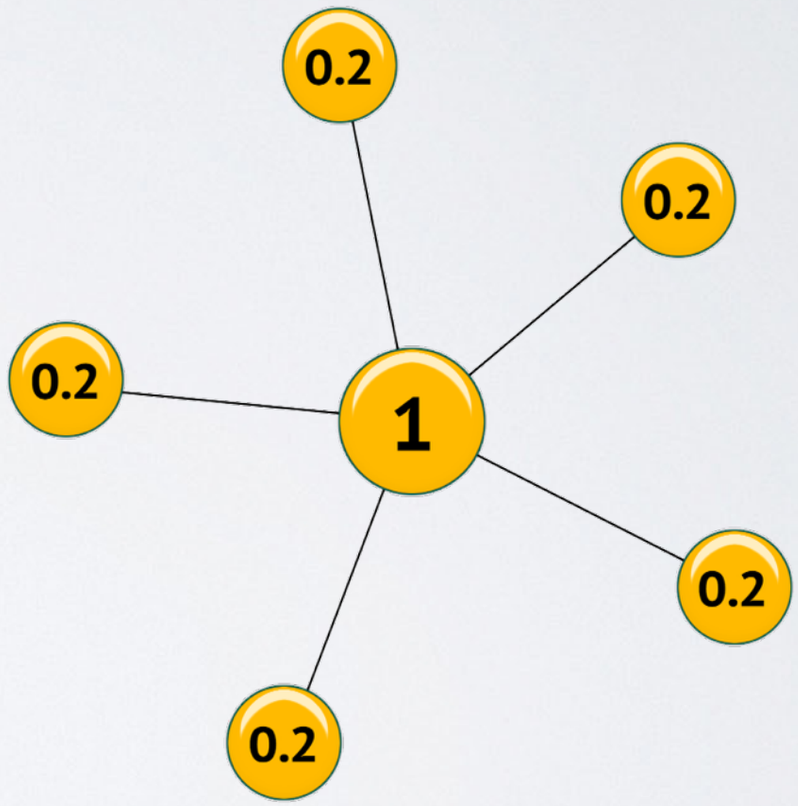


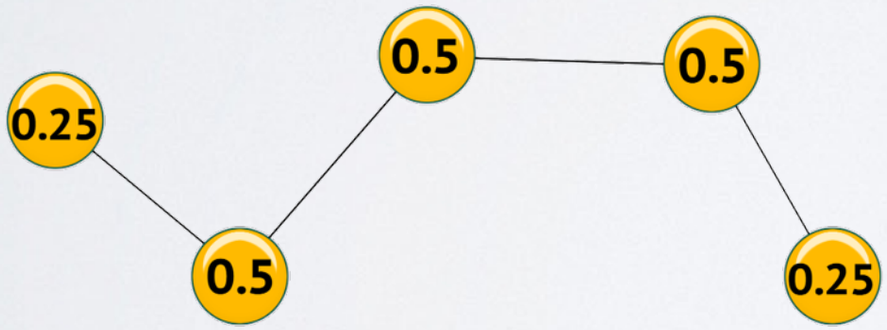




Normalized Degree Centrality :

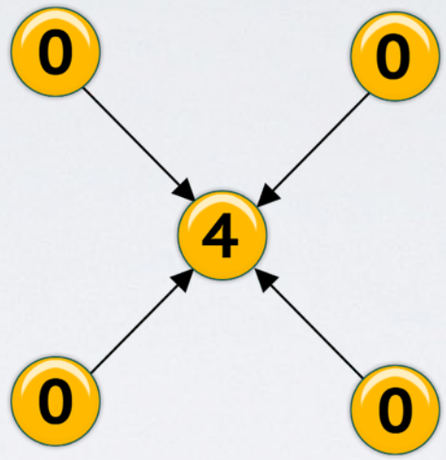
Degree Centrality can be normalized by dividing by *(N-1)* the max possible number of degree for a given node.



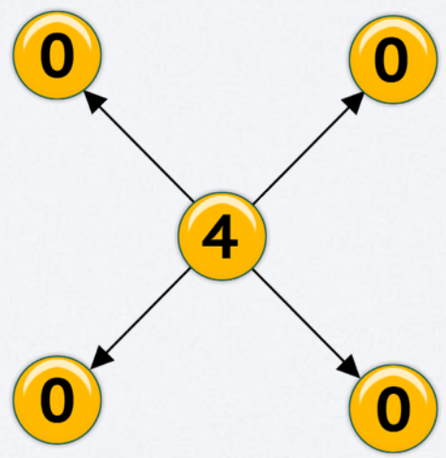


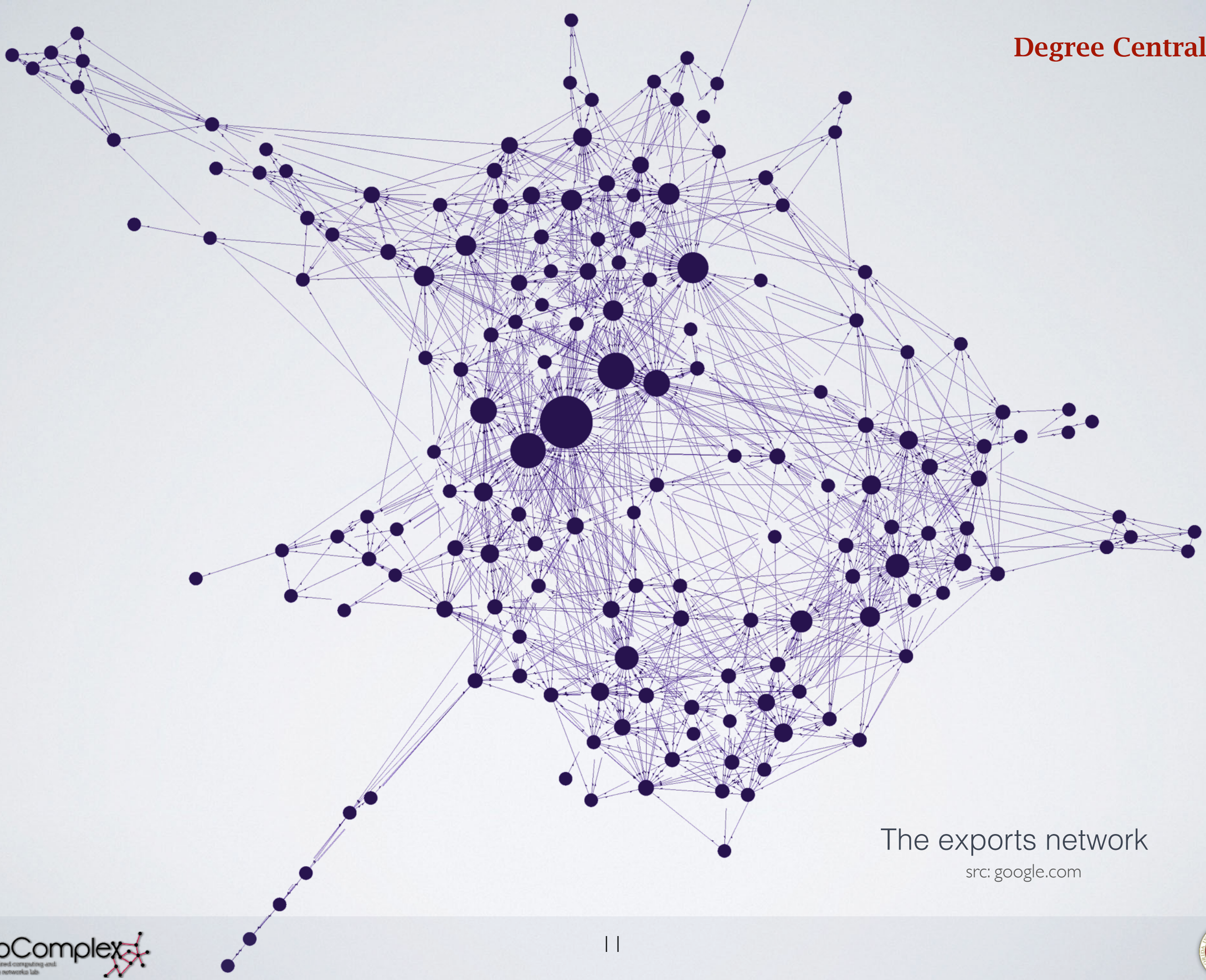
- The degree centrality for a ***directed graph*** is defined as:

  1. ***Indegree*** centrality : (# of incoming links for a node)



  2. ***Outdegree*** centrality (# of outgoing links for a node)

