Social Media Mining

MK Data Mining 2

M. N. Fakhruzzaman, S.Kom., M.Sc.

Program Studi S1 Teknologi Sains Data

Fakultas Teknologi Maju dan Multidisiplin

Universitas Airlangga Indonesia

Outline

- Text mining in Social media
- Social Network analysis

Text Mining In Social Media

- People use social media to communicate
- Social media provides rich information of human interaction and collective behavior
- Traditional Media vs. Modern Social Media
- Information in most social media sites are stored in text format
- Text Mining can help deal with textual data in social media for research
- But, not only text can be mined from social media

Distinct Aspects of Text in Social Media

- Textual data provides insights into social networks
- Textual data also presents new challenges:
  - Time Sensitivity (outdated post may not be relevant anymore)
  - Short Length (typical characteristics in facebook statuses and tweets)
  - Unstructured Phrases (slangs, alay, typing wilithekid)

Aspect #1: Time Sensitivity

- Social media's real-time nature
  - Example: Twitter users tweet about specific hot issue in a specific timeframe, and trend changes overtime
- Large number of real-time updates from Facebook and Twitter contain abundant information
  - Information → detection and monitoring of an event
  - Use data to track a user's interest in an event
- A user is connected and influenced by his/her friends
  - Example: People will not be interested in a movie after several months, while they may be interested in another movie released several years ago because of the recommendation from his friends

Aspect #2: Short Length

- Certain social media websites have restrictions on the length of user's content
  - Twitter's 140 characters rule
  - Windows Live Messenger's 128 character personal status
- Short Messages → people become more efficient with their participation in social media applications
- Short Messages also bring new challenges to text mining

Aspect #3: Unstructured Phrases

- Variance in quality of content makes the tasks of filtering and ranking more complex
- Computer software have difficulties to accurately identify semantic meaning of new abbreviations or acronyms

Applying Text Mining in Social Media

- Certain aspects of textual data in social media presents great challenges to apply text mining techniques

Event Detection

- Event Detection aims to monitor a data source and detect the occurrence of an event that is captured within that source
  - Monitor Real-Time Events via Social Media
  - Example: Detecting earthquake when people are posting live-situation through microblogging like Twitter & Facebook
- Improve traditional news detection
  - Large number of news are generated from various new channels, but only few receive attention from users
  - Some news channels are identified to be malicious and full of hoaxes

Collaborative Question Answering

- Collaborative question answering services bring together a network of self-declared “experts” to answer questions posted by other people
- Through text mining, a tremendous amount of historical QA pairs have built up their databases, and this transformation gives users an alternative place to look for information, as opposed to a web search
- The corresponding best solutions could be explicitly extracted and returned

Social Tagging

- A method for Internet users to organize, store, manage and search for tags / bookmarks (also as known as social bookmarking) of resources online
- Social Tagging vs. File Sharing
- Through text mining, it helps to improve the quality of tag recommendation
  - Facebook’s tag recommendation of a photo
- Utilize social tagging resources to facilitate other applications
  - Web object classification, document recommendation, web search quality

What can you do with it?

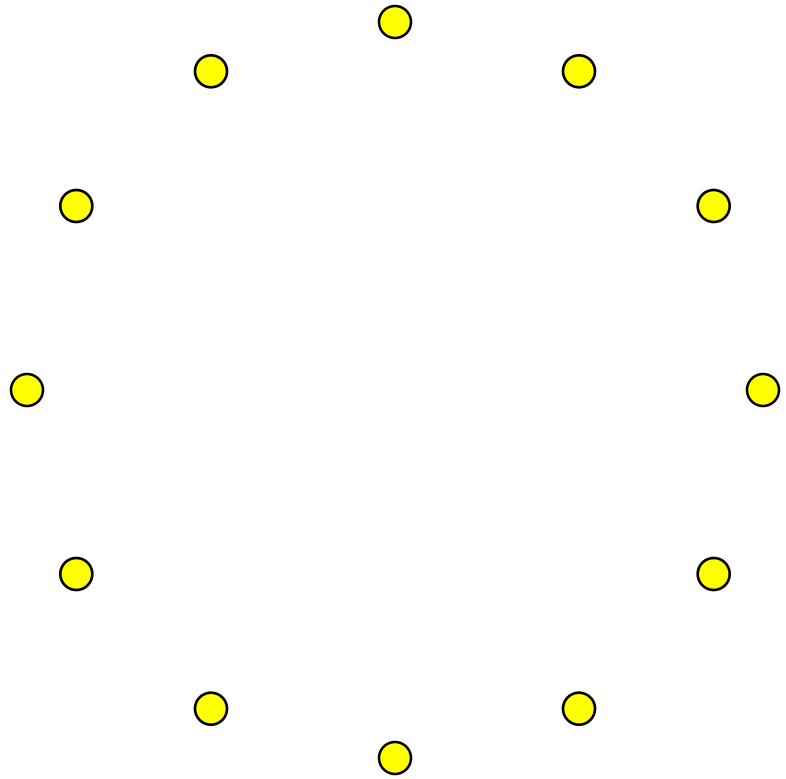
- Doing a lexicon based sentiment analysis (using VADER or TextBlob)
- TF-IDF Clustering to detect topics or sentiments (or using another feature extraction method)
- Generate social network to learn how human communicate via social media

