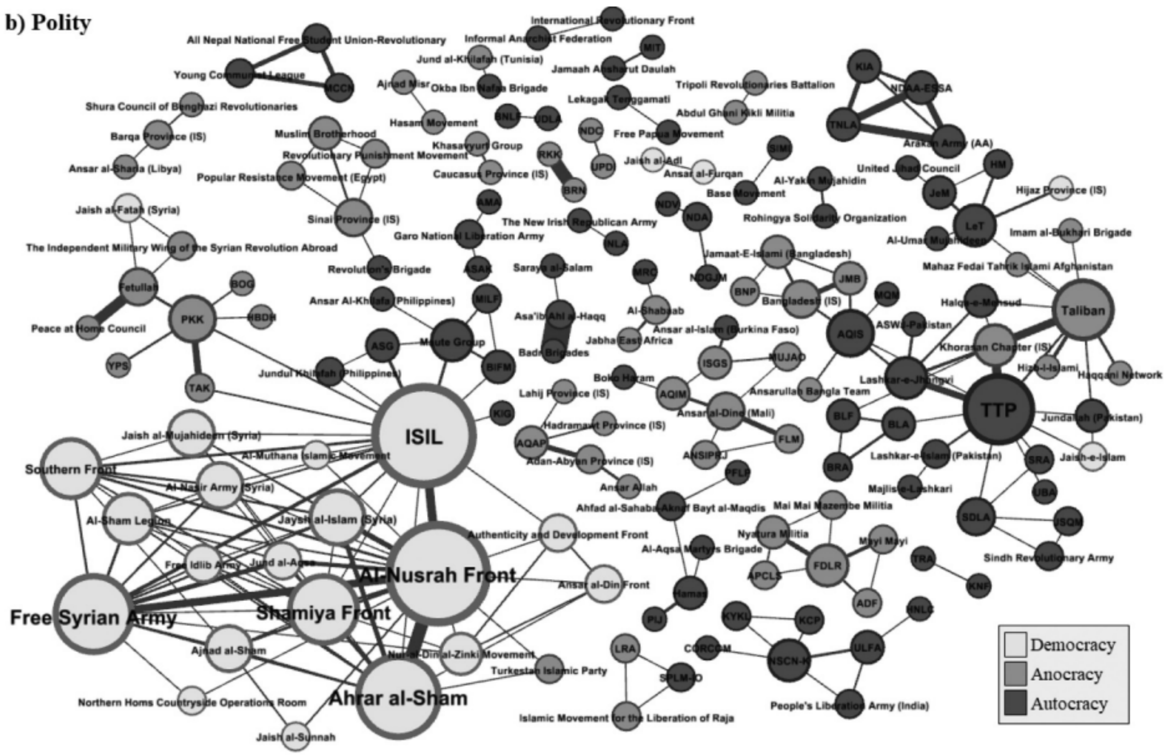
Application of Big Data in Social Science

week 10

Heidi Hyeseung Choi
Fall 2022
HYSIS
heidichoi@hanyang.ac.kr

Announcements

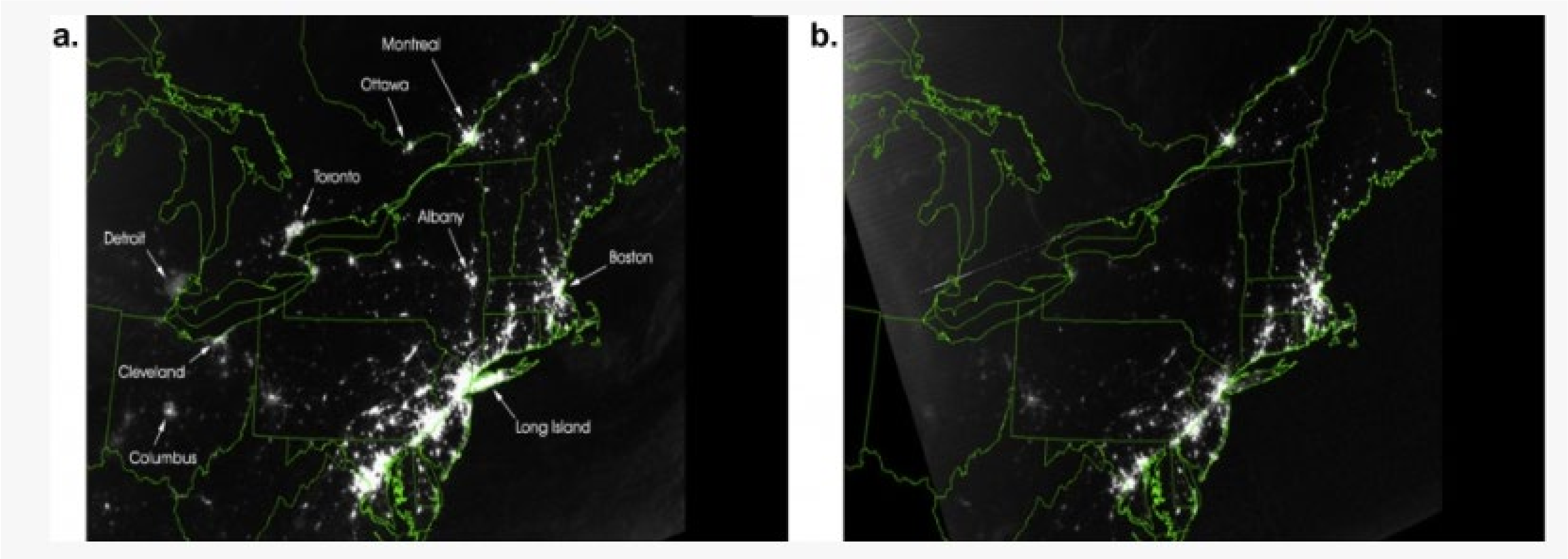
~~9.1~~	~~1~~	~~Introduction~~
~~9.8~~	~~2~~	~~Web Scraping~~
~~9.15~~	~~3~~	
~~9.22~~	~~4~~	~~Natural Language Processing~~
~~9.29~~	~~5~~	
~~10.6~~	~~6~~	~~Text Analysis~~ ~~(recorded lecture on week 7)~~
~~10.13~~	~~7~~	
~~10.20~~	~~8~~	~~Mid term exam (as school schedule)~~
~~10.27~~	~~9~~	~~Midterm review & word cloud~~
11.3	**10**	**Social Network Analysis**
11.10	11	Machine Learning: Supervised Learning
11.17	12	
11.24	13	Machine Learning: Unsupervised Learning
12.1	14	
12.8	15	Data Visualization
12.15	16	Final Exam



Interconnectivity.

- Vulnerability due to interconnectivity: cascading failures

2003 North American Blackout



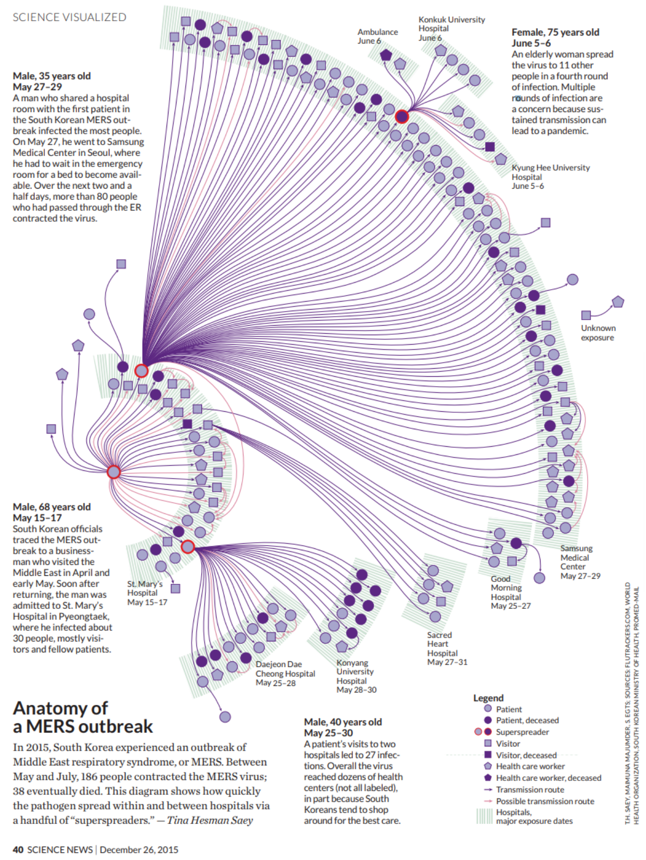
a real image of the US Northeast on August 14, 2003, before and after the blackout that left without power an estimated 45 million people in eight US states and another 10 million in Ontario.