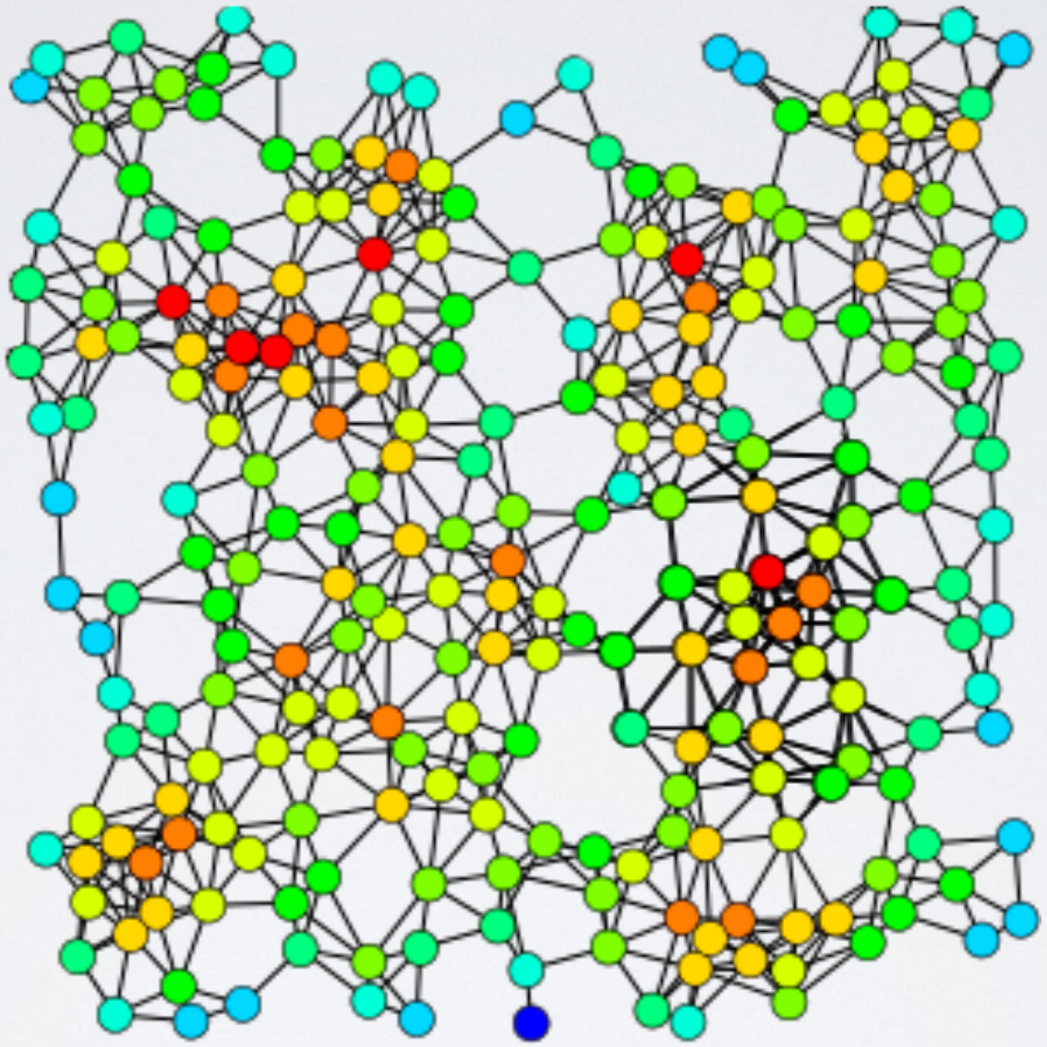




The exports network

src: google.com

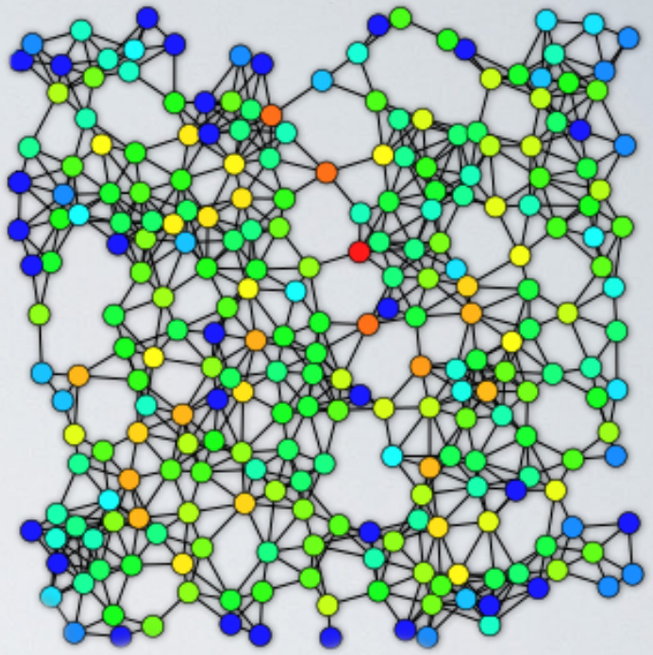
Is counting the links for a given node is always enough to find the central node within a network ?



src: google.com

A node with high degree centrality value cannot always be more central than a node with a lower degree

Betweenness Centrality



src: google.com

- Freeman was the first one who introduced the concept of the betweenness centrality in the1970s to study social networks

- The importance of a node relies more on how much the node is used to join any other nodes

- The idea of betweenness for a given node *(i)* is that how many pairs of nodes in the network are connected through the shortest path to each other passing that node *(i)*.

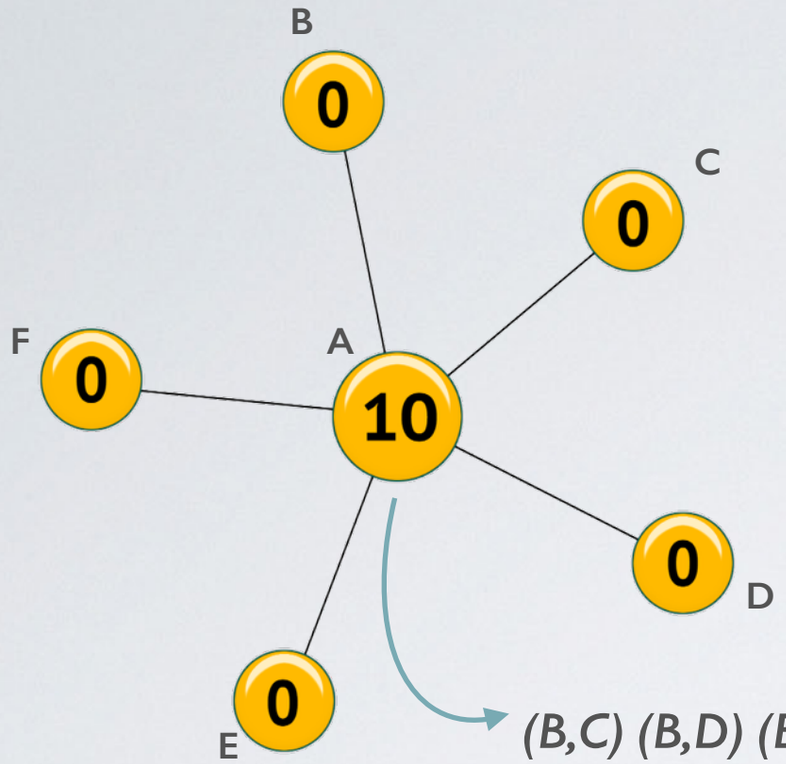
- To calculate the betweenness for a given node *(i)* :
  - how many short paths are there between each two nodes except node *(i)*
  - how many one of these paths are passing through the node *(i)*

- The formula is :

$$C_B(i) = \sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}$$

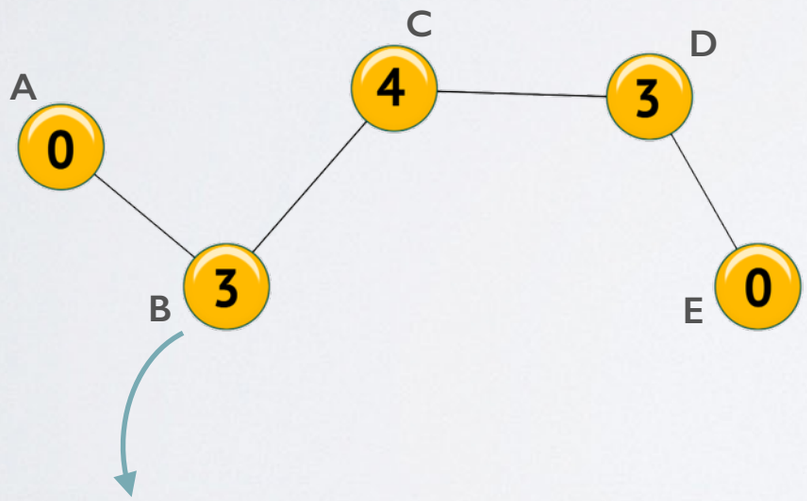
- Based on Freeman conception:

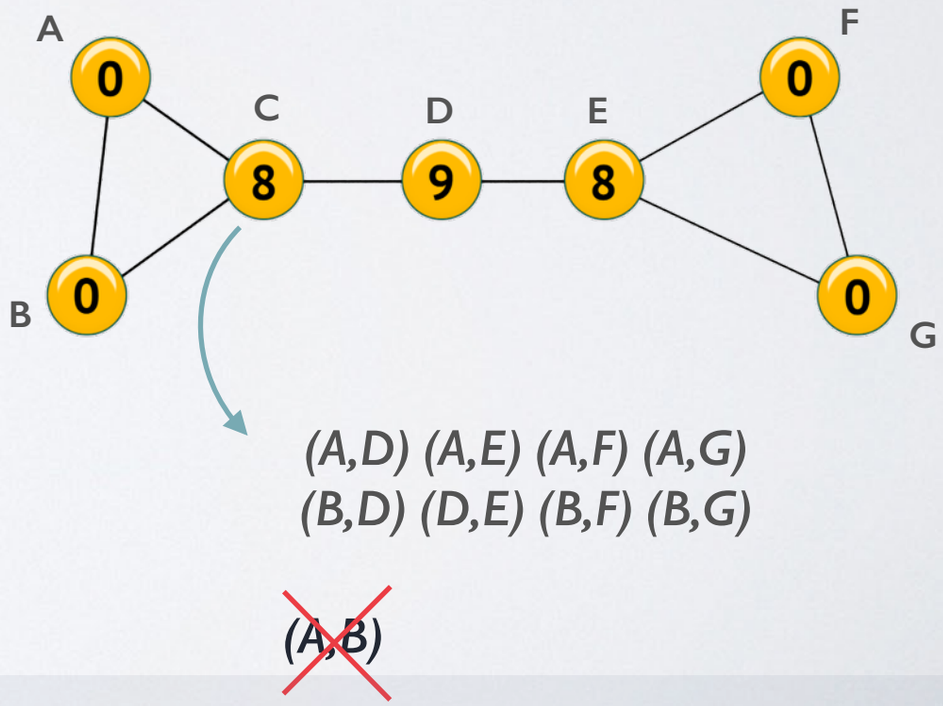
  The nodes that have high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have high betweenness value.