Social Network Analysis

- Using graph theory to model user interaction on social media
- Social network analysis is a method by which one can analyze the *connections* across individuals or groups or institutions.

Society as a Graph

People are represented as *nodes*.

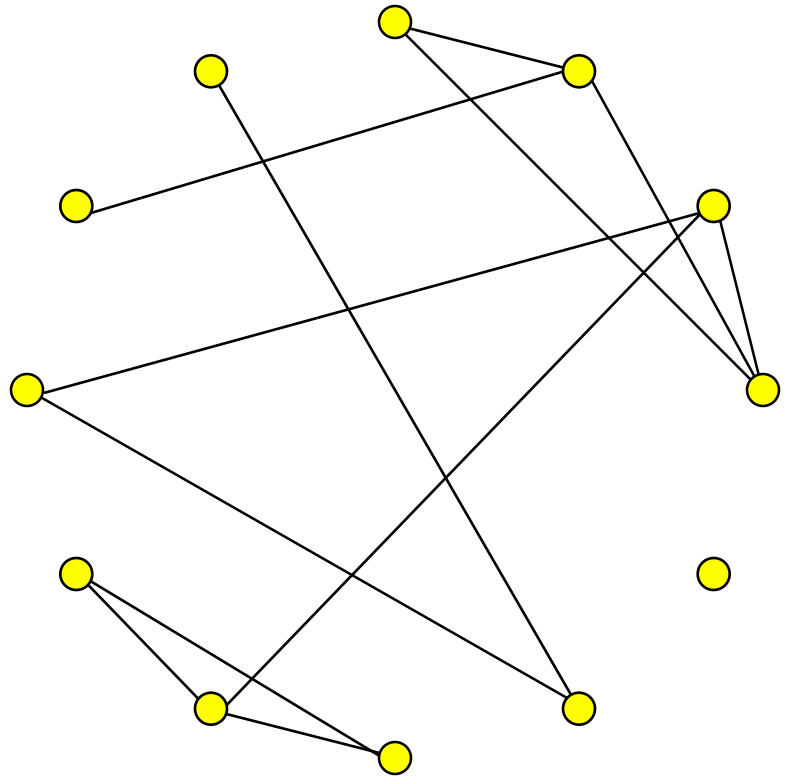


Society as a Graph

People are represented as *nodes.*

Relationships are represented as *edges.*

(Relationships may be acquaintanceship, friendship, co-authorship, mentions, follows etc.)



Allows analysis using tools of mathematical graph theory

Some concepts

- Before we discuss “the strength of weak ties” and “small worlds”, let’s just go over some basic concepts.

- A **node or vertex** is an individual unit in the graph or system. (If it is a network of legislators, then each node represents a legislator).

- A **graph or system or network** is a set of units that may be (but are not necessarily) connected to each other.

Some concepts

- An “edge” is a connection or tie between two nodes.
- A **neighborhood** N for a vertex or node is the set of its immediately connected nodes.
- **Degree**: The degree k_i of a vertex or node is the number of other nodes in its neighborhood.

Some concepts

- In an **undirected** graph or network, the edges are reciprocal—so if A is connected to B, B is by definition connected to A. (e.g. Friendship in facebook)

- In a **directed** graph or network, the edges are not necessarily reciprocal—A may be connected to B, but B may not be connected to A (think of a graph with arrows indicating direction of the edges.) (e.g. mentions, follows in twitter / insta)