- Cyber attacks
- Financial systems in 1997
- 2009-2011 financial meltdown: US credit crisis as the trigger.
- Money supply of terrorist organizations. Etc.

need to analyze the structure of the network!!

Network science?

- What is the overall structure of the network?
- Who are the important people, or hubs, in the network?
- What are the subgroups and communities in the network?



SCIENCE VISUALIZED

Male, 35 years old
May 27–29
A man who shared a hospital room with the first patient in the South Korean MERS outbreak infected the most people. On May 27, he went to Samsung Medical Center in Seoul, where he had to wait in the emergency room for a bed to become available. Over the next two and a half days, more than 80 people who had passed through the ER contracted the virus.

Female, 75 years old
June 5–6
An elderly woman spread the virus to 11 other people in a fourth round of infection. Multiple rounds of infection are a concern because sustained transmission can lead to a pandemic.

Male, 68 years old
May 15–17
South Korean officials traced the MERS outbreak to a businessman who visited the Middle East in April and early May. Soon after returning, the man was admitted to St. Mary's Hospital in Pyeongtaek, where he infected about 30 people, mostly visitors and fellow patients.

Anatomy of a MERS outbreak

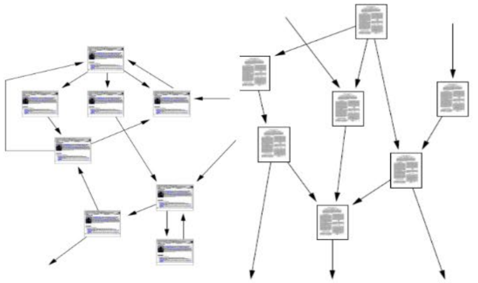
In 2015, South Korea experienced an outbreak of Middle East respiratory syndrome, or MERS. Between May and July, 186 people contracted the MERS virus; 38 eventually died. This diagram shows how quickly the pathogen spread within and between hospitals via a handful of "superspreaders." *— Tina Hesman Saey*

Male, 40 years old
May 25–30
A patient's visits to two hospitals led to 27 infections. Overall the virus reached dozens of health centers (not all labeled), in part because South Koreans tend to shop around for the best care.

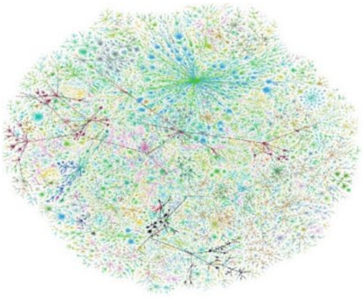
Legend
- ● Patient
- ● Patient, deceased
- ● ● Superspreader
- □ Visitor
- ■ Visitor, deceased
- ⬠ Health care worker
- ⬟ Health care worker, deceased
- → Transmission route
- → Possible transmission route
- ▥ Hospitals, major exposure dates

T.H. SAEY, MAIMUNA MAJUMDER, S. EGTS; SOURCES: FLUTRACKERS.COM, WORLD HEALTH ORGANIZATION, SOUTH KOREAN MINISTRY OF HEALTH, PROMED-MAIL

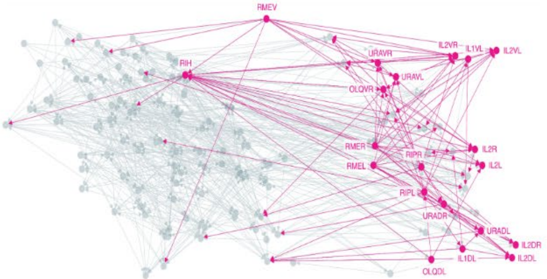
40 SCIENCE NEWS | December 26, 2015



Information networks:
Web & citations

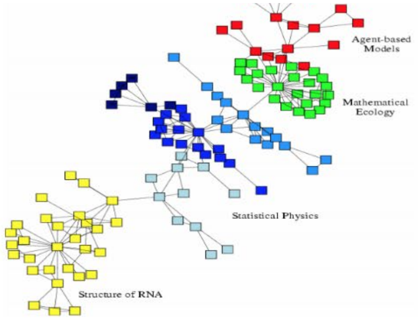


Internet

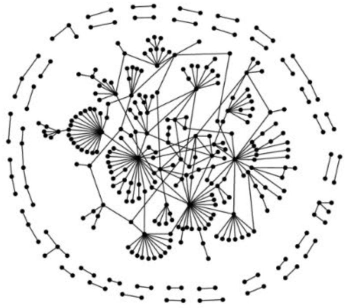


Networks of neurons

Social networks



Economic networks



Biomedical networks

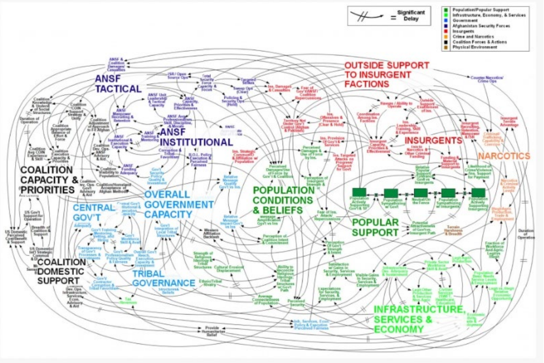


Image 1.6

The Network Behind a Military Engagement

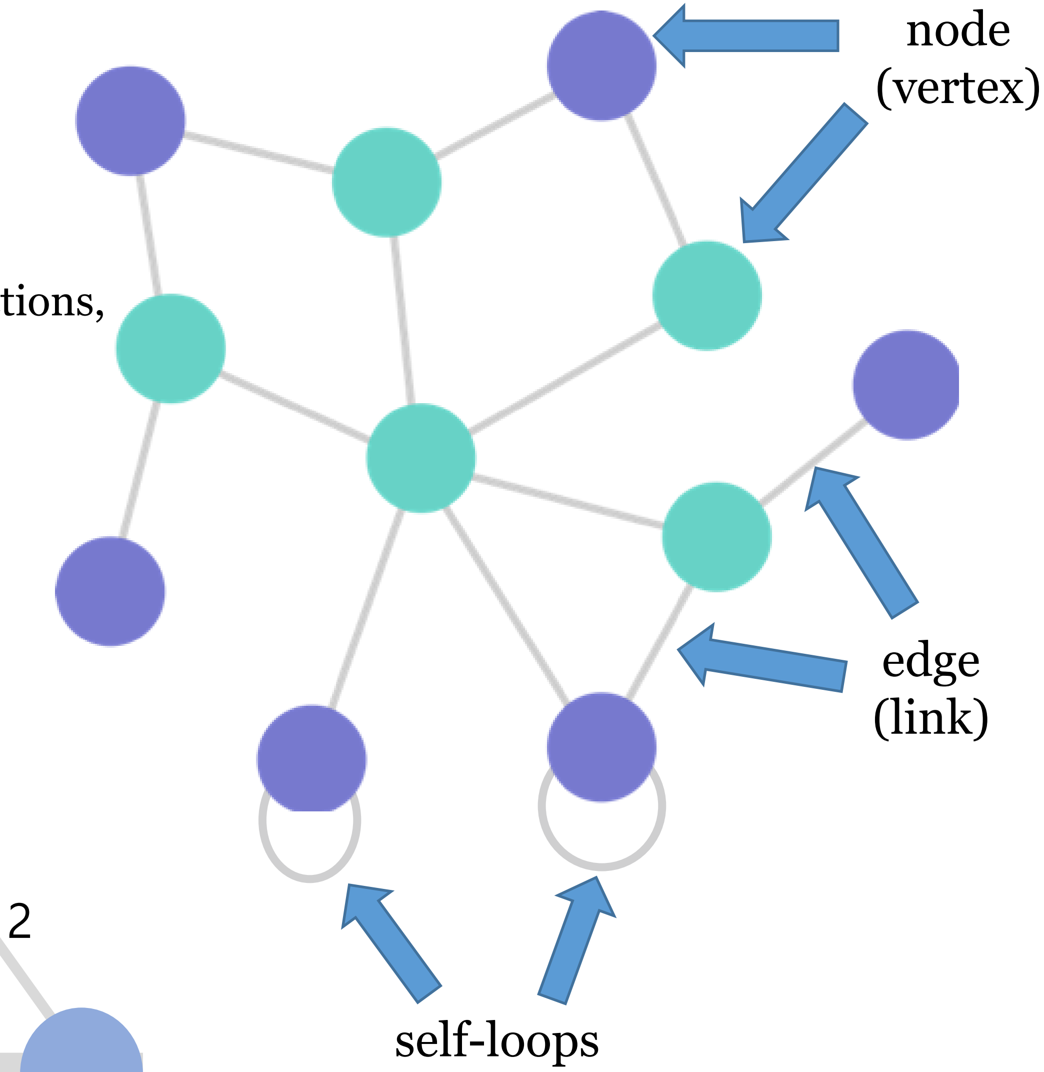
What is network science?