The highest possible value of betweenness is:

$$\frac{(n-1)(n-2)}{2} \quad \text{or} \quad \binom{n-1}{2}$$





Normalized Betweenness Centrality :

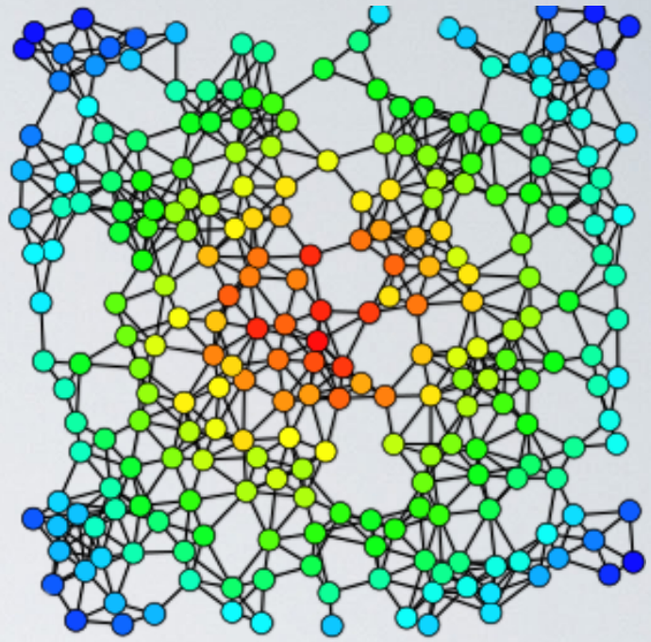
$$C'_B(i) = \frac{C_B(i)}{\left[(n-1)(n-2)\right]/2}$$

$\left[(n-1)(n-2)\right]/2$ is the max number of paris of nodes in an undirected graph except node *(i)*

$(n-1)(n-2)$ is the max number of paris of nodes in a directed graph except node *(i)*

Closeness Centrality

- A node is considered “important” if it is relatively close to all other nodes.



src: google.com

- Beauchamp (1965) (is a scientist in Business and Information systems) says:

> *“Closeness centrality is based on the idea that nodes with short distance to other nodes can spread information very productively through the network”*

src: Landherr, A., Friedl, B., Heidemann, J. (2010) “A Critical Review of Centrality Measures in Social Networks”.

- To calculate the closeness lets define the farness
- The *farness* of a node *(i)* is equal to the sum of shortest paths' lengths from the given node *(i)* to all other nodes.

$$Farness(i) = \sum_{j=1}^{n} d(i,j)$$

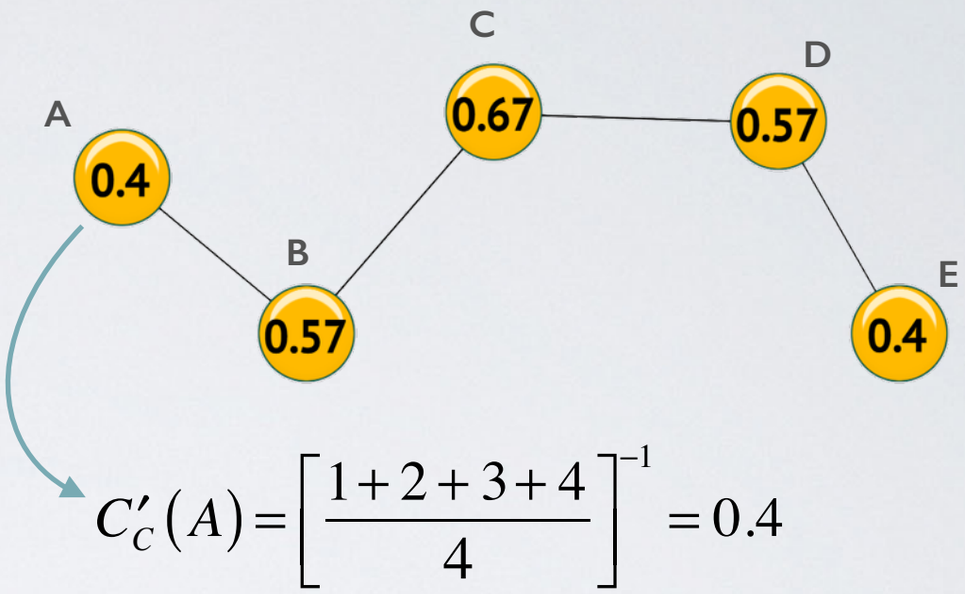
- The closeness of a node *(i)* is the inverse of the farness

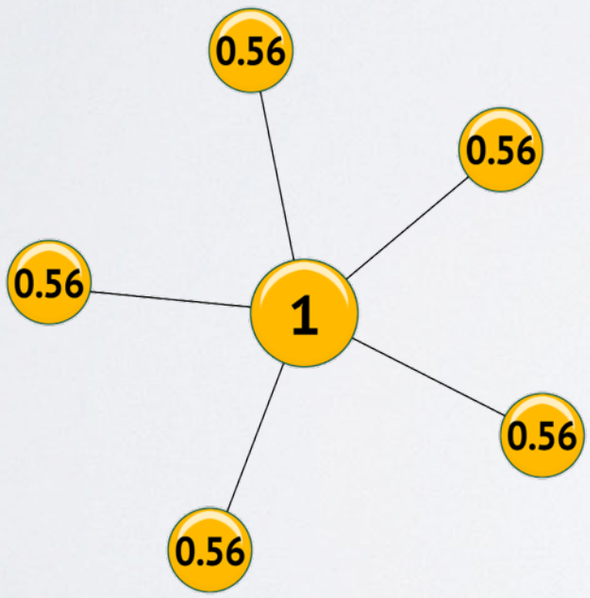
$$C_C(i) = \left[\sum_{j=1}^{n} d(i,j) \right]^{-1}$$

- The lower the total distance of a node to other nodes is, the higher its closeness centrality value.

Normalized Closeness Centrality :

$$C'_C(i) = \left[\frac{\sum_{j=1}^{n} d(i,j)}{(n-1)} \right]^{-1}$$



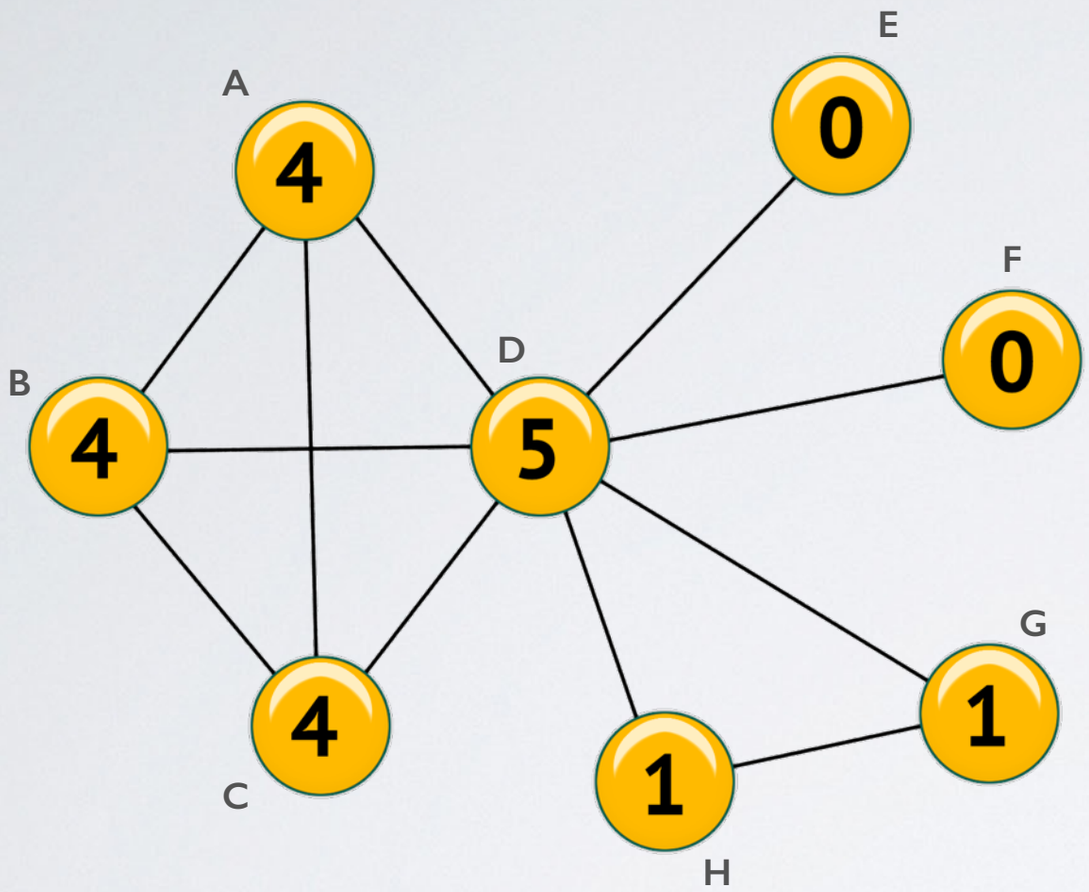


Cross-Clique Centrality

- Measures the connectivity of a node to different cliques

- For node *(i)* the cross-clique is denoted as ***X(i)*** which is equal to the number of cliques that node *(i)* belongs to