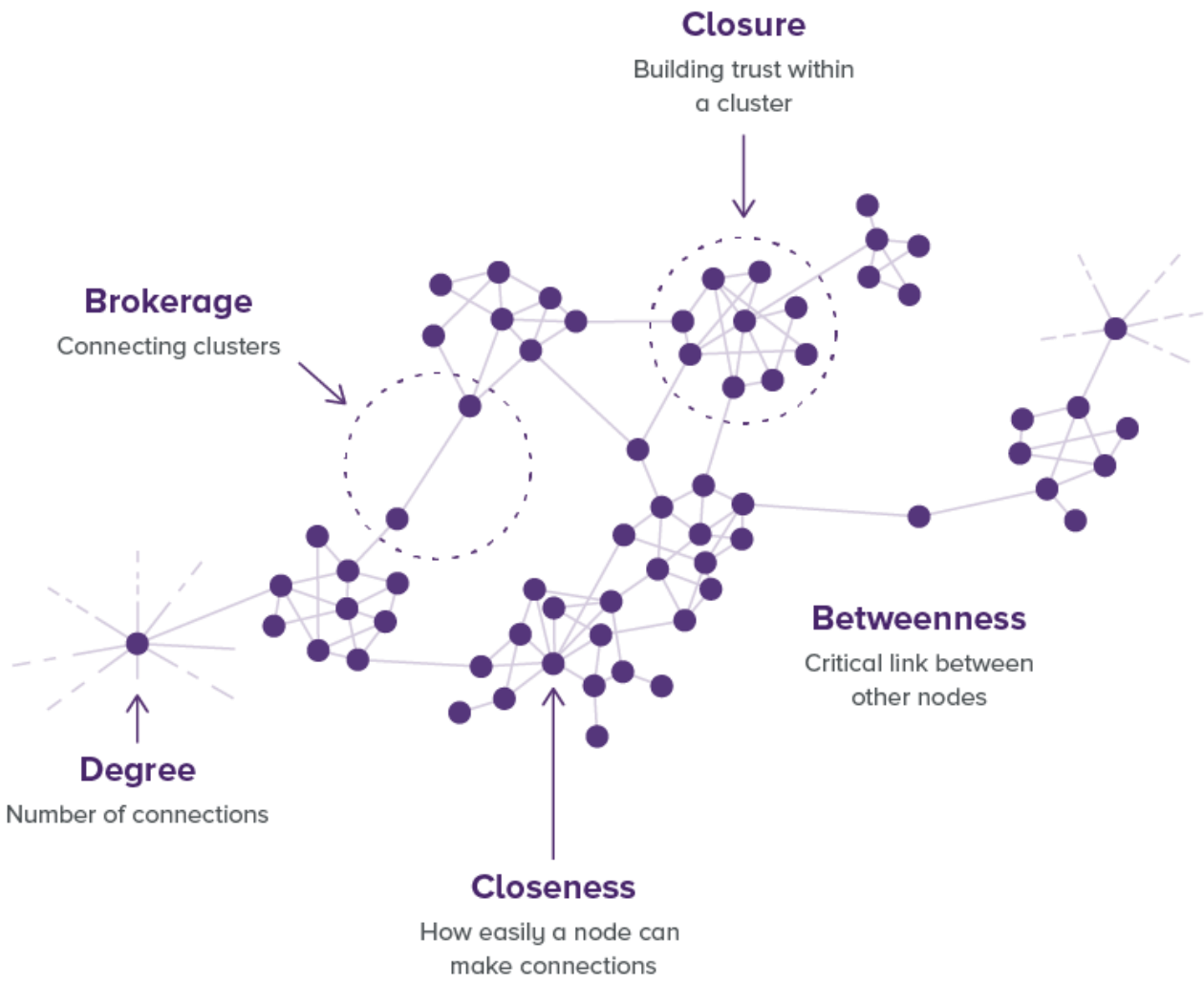
- Okay, now let's discuss the meaning of the "strength of weak ties"....

The Strength of Weak Ties

- Granovetter's "The Strength of Weak Ties" (considered one of the most important sociology papers written in recent decades) argued that "weak ties" could actually be more advantageous in politics or in seeking employment than "strong ties", because weak ties allowed an individual to reach a higher number of other individuals.

The Strength of Weak Ties

- Granovetter observed that the presence of weak ties often reduced path lengths (distance) between any two individuals—which led to quicker diffusion of information.
- The weak node can act as a connector or bridge between two other nodes (or maybe communities / clusters)



Small Worlds---Intro

Next, let's consider the related concept of "small worlds", another concept that has emerged in network analysis.

But for some background, let's discuss some different possible types of graphs, plus the concepts of "clustering" and "diameter".