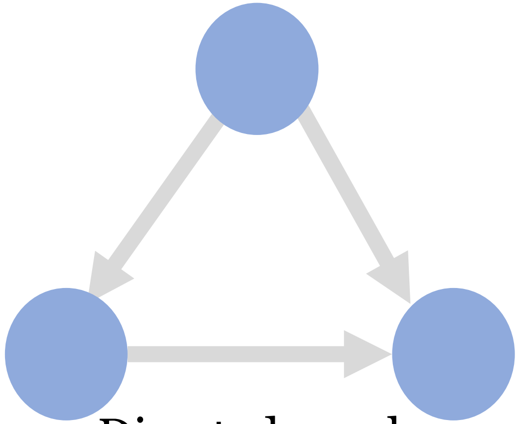
- an academic field which studies complex networks focusing on the study of presence of patterns of connection(structure) in a wide range of physical and social phenomena.
  - Eg) telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks,
- Draws on theories and methods from diverse discipline:
  - Graph theory: mathematics that study graphs
  - Statistical mechanics: physics
  - Data mining and information visualization: computer science
  - Inferential modeling: statistics
  - Social structure: sociology

Social Network Analysis(SNA) focuses on analyzing the patterns of relationships of people. (social structures, human behaviors, etc) from political science + sociological perspective.

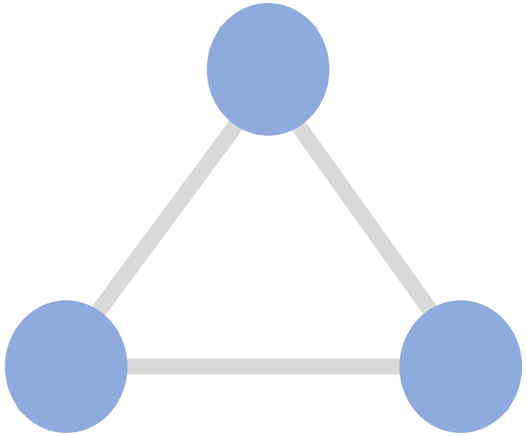
Basic definitions

- Node: actors
  - (individuals, organization, molecules, web documents, etc)
- Edge: relation that connects two nodes.
  - (family, professional, friendship (relations), alliances, chemical reactions, URLs, etc)
- Edges can be directed or undirected.
  - Directed graph: one-way relationship
  - Undirected graph: two-way relationship
- Weighted graph: graph with weights
- Self-loop: when edge connects to itself.

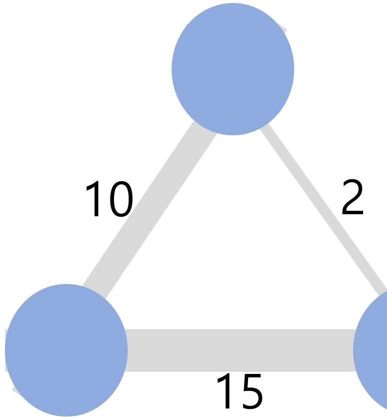




Directed graph



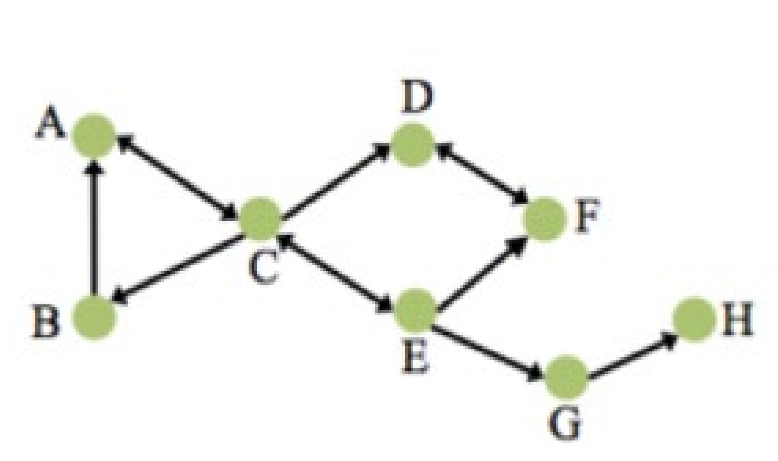
Undirected graph



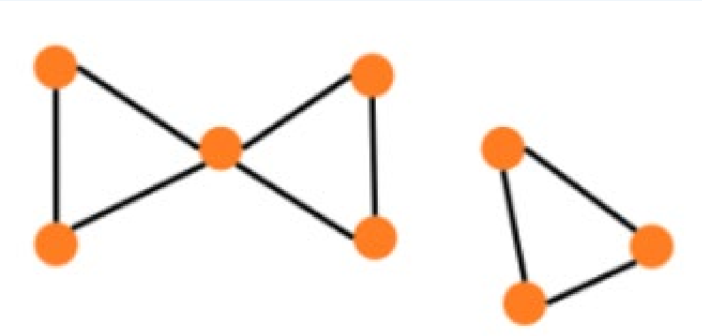
Undirected weighted graph

Basic concepts

- Network attributes: properties of the network elements.
  - Node attribute (eg. political affiliation, gender, nationality etc)
  - Edge attribute (eg. weight)
- Degree: the number of edges that it has to other nodes in the network.
  - directed graph: in-degree & out-degree
  - undirected graph: degree
- Path: a sequence of nodes in which each node is connected by an edge to the next.
- Path length: the distance between two nodes, measured as the number of edges between them.
- Components (islands): nodes that are disconnected from each other.
- Giant component: largest component within a network



A to F:
Path: ACDF, ACEF
Path length: 3, 3



Network Centrality measures

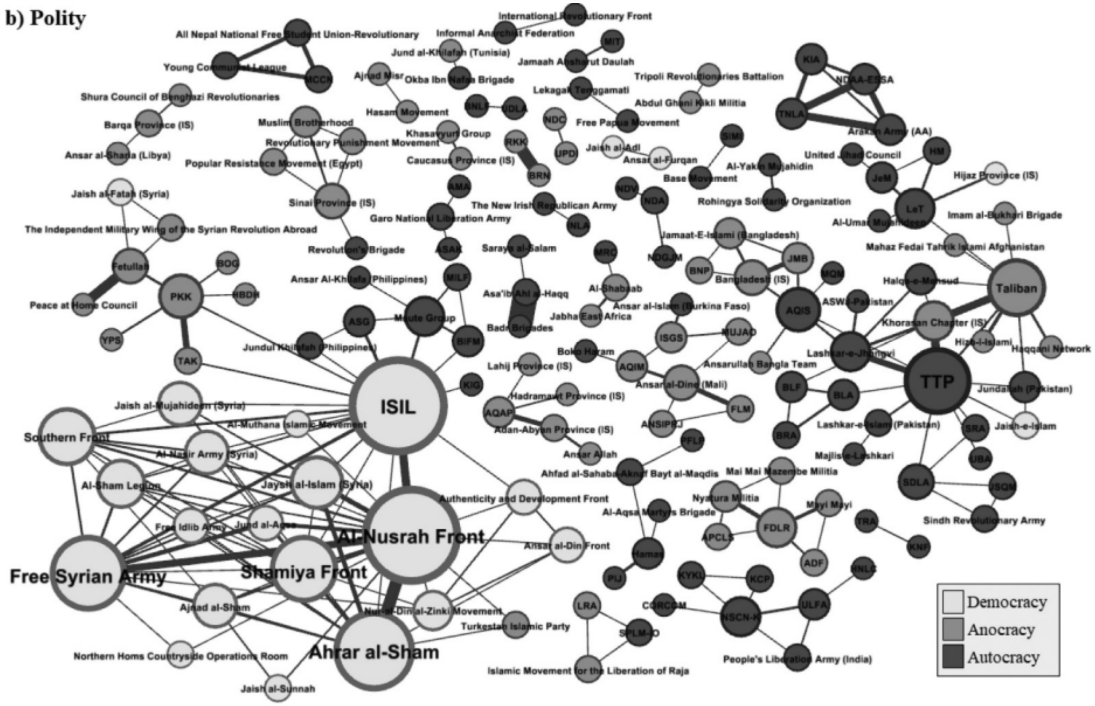
Question:

Within the network, which nodes are more important or influential?

Centrality measure is used to detect:

- How influential a person is within a social network?
- How efficiently a road is in a transportation network?