- *Imagine a person who has cross-connections in his community and would like to travel to another place. If a disease has been spread in his community, he will be a potential disease carrier. In such case this person should be monitored*



Lets consider the cliques with more than two nodes

{A,B,C} {A,B,D} {A,C,D} {B,D,C} {D,H,G} {A,B,C,D}

Node (D) belongs to 5 cliques

$$X(D) = 5$$

Centralization :

- We also can calculate the centrality for the whole graph
- Freeman's general formula for centralization :

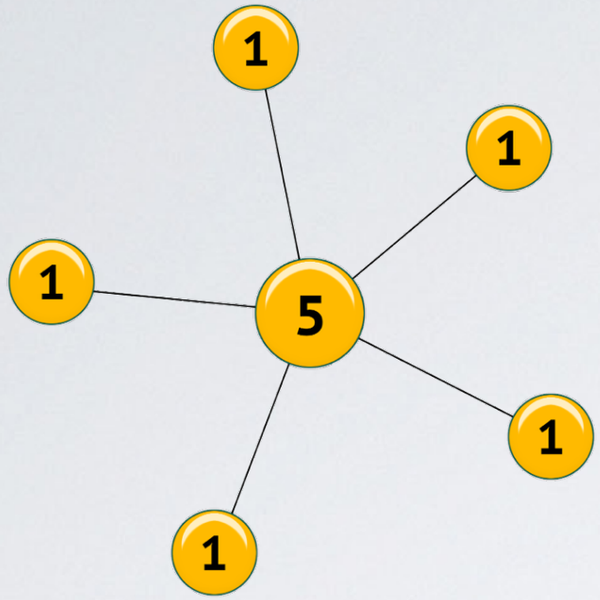
$$C_x = \frac{\sum_{i=1}^{g}\left[C_x\left(n^*\right) - C_x(i)\right]}{\left[(N-1)(N-2)\right]}$$

$C_x\left(n^*\right)$ the max value of centrality in the graph

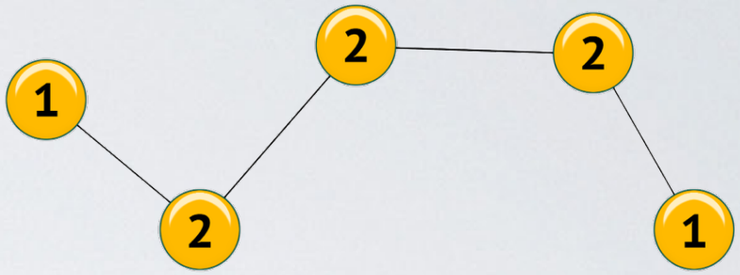
- The value of centralization is between 0 and 1:

0 : it means that the centrality of all nodes are equal (low centralization)

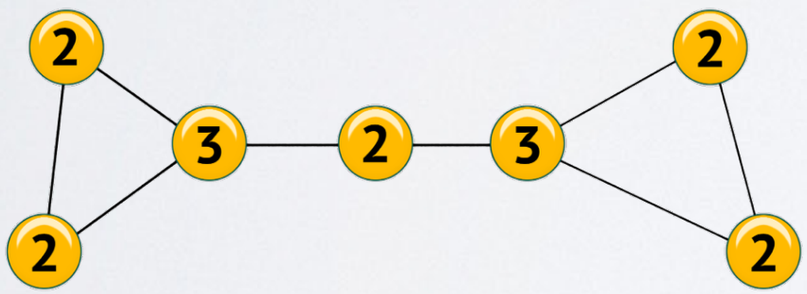
1 : it means that the shape of the graph is star-like. (High centralization)



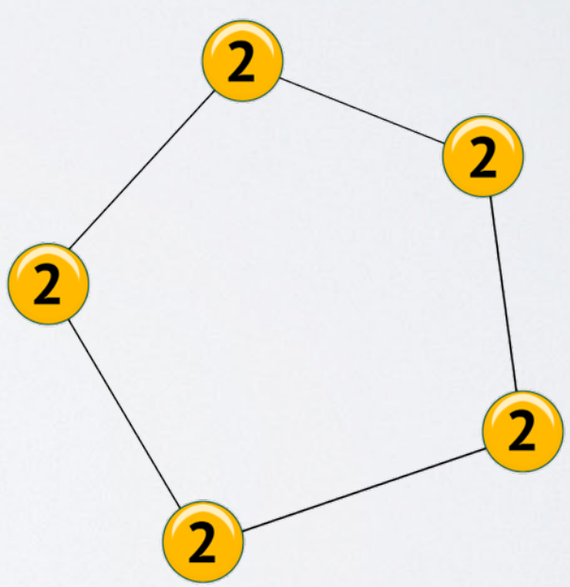
$$C_D = \frac{5(5-1)}{[5\times 4]} = 1.00$$



$$C_D \approx 0.167$$



$$C_D \approx 0.167$$



$$C_D = 0$$

Applications

Metro Systems Centrality

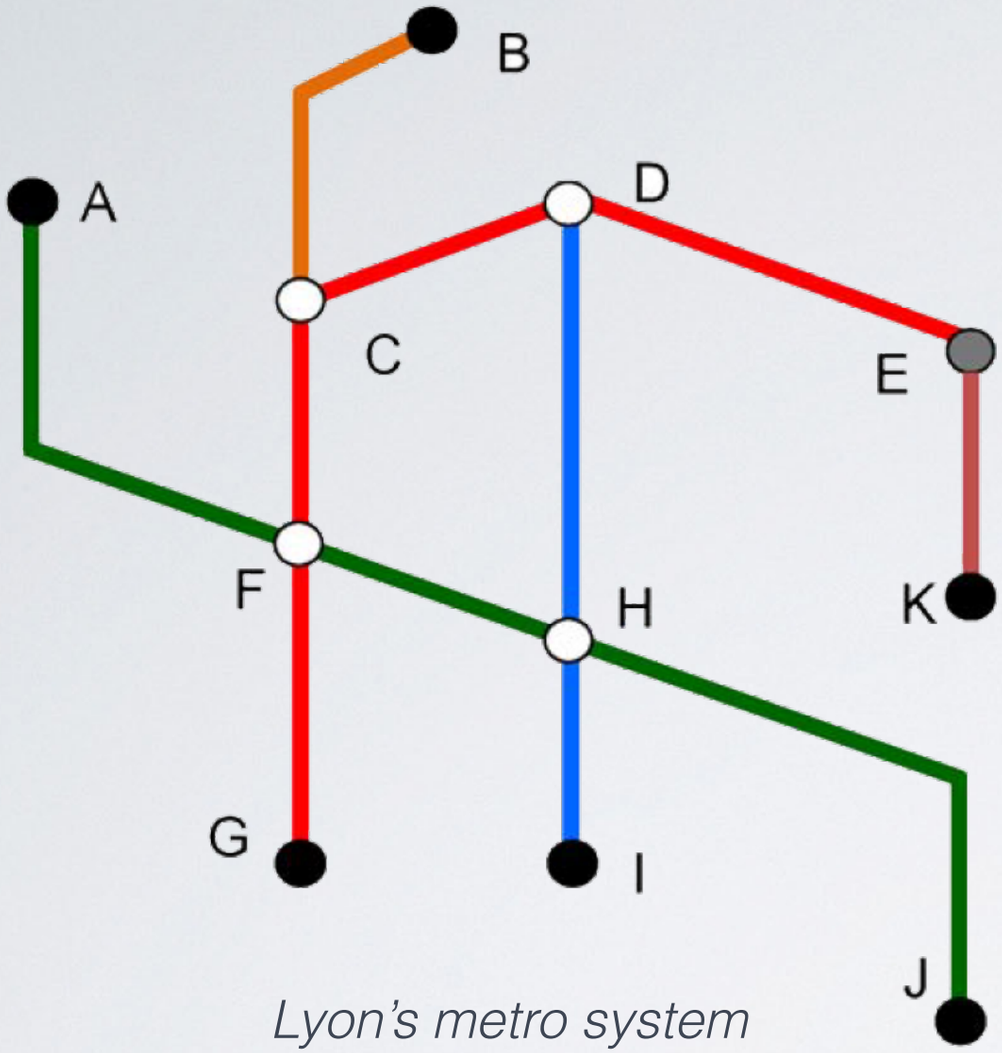
- *Since the role of public transport is increasing day after day, new methods and techniques are needed to better plan the systems.*

- *There is a study (Derrible S. 2012) tries to improve the work of public transit systems by examining an important aspect in transit network that is "centrality".*

- *In this study the researcher measured the betweenness centralities of 28 worldwide metro stations*

The study showed that:

- *In public transportation systems a station might be used heavily because its location **BUT** there might be another station used even more than that station because it is a transfer point to other locations.*



Lyon's metro system

src: Derrible, S. (2012).
"Network Centrality of Metro Systems"

- *Station F is located at Place Bellecour (a major shopping center) and is heavily used because of that.*
- *Meanwhile, station H is not located in a major area BUT it is also heavily used because it's a node to get to node F and node D.*