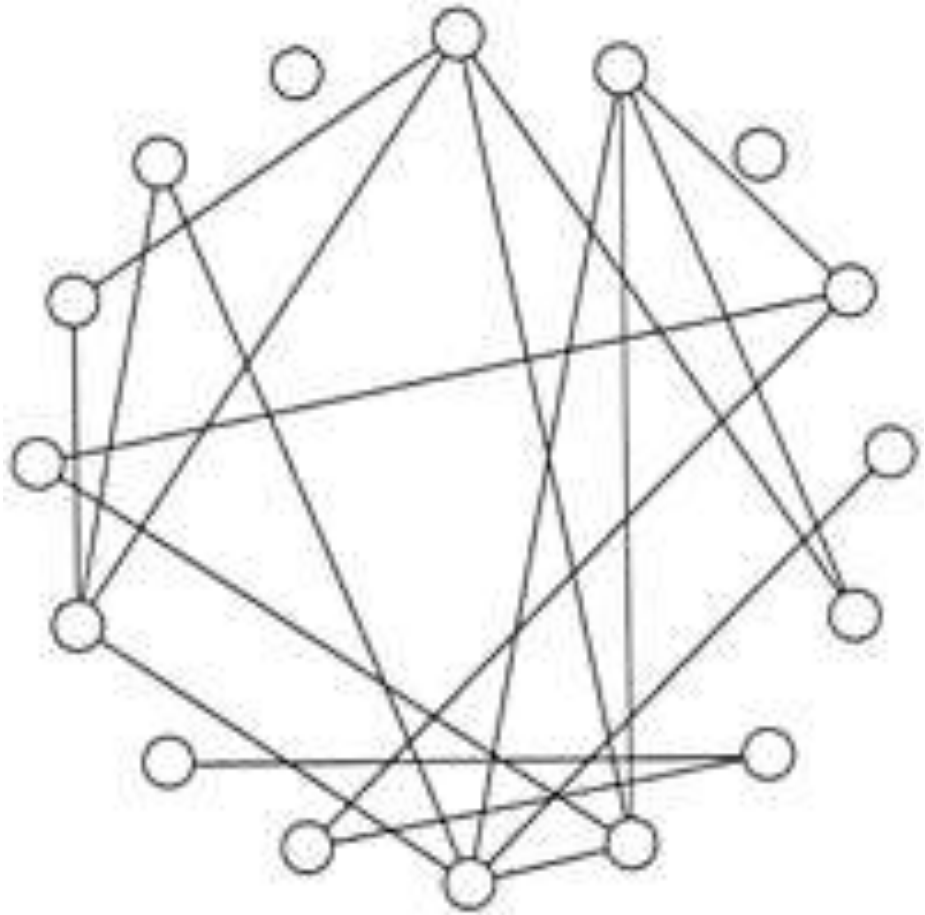
Two possible graphs (almost at opposite ends of a spectrum) are "random graphs" and "regular graphs". A "small world" can be thought of in-between a random and a regular graph.

Background→Random Graphs

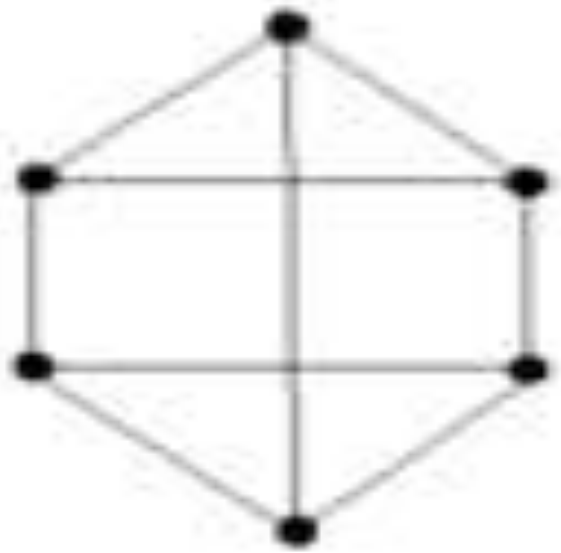
- In a **random graph,** each pair of vertices *i, j* has a connecting edge with an independent probability of p

- This graph has 16 nodes, 120 possible connections, and 19 actual connections—about a 1/7 probability than any two nodes will be connected to each other.

- In a random graph, the presence of a connection between A and B as well as a connection between B and C will not influence the probability of a connection between A and C.



Background→Regular Graphs



- A **regular graph** is a network where each node has the same number (k) of neighbors (that is, each node or vertex has degree k).
- A k-degree graph is seen at the left. k = 3 (each node is connected to three other nodes—that is, there are three nodes in each node's neighborhood.)

Clustering Coefficients

- This formula (on the right) is for the total number of possible connections for an undirected matrix. (Think in terms of a matrix—the total number of possible connections is half of the total # of cells, after subtracting the diagonal.)

$$\tau_G(v) = C(k_i, 2) = \frac{1}{2} k_i(k_i - 1).$$

A Very Simple Example

	A	B	C	D
A		1	0	1
B	1		1	0
C	0	1		0
D	1	0	0	

- Four legislators—whether they serve on at least one committee together.
- This is an undirected matrix—if legislator A serves with legislator B on a committee, then legislator B serves with legislator A on a committee.