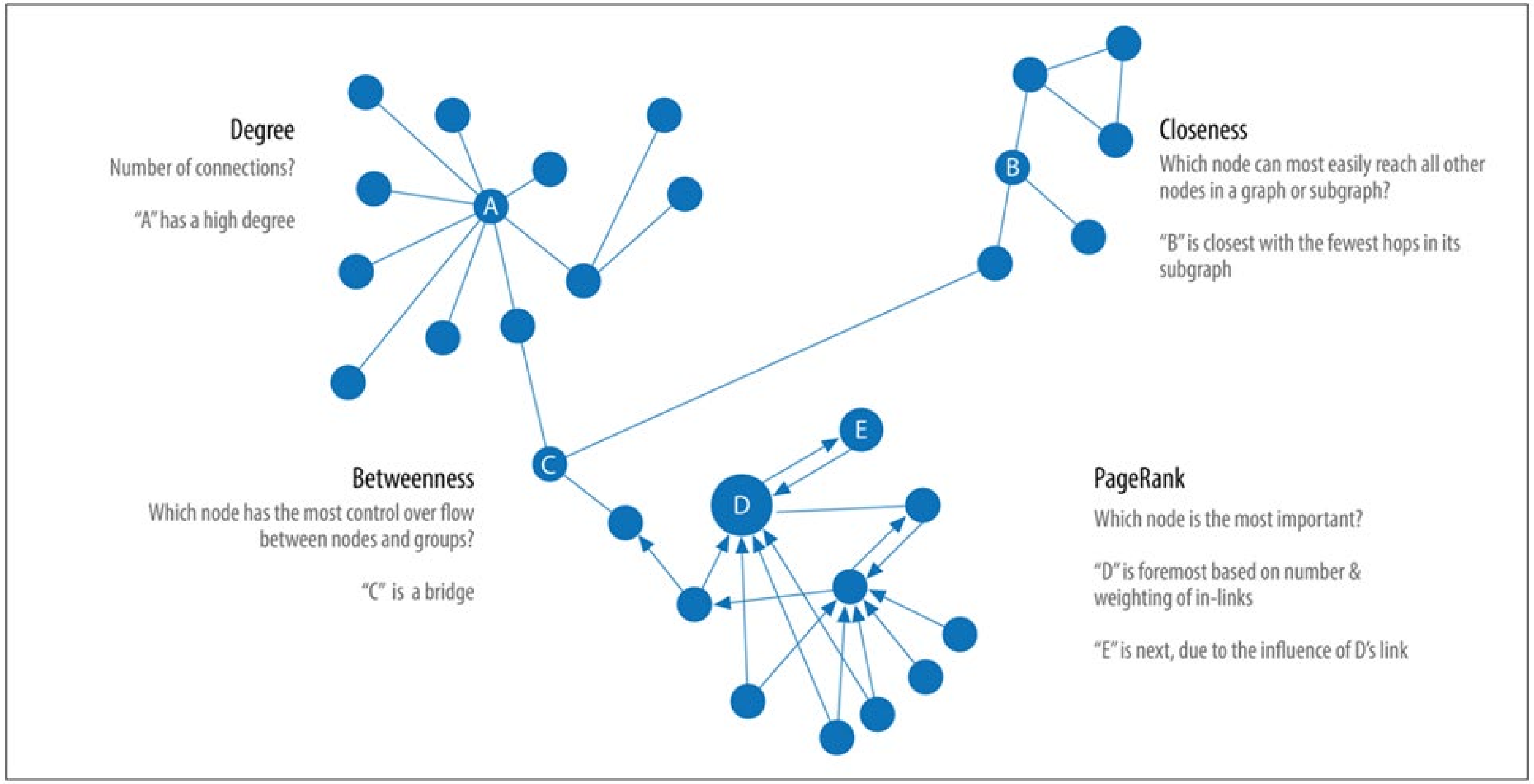
1. Degree centrality
2. Closeness centrality
3. Betweenness centrality
4. Eigenvector centrality



b) Polity

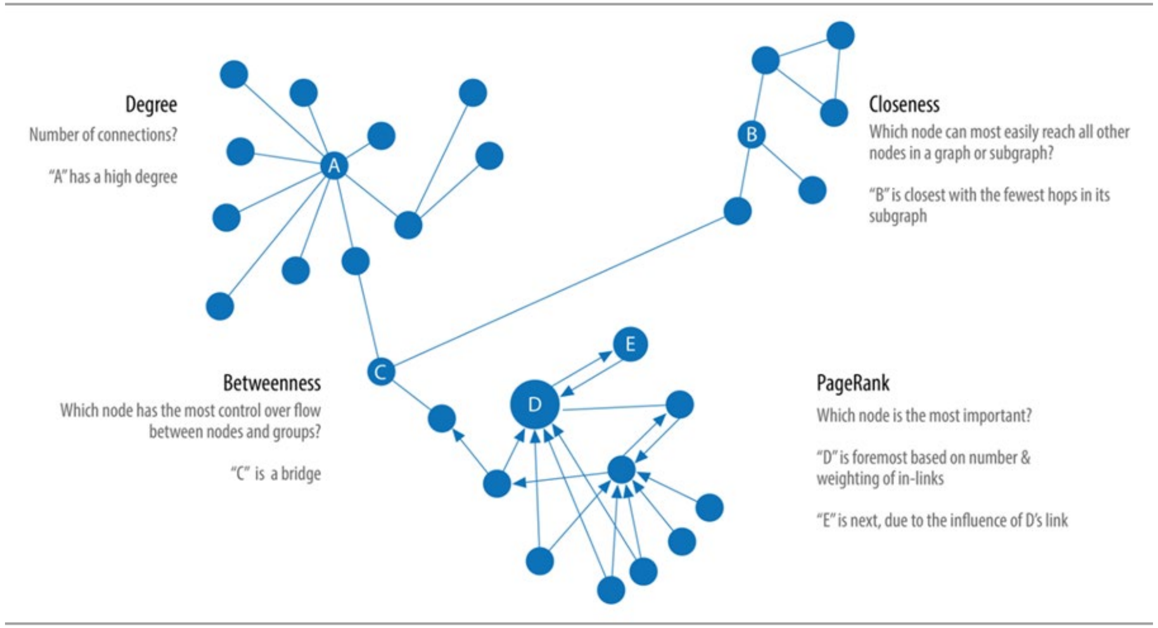
Degree centrality: most connected

- The nodes with higher degree is considered important.
- Degree centrality measure is useful for finding very connected individual, popular ones that are likely to hold a lot of information


Degree
Number of connections?
"A" has a high degree
A
B
Closeness
Which node can most easily reach all other nodes in a graph or subgraph?
"B" is closest with the fewest hops in its subgraph
C
Betweenness
Which node has the most control over flow between nodes and groups?
"C" is a bridge
D
E
PageRank
Which node is the most important?
"D" is foremost based on number & weighting of in-links
"E" is next, due to the influence of D's link

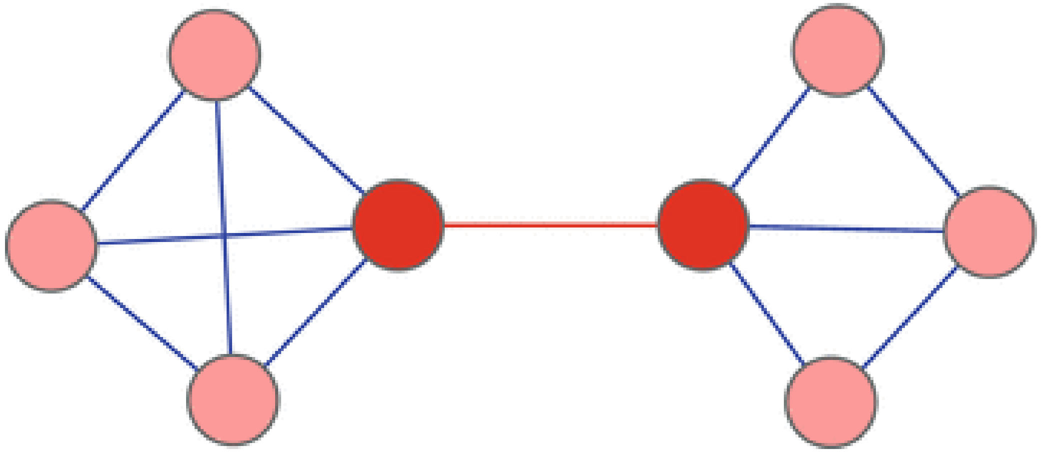
Closeness centrality: the closest to the rest of the nodes