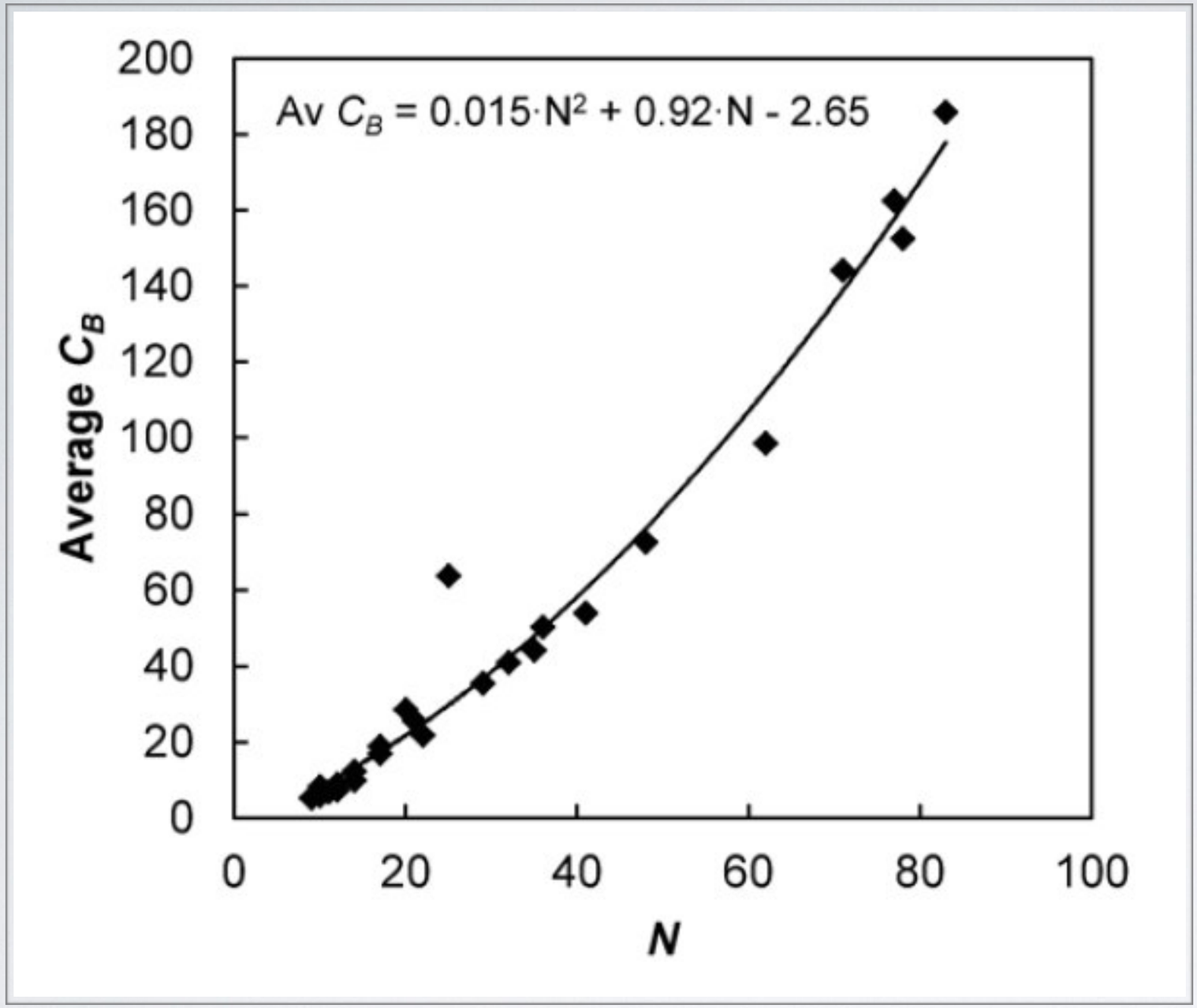
- $C_B(F) = C_B(H)$
- *H and Fare equally centered (both are "important" nodes).*

Applications

Metro	Nodes	Links	Betweenness Centrality C_B				Quadratic Coefficients	
			Min*	Max	Ave	Sum	$\|a_n\|$	$\|a_o\|$
Athens	9	9	7.00	15.00	5.33	48	0.03125	1.50
Brussels	9	9	7.00	19.00	5.11	46	0.03261	1.50
Lyon	10	10	11.00	18.00	5.80	58	0.03017	1.75
Montreal	10	10	11.00	18.00	5.80	58	0.01742	1.75
Toronto	10	9	8.00	26.00	8.20	82	0.03017	1.43
Bucharest	11	12	6.00	19.00	6.82	75	0.01643	1.23
Lisbon	11	11	13.50	21.50	6.82	75	0.02167	1.63
Singapore	12	13	10.00	26.00	8.92	107	0.01335	1.43
Buenos Aires	12	13	17.50	34.00	7.33	88	0.00426	0.38
Milan	14	15	23.00	39.00	12.36	173	0.00888	1.54
St Petersburg	14	16	22.50	25.00	9.93	139	0.00058	0.08
Hong-Kong	17	18	15.00	71.00	18.94	322	0.01046	3.37
Washington DC	17	18	19.00	71.00	16.94	288	0.00544	1.57
Stockholm	20	19	35.00	113.00	28.60	572	0.00920	5.27
Boston	21	22	37.00	102.00	25.62	538	0.00592	3.18
Shanghai	22	28	13.83	93.22	21.82	480	0.00441	2.12
Chicago	25	57	23.00	221.00	63.92	1598	0.00397	6.34
Barcelona	29	42	9.12	163.07	35.45	1028	0.00218	2.25
Berlin	32	43	22.10	110.40	40.97	1311	0.00166	2.17
Mexico City	35	52	12.42	129.22	44.20	1547	0.00111	2.06
Osaka	36	51	13.25	153.00	50.25	1809	0.00133	2.01
Moscow	41	62	25.52	177.26	54.10	2218	0.00106	2.36
Madrid	48	79	2.03	265.19	72.77	3493	0.00062	2.18
Tokyo	62	107	9.50	452.55	98.56	6111	0.00045	2.76
Seoul	71	111	21.41	467.57	144.10	10231	0.00034	3.50
New York City	77	109	10.74	683.15	162.45	12509	0.00036	4.46
Paris	78	125	40.07	630.73	152.50	11895	0.00025	3.02
London	83	121	7.69	1240.29	185.84	15425	0.00030	4.59

src: Derrible, S. (2012). "Network Centrality of Metro Systems"

- *The data shows that the betweenness is tend to increase by increasing the number of nodes*



src: Derrible, S. (2012).
"Network Centrality of Metro Systems"

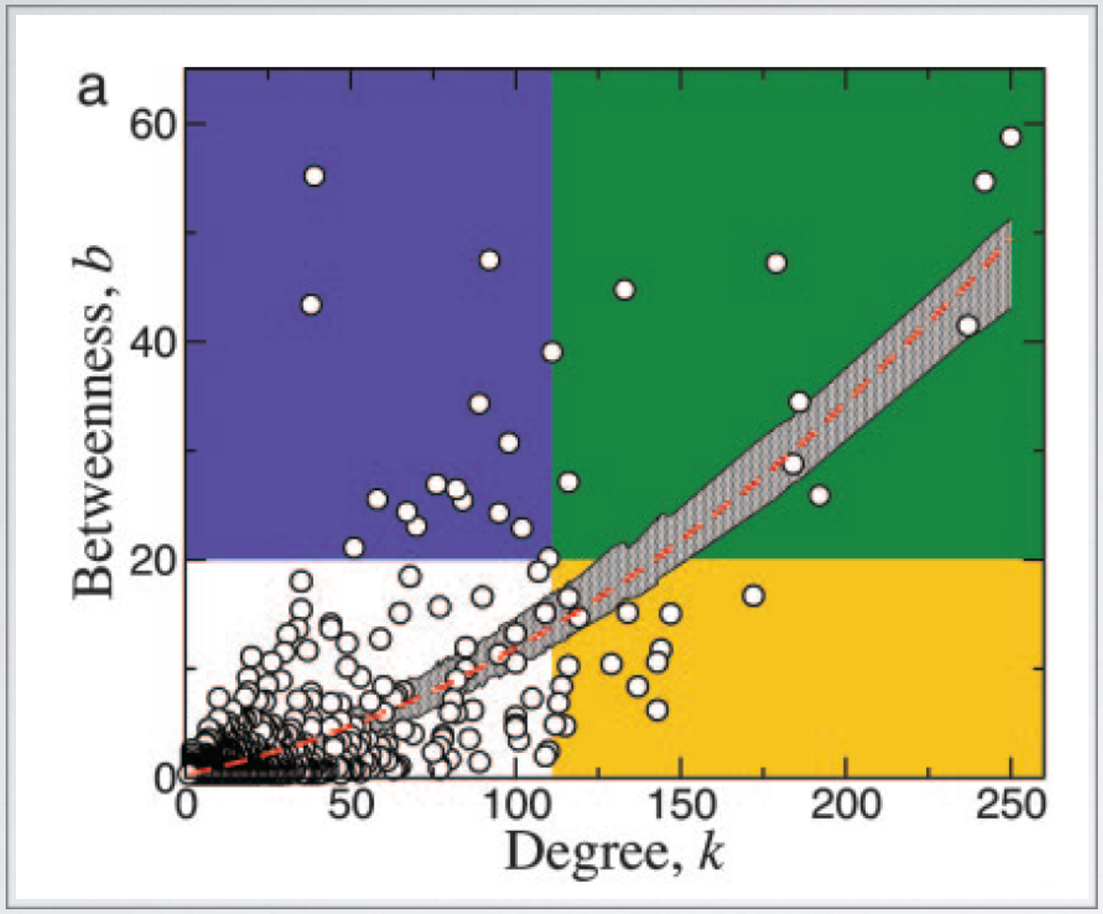
- *The increase of short-paths grows faster with the increase of the nodes*
- *Only one system (Chicago metro system) doesn't fit the regression*

The Worldwide Air Transportation Centrality

- *This is a study that analyzes the global structure of air transportation system.*

- *They measured the centrality of each airport using randomized-betweenness to compare between the most connected cities and the most central cities.*

- *The conclusion of the study shows that most of the connected airports aren't the most central.*



src: Guimerà, R., Mossa, S., Turtschi, A., & Amaral, L. A. N. (2005). "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles"

- *The betweenness is described as a quadratic function of the degree (dashed line). The study showed that 95% of all data falls in the gray area.*
- *HOWEVER, we can find some other cities that have low degree and high betweenness (the blue area)*
- *In contrast the yellow region shows that some of the airports are well connected but not well central.*
- *And there are just few airports that are highly connected and have high betweenness centrality (the green area)*

src: Guimerà, R., Mossa, S., Turtschi, A., & Amaral, L. A. N. (2005). "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles"

- *The most connected airports*

- *The most central airports*

Protection against worms in Online Social Networks

- Social networks are ideal places for malware creators to propagate worms through the network.

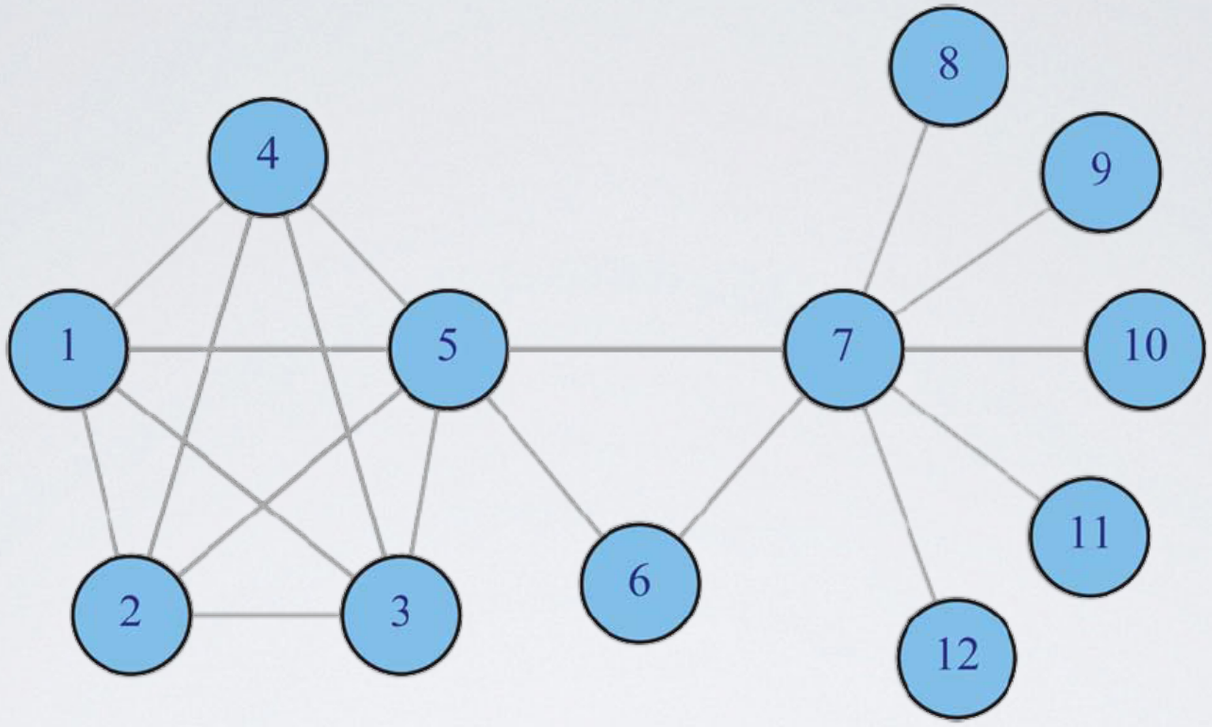
- Worms can hit the network and propagate faster through the central nodes.

  Example: first worm was called "SAMY" hit Myspace.com in 2005 and hit about one million accounts in 24 hours.

- This study is about detecting malware in a OSN.

- Start searching from the more central nodes (candidates) to scan the posts in order to detect any malicious code.

- Selecting more candidates will lead to detect a malware faster, considering that scanning each node will be done in a fixed period of time.



src: Faghani, M.R (2013). "A Study of XSS Worm Propagation and Detection Mechanisms in Online Social Networks"

Metric	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9	v_{10}	v_{11}	v_{12}
Degree	4	4	4	4	6	2	7	1	1	1	1	1
Closeness	0.47	0.47	0.47	0.47	0.68	0.55	0.73	0.44	0.44	0.44	0.44	0.44
betweenness	0	0	0	0	28	0	40	0	0	0	0	0
PageRank	0.09	0.09	0.09	0.09	0.14	0.06	0.23	0.04	0.04	0.04	0.04	0.04
Cross-Connectivity	11	11	11	11	12	1	1	0	0	0	0	0

- *Node (5) and node (7) are the best candidates to start monitoring*

- To see the effectiveness of the centrality measures in this study in term of malware detection time:

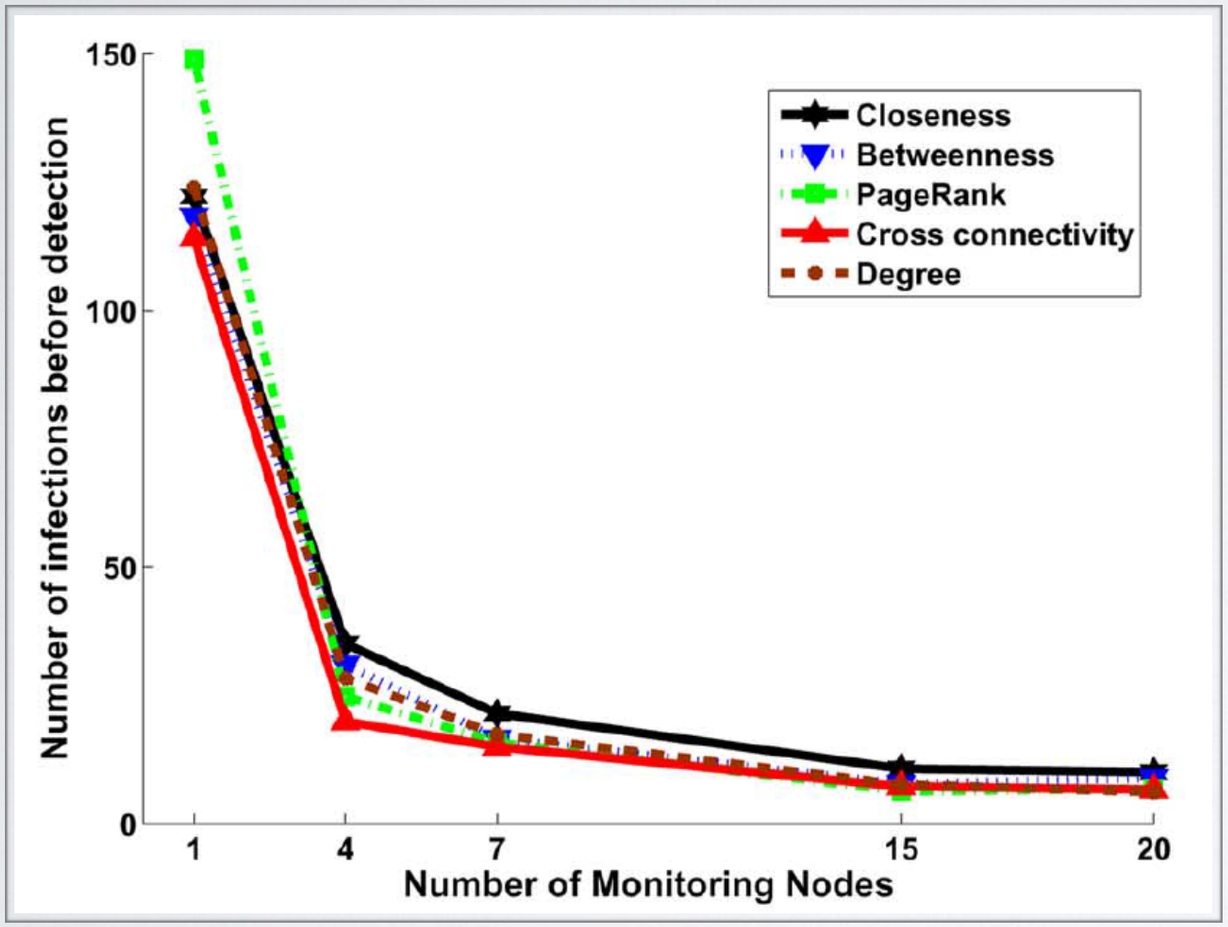
  - *Based on each one of the centrality measures, select the top 1,4,7,15,20 candidates, respectively.*

  - *The monitoring system starts checking these nodes' and their friends' activities.*