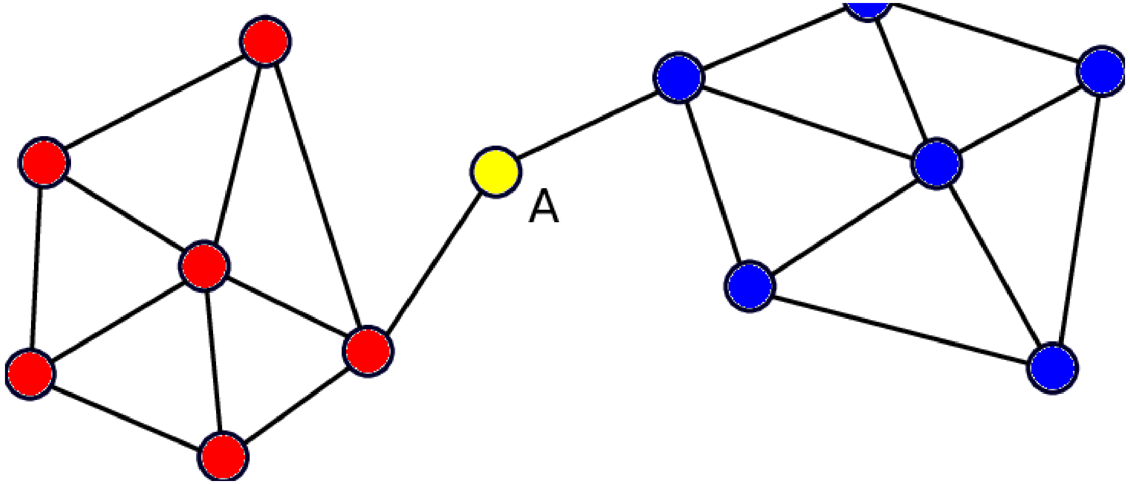
- scores each node based on their ‘closeness’ to all the other nodes in the network.
- Calculates the shortest paths between all nodes, and then assigns each node a score based on its sum of shortest paths.
- Nodes that are more ‘closer’ to other nodes have smaller distances
- This measure is useful when finding the one who are best placed to influence the entire network quickly.
- Helps find a good ‘broadcaster’.


Degree
Number of connections?
“A” has a high degree
A
Closeness
Which node can most easily reach all other nodes in a graph or subgraph?
“B” is closest with the fewest hops in its subgraph
B
Betweenness
Which node has the most control over flow between nodes and groups?
“C” is a bridge
C
D
E
PageRank
Which node is the most important?
“D” is foremost based on number & weighting of in-links
“E” is next, due to the influence of D’s link

Betweenness centrality: through which passes more info

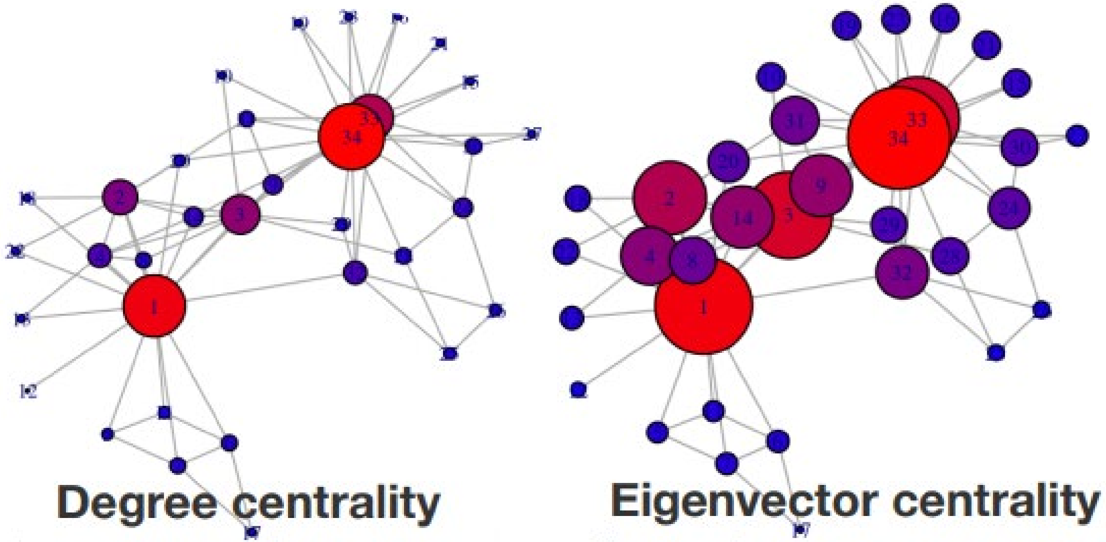
- Measures the number of times a node lies on the shortest path between other nodes.
- This measure shows which nodes work as ‘bridges’ between nodes in a network.
- normally used to find the individual with the most influence in the flow of the network.
- eg) analyzing communication dynamics





Eigenvector Centrality: connected to other important nodes

- measures a node's 'influence' based on the number of links it has to other nodes in the network.
- It helps identify nodes with influence over the whole network, not just the ones that are directly connected to it.
- It is considered as a good 'all-round' SNA score, handy for understanding human social networks, and networks like malware propagation.

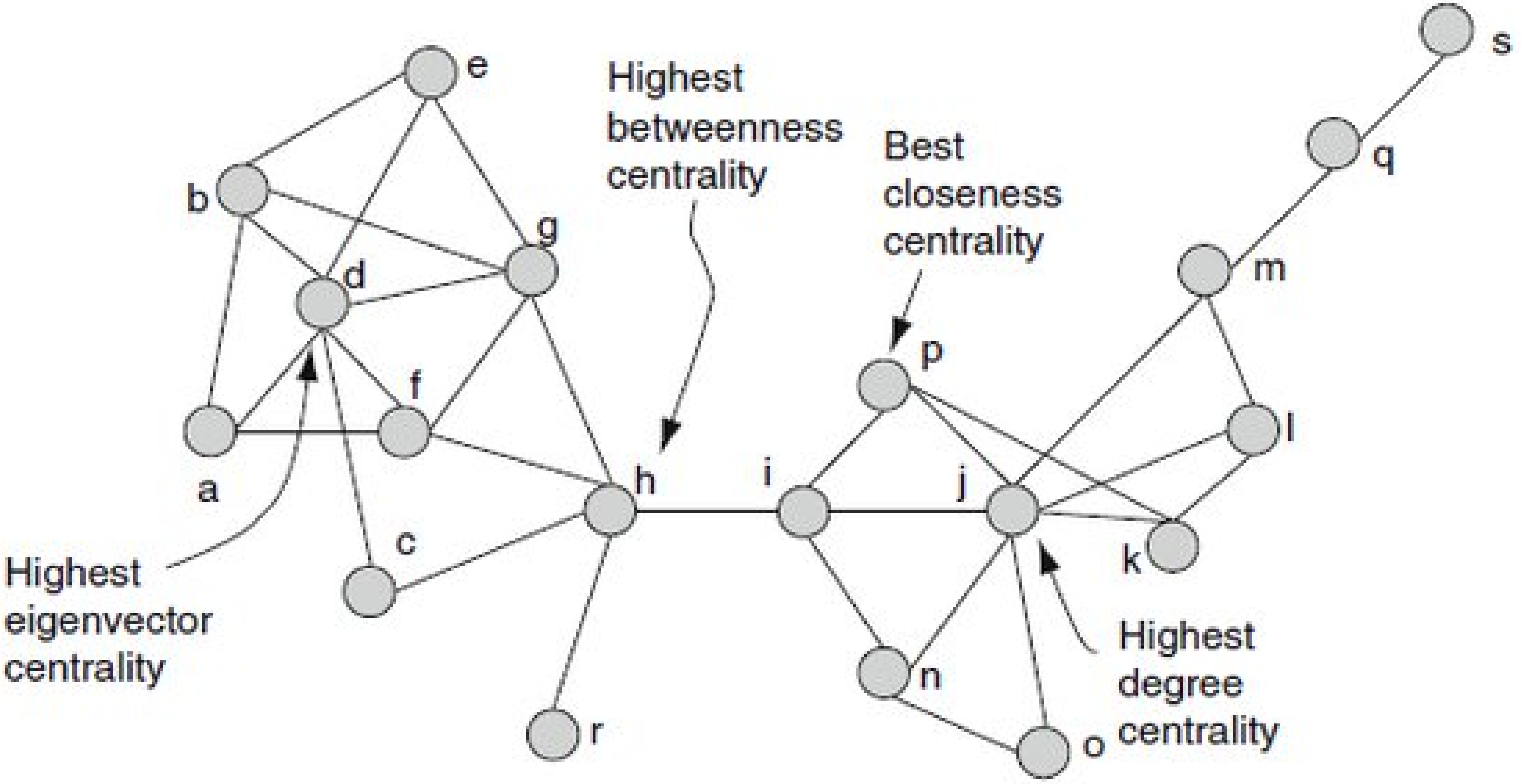


Degree centrality

Eigenvector centrality

Jure Leskovec, Stanford CS224W: Social and Information Network Analysis, http://cs224w.stanford.edu

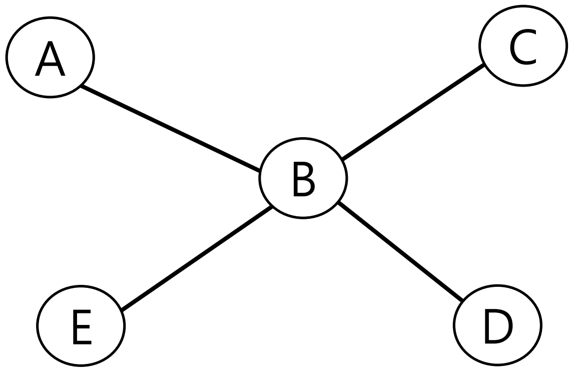
Putting four centrality measures together.