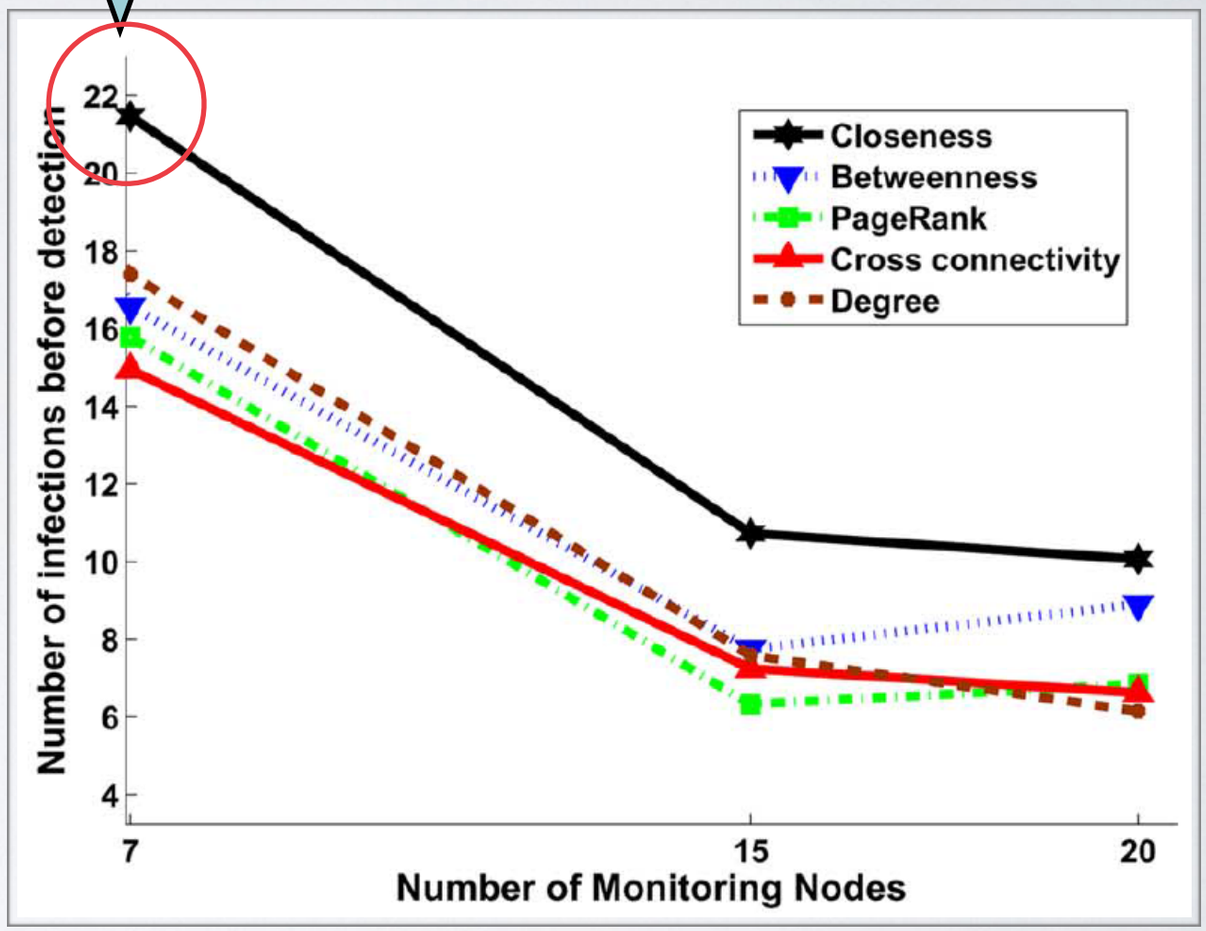
  - *When the first post that contains the malicious code is being detected, the system stops scanning.*

  - *Then the system counts the number of the nodes that being scanned and were infected at that point.*

- The Results for a network with 10,000 nodes

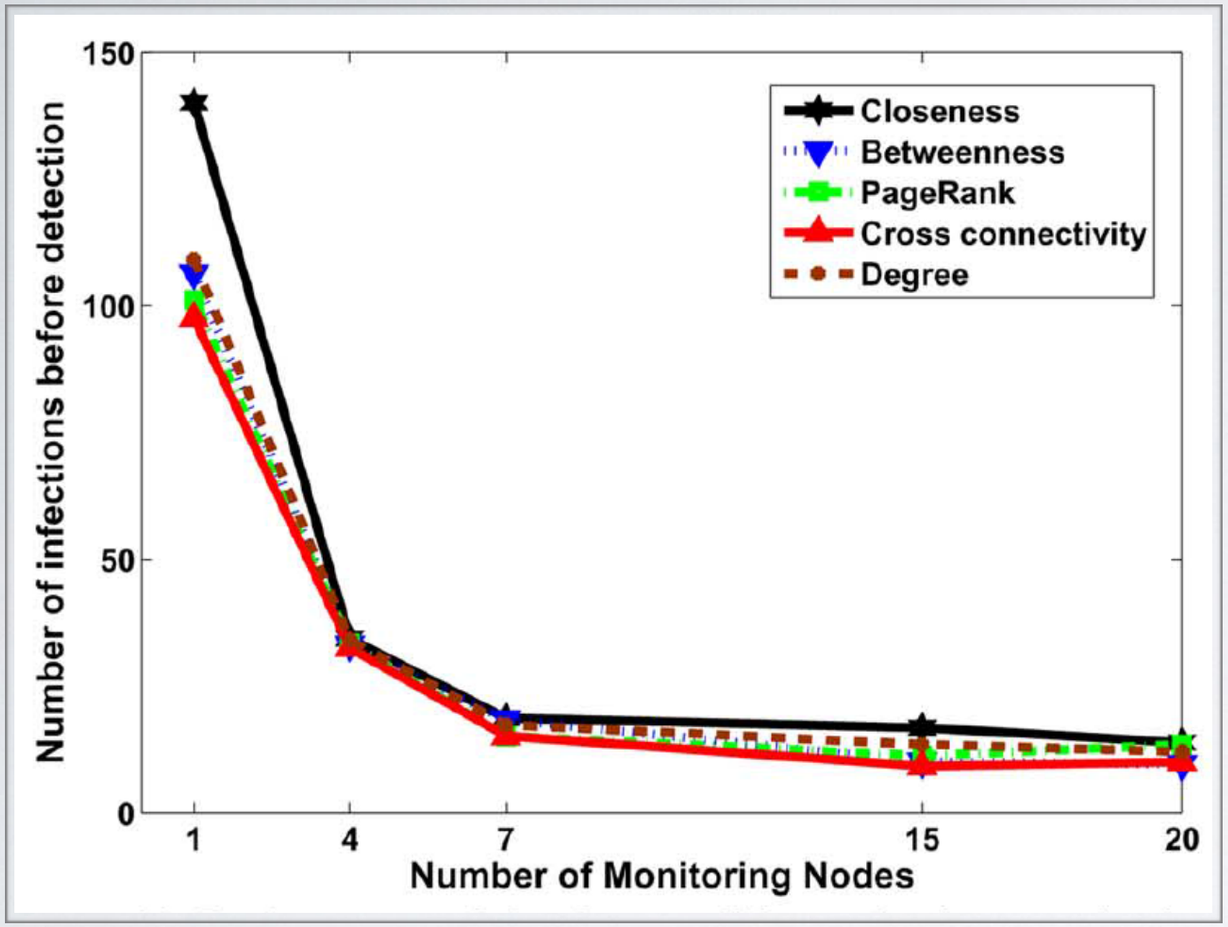


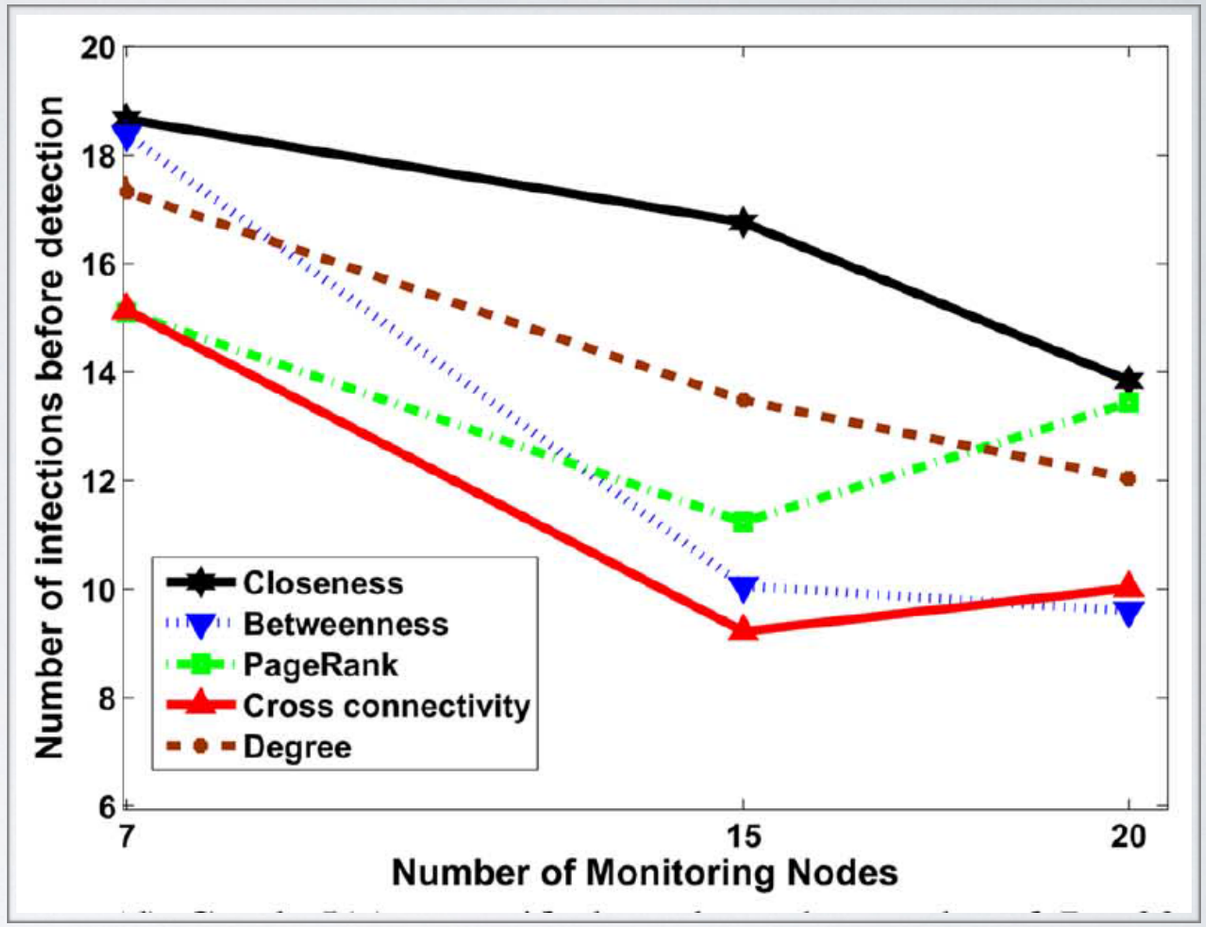
When 7 nodes were selected using CC, the system detected the malicious code after scanning 22 infected nodes



src: Faghani, M.R (2013). "A Study of XSS Worm Propagation and Detection Mechanisms in Online Social Networks"

- The Results for a network with 20,000 nodes





src: Faghani, M.R (2013). "A Study of XSS Worm Propagation and Detection Mechanisms in Online Social Networks"

- The performance of all centrality measures is almost same.

Social Graphs in Movies

- *Movie galaxies is an interesting project of discovering the social graph in movies.*

moviegalaxies

- *Many of networks are constructed out of movies' scripts.*
- *It's a visualization of characters interaction as a social graph.*
- *They used Betweenness and Degree centralities to show the relationship of each node to all the network's nodes*
- *http://moviegalaxies.com*

K-Path Centrality

- One of the variances of degree centrality
- Proposed by Sade (1989).
- It's basically counts all paths of length (k or less) that start from or end to a given node.
- When:
  A. k=1 (min value):
     the measure is identical to degree centrality.

  B. k = n-1 (max value)
     the measure counts the total number of paths of any length that originate at a given node.

Types of K-Path Centrality

1. Edge-disjoint k-path:
    - Counting paths that share no-edges and start from or end in a given node.
    - According to Ford and Fulkerson (1956) the minimum number of edges that must be deleted in order to disconnect two nodes is equal to the number of edge-disjoint paths that link two nodes.
    - So, edge-disjoint path centrality can measure the difficulty of separating two nodes.

2. Vertex-disjoint k-path centrality:
    - Counting paths that share no vertex (excepts two ends nodes).
    - According to Menger (1927) the number of vertex-disjoint paths is equal to number of nodes that must be removed in order to isolate two nodes.

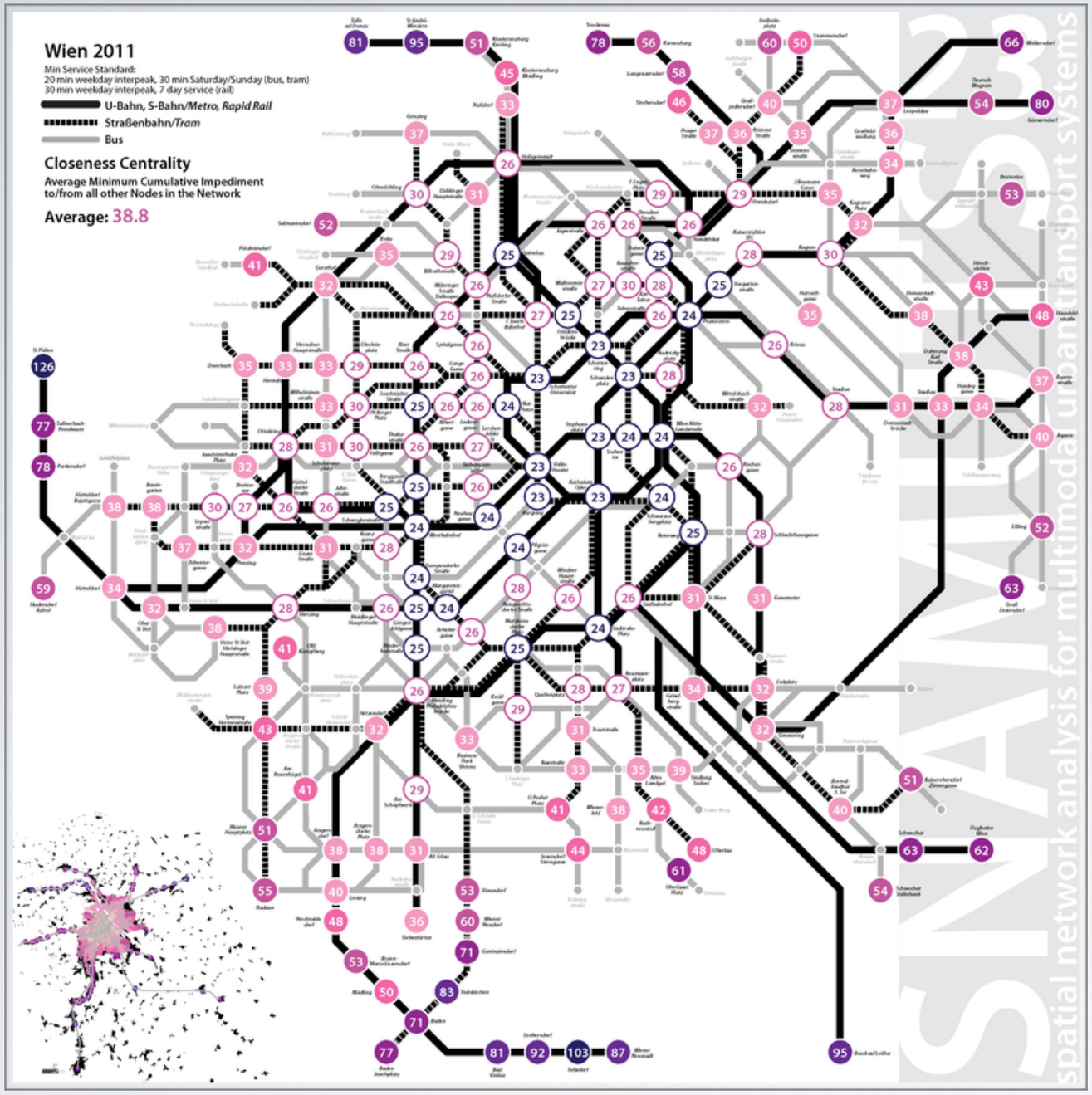
Urban Transport Systems

- *SNAMUTS is a project of analyzing networks of urban transport systems*

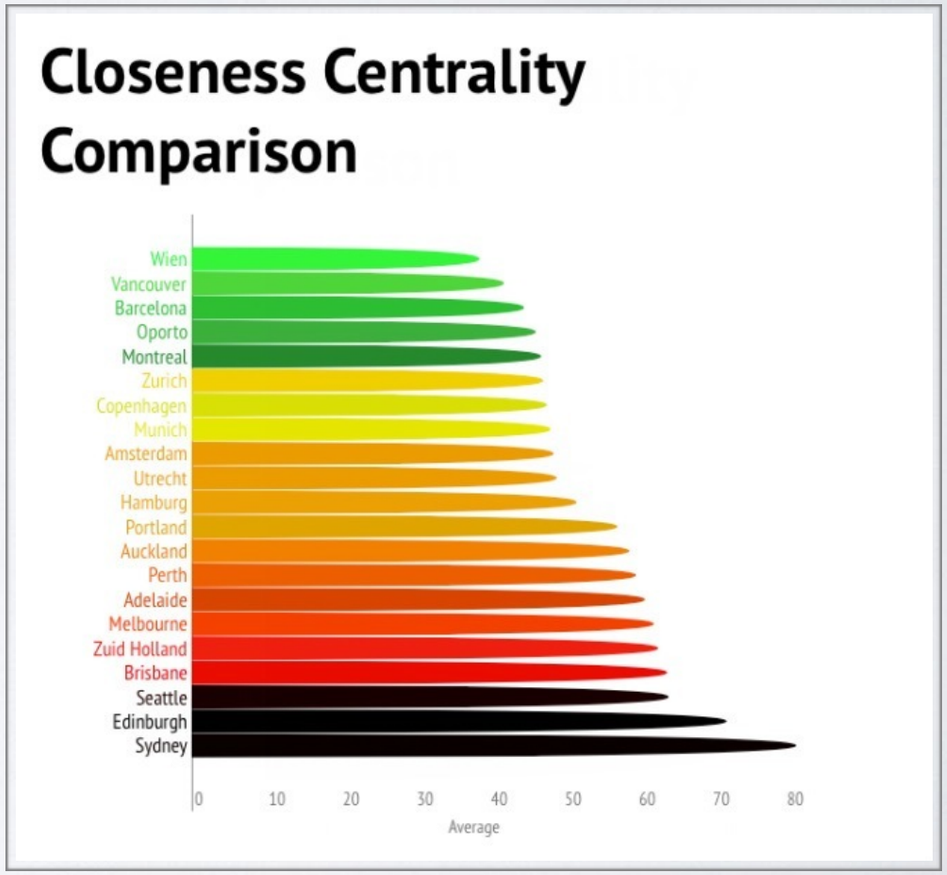
src: snamuts.com

- *This project compares between 21 cities transport system around the world (including bus, train, metro) to find which system has the best performance based on closeness centrality*

- *The closeness centrality describes how much easy is the movement along the network of public transportation in term of speed and the travel time*



src: snamuts.com



Questions ?