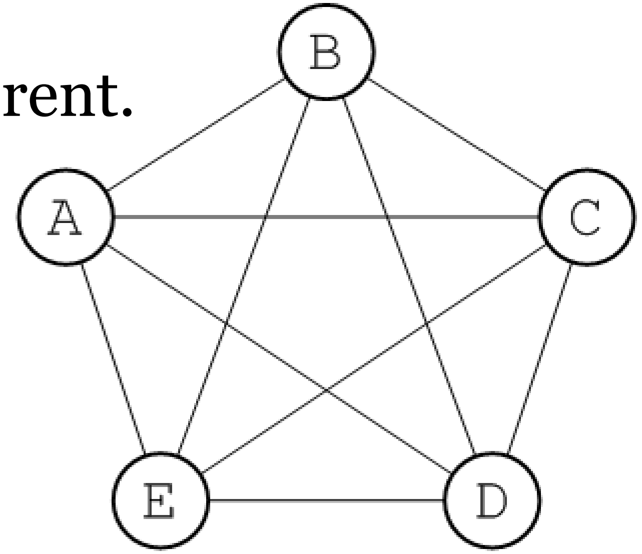


s
e
Highest betweenness centrality
Best closeness centrality
q
b
g
m
d
p
f
l
a
h
i
j
k
c
Highest eigenvector centrality
n
Highest degree centrality
r
o

Network density

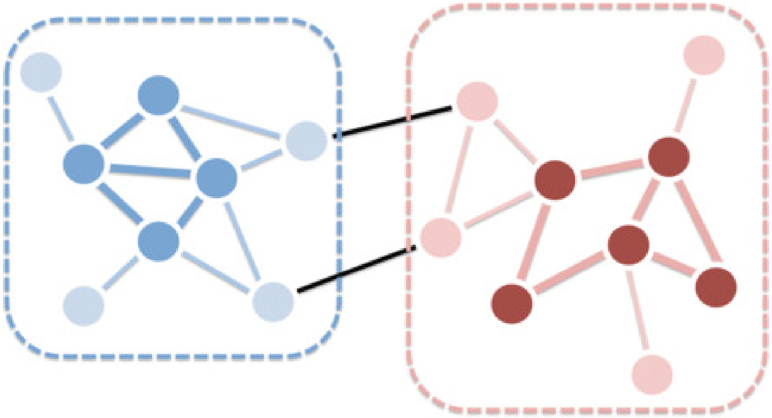
- Calculates how dense your network is.
- It is calculated by the ratio of actual edges in the network to all possible edges in the network.
- On a scale of 0 to 1, a complete network (where all nodes are connected to all other nodes) equals to 1.
- Density can be a good information because it helps us understand how connected the network is.
- Also it can tell us how the network structures are different.

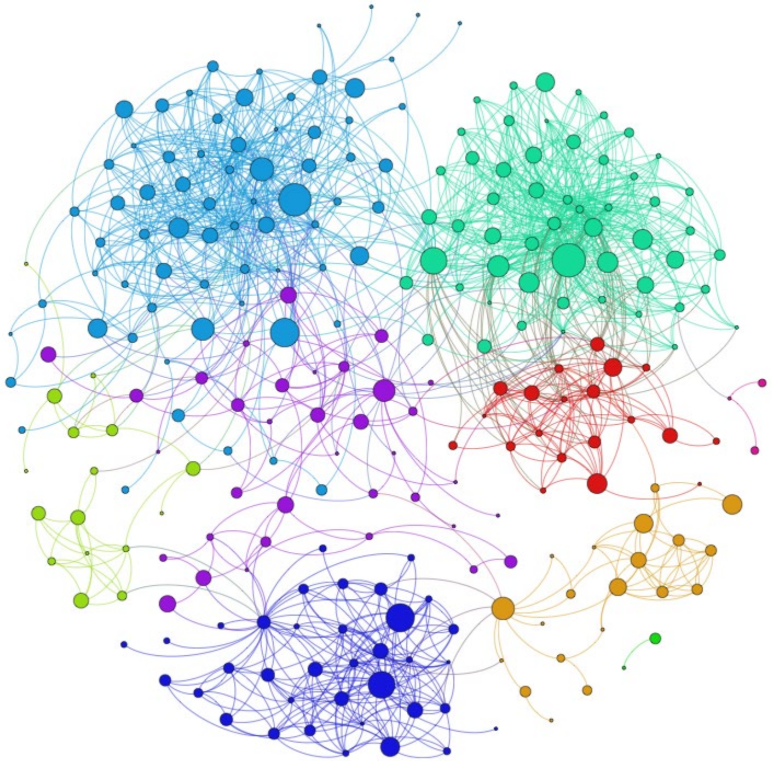




Modularity: community detection

- Measures the strength of division of a network into groups (clusters or communities).
- Network with high modularity score have dense connections between the nodes within groups, but sparse connections between nodes in different modules.
- This is useful when we look at social community, where faster transmission of information or rumor among them than a loosely connected community.

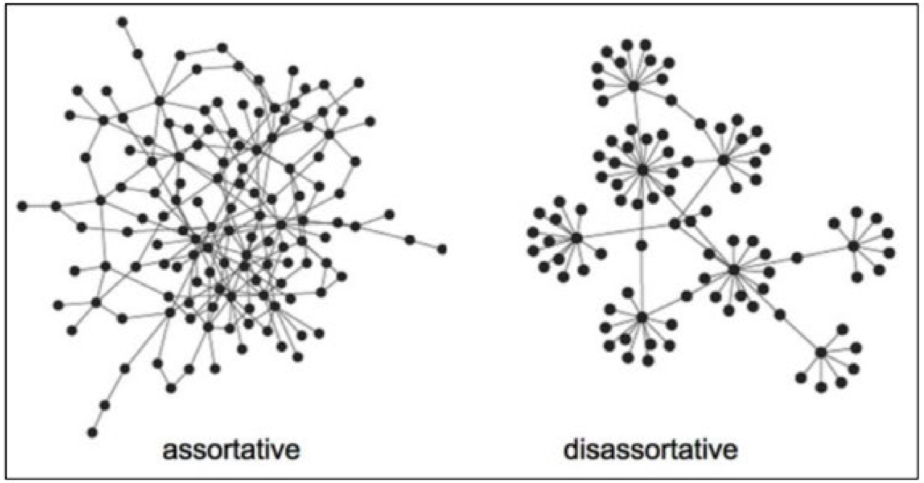


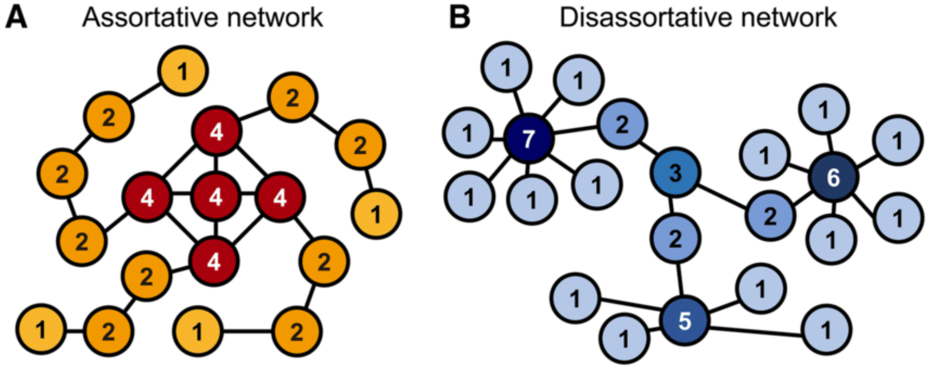


Network Assortativity-topology

- **assortative network:**
  - similar nodes may be more likely to attach to each other than dissimilar ones.
  - hubs are connected to hubs.
  - social networks tend to be assortative.

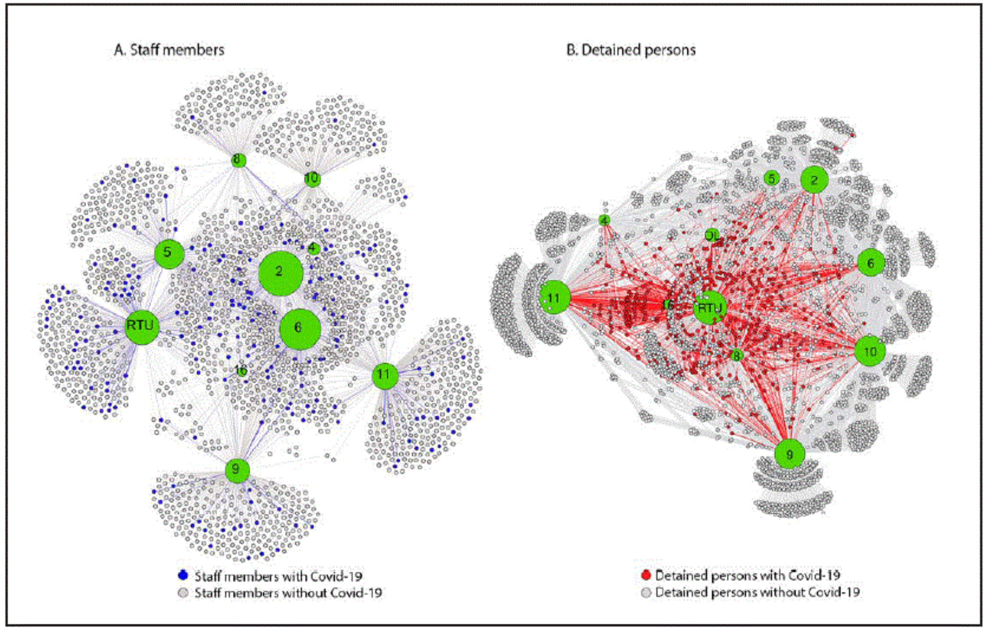
- **disassortative network:**
  - hubs (highly connected nodes) tend to avoid each other, and tend to link to lower-degree nodes.
  - display hub and spoke character.
  - eg) biological (metabolic, protein interaction, predator-prey) and technological networks (www, internet).



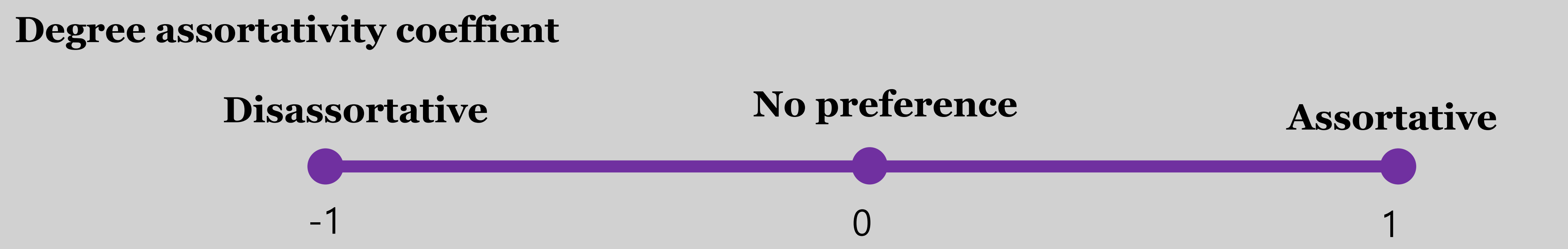


Network Assortativity-topology

- **Degree assortativity coeffient: -1~0~1**
  - positive (assortative mixing) value:
    - high degree nodes tend to connect to high degree nodes. low degree nodes tend to connect to low degree nodes.
  - negative (disassortative mixing) value:
    - high degree nodes tend to prefer low degree nodes. and vice versa.
  - close to 0:
    - it does not have any preference. this network is not assortative (different from disassortative)


A. Staff members
B. Detained persons
● Staff members with Covid-19
○ Staff members without Covid-19
● Detained persons with Covid-19
○ Detained persons without Covid-19
RTU
2
5
6
8
9
10
11

Network Characteristics and Visualization of COVID-19 Outbreak in a Large Detention Facility in the United States — Cook County, Illinois, 2020



Degree assortativity coeffient

Disassortative **No preference** **Assortative**

-1 0 1

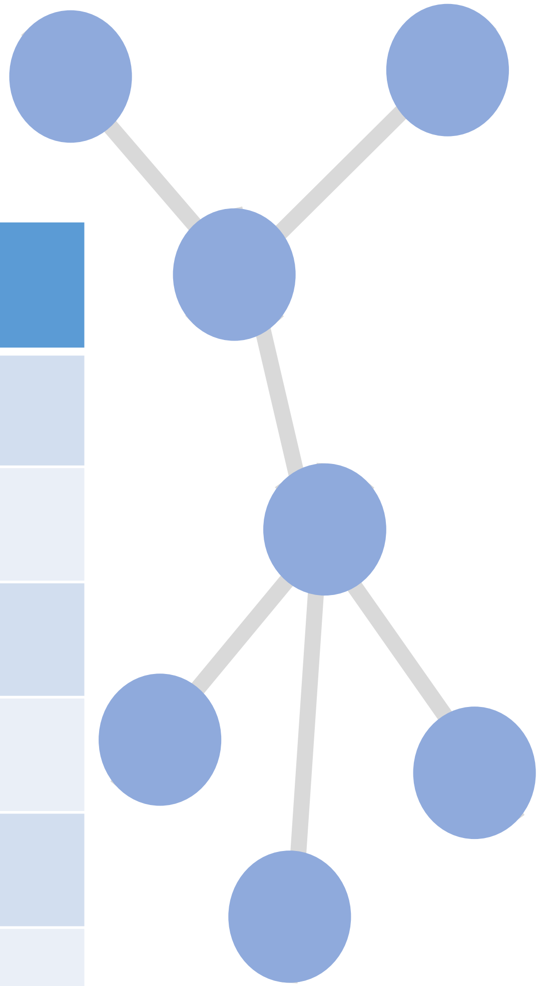
Data composition for network analysis

Node list with attributes

Node	Attribute 1 (gender)	Attribute 2 (age)	…
A	M	12	
B	F	16	
C	M	22	
D	M	27	
E	F	32	
F	F	26	
G	F	42	

Edge list

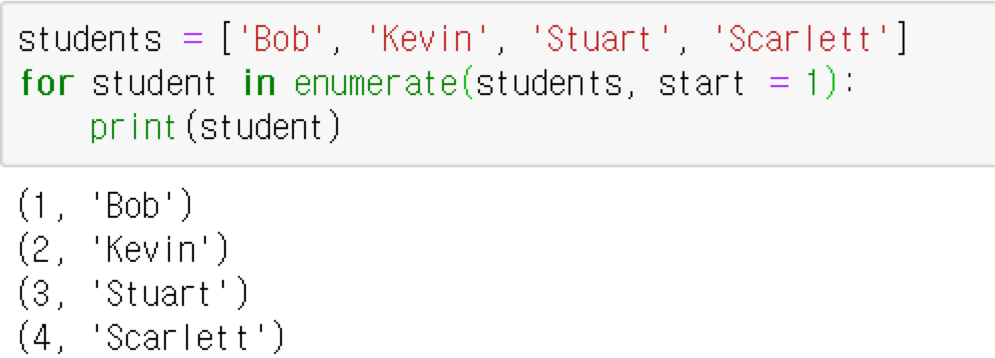
Source (origin) Side A	Target (destiny) Side B	Weight
A	B	2
B	C	3
C	D	1
D	F	2
E	A	1
F	B	1
G	E	1



enumerate() function

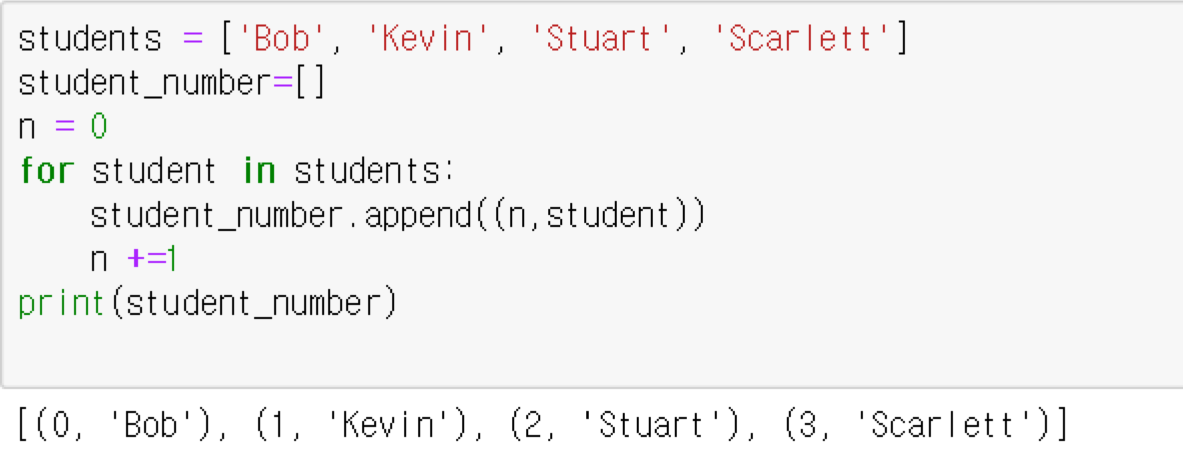
- enumerate function assigns an index to each item in a iterable object that can be used to reference the item later.

>>>enumerate(iterable, start=0)



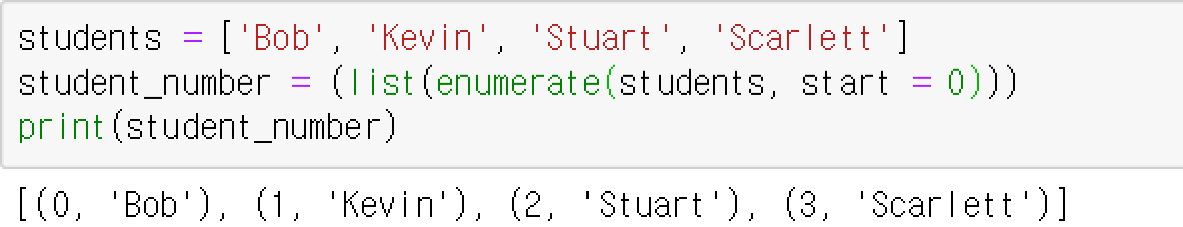
```
students = ['Bob', 'Kevin', 'Stuart', 'Scarlett']
for student in enumerate(students, start = 1):
    print(student)
```

```
(1, 'Bob')
(2, 'Kevin')
(3, 'Stuart')
(4, 'Scarlett')
```



```
students = ['Bob', 'Kevin', 'Stuart', 'Scarlett']
student_number=[]
n = 0
for student in students:
    student_number.append((n,student))
    n +=1
print(student_number)
```

```
[(0, 'Bob'), (1, 'Kevin'), (2, 'Stuart'), (3, 'Scarlett')]
```



```
students = ['Bob', 'Kevin', 'Stuart', 'Scarlett']
student_number = (list(enumerate(students, start = 0)))
print(student_number)
```

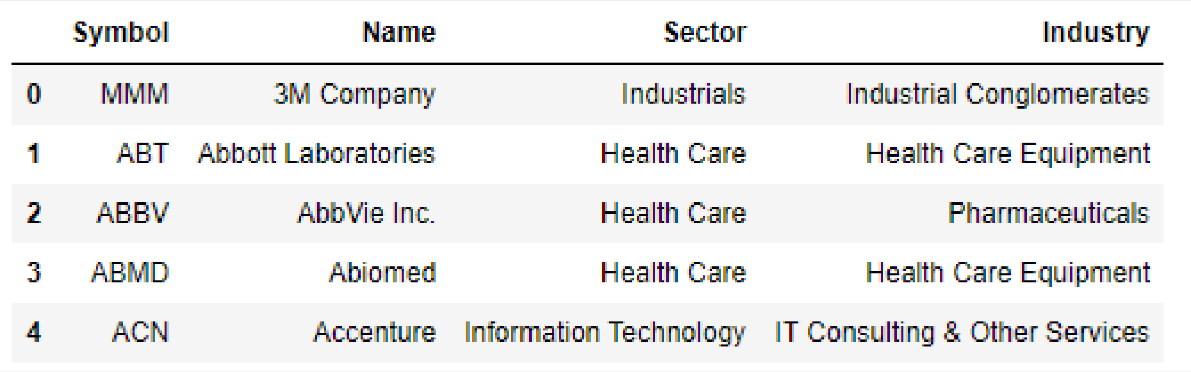
```
[(0, 'Bob'), (1, 'Kevin'), (2, 'Stuart'), (3, 'Scarlett')]
```

set_index().to_dict() of pandas

We can use this if you want to make DataFrames into dictionaries!

- It is composed of two parts.

1. set_index()
   - Is used when we want to set the DF index using existing columns.

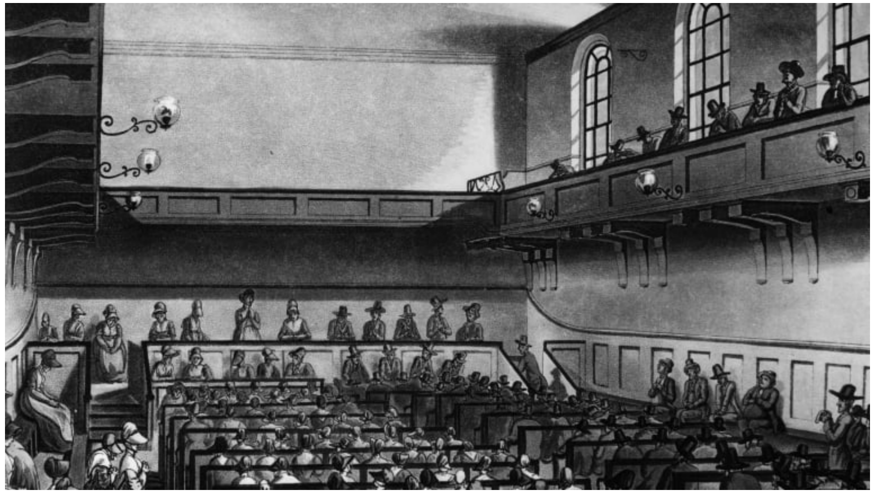


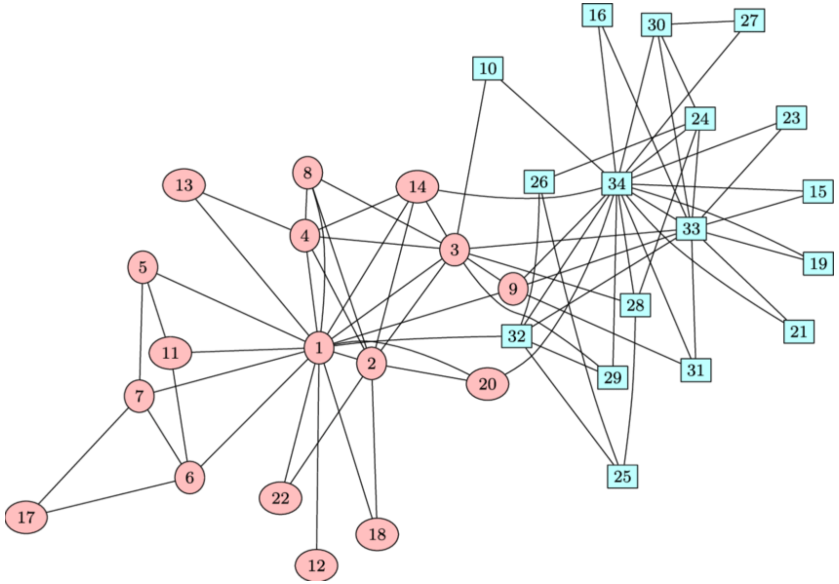
	Symbol	Name	Sector	Industry
0	MMM	3M Company	Industrials	Industrial Conglomerates
1	ABT	Abbott Laboratories	Health Care	Health Care Equipment
2	ABBV	AbbVie Inc.	Health Care	Pharmaceuticals
3	ABMD	Abiomed	Health Care	Health Care Equipment
4	ACN	Accenture	Information Technology	IT Consulting & Other Services

2. To_dict()
   - Converts dataframe to a dictionary.
   - Parameters: dict(default), index, etc
     - dict = **{column : {index: value}}**
       - **{‘Symbol’: [0: ‘MMM’, 1: ‘ABT’ …}**
     - Index = {index : {column1 : value1, column2 : value2}}
       - {0: {‘Symbol’: ‘MMM’, ‘Name’: ‘3M Company’, ‘Sector’: ‘Industrials’….},
         1: {‘Symbol’: ‘ABT’, ‘Name’: ‘Abbott Laboratories’, ‘Sector’: ‘Health Care’….}}

Datasets we will be using today

- Zachary’s Karate Club
  - Dataset of friendship between the 34 members of a karate club at a US university in 1977.
  - Node: member
  - Edge: a tie between two members.
  - The Karate Club split into two after an argument between two teachers.
- Quakers dataset
  - historically Protestant Christian set of denominations, a.k.a. “Religious Society of Friends”.
  - Network of early Quakers.





Packages to install

```
>>> pip install pyvis
>>> pip install networkx
```