A Very Simple Example

	A	B	C	D
A		1	0	1
B	1		1	0
C	0	1		0
D	1	0	0	

- The possible number of connections in this matrix is 6.
- K=4 legislators.
- ½ * k * (k-1) = ½ * 4 * 3
- = 6

A Very Simple Example

	A	B	C	D
A		1	0	1
B	1		1	0
C	0	1		0
D	1	0	0	

- The clustering coefficient for legislator A is 2/3 – s/he is "connected to" two out of a possible 3 other legislators. The same is true of legislator B.
- Legislators C and D each have a clustering coefficient of 1/3.

A Very Simple Example

	A	B	C	D
A		1	0	1
B	1		1	0
C	0	1		0
D	1	0	0	

- The average of those four clustering coefficients is .5.
- And note that across the entire network, .5 (3 of 6) of all possible connections are actually made.

Clustering Coefficients

- This is the formula the clustering coefficient for the system. N=number of nodes. C=clustering coefficient for each node *i.*

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i .$$

Clustering Coefficient

- Note that the clustering coefficient for undirected graphs is a bit different than the clustering coefficient for directed graphs—there are twice as many possible ties, a non-reciprocated edge counts for one tie, and a reciprocated edge counts for two ties.

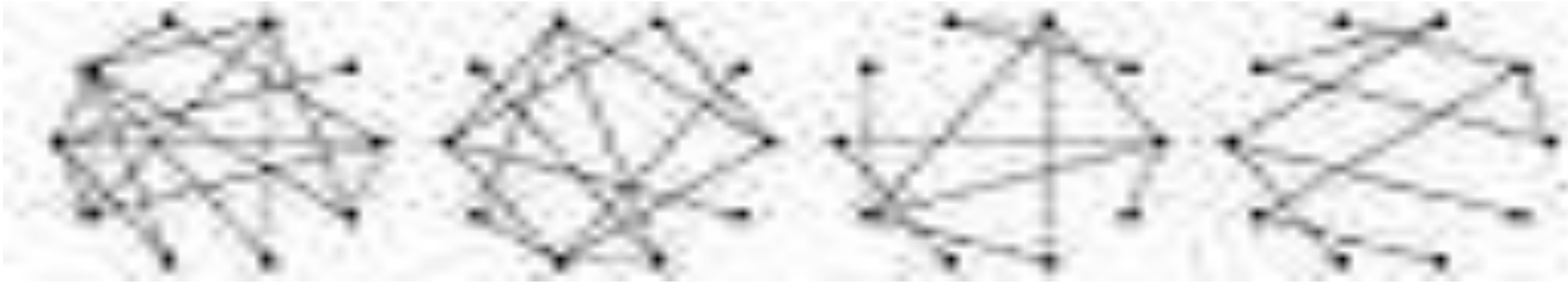
Clustering Coefficient

- So, in an undirected graph, if a node is connected to four other nodes—and among those four, only the first and the third are connected—the clustering coefficient is 1/6. (1 actual connection out of 6 possible connections.)

- Clustering refers to how connected your neighbors are to each other (relative to how connected they could be)

- Now let's talk about **network diameter**.

Graph Diameter



- The graph diameter is the “longest shortest path” between any two vertices or nodes.
- The graphs above have diameters of 3, 4, 5, and 7, respectively.
- The graph on the right has a relatively large diameter, because it takes (at most) 7 edges to travel between one node to another. (the two nodes at the very bottom of the network are not very closely connected)

Network Measures

- Size
  - Number of nodes
- Density
  - Number of ties that are present the amount of ties that could be present
- Out-degree
  - Sum of connections from an actor to others
- In-degree
  - Sum of connections to an actor

Distance

- Walk
  - A sequence of actors and relations that begins and ends with actors
- Geodesic distance
  - The number of relations in the shortest possible walk from one actor to another
- Maximum flow
  - The amount of different actors in the neighborhood of a source that lead to pathways to a target

Some Measures of Power & Prestige (Centrality Measure)

(based on Hanneman, 2001)

- Degree
  - Sum of connections from or to an actor
    - Transitive weighted degree→Authority, Hub, Influentiality
- Closeness centrality
  - Distance of one actor to all others in the network
- Betweenness centrality
  - Number that represents how frequently an actor is between other actors' geodesic paths

Cliques and Social Roles

(based on Hanneman, 2001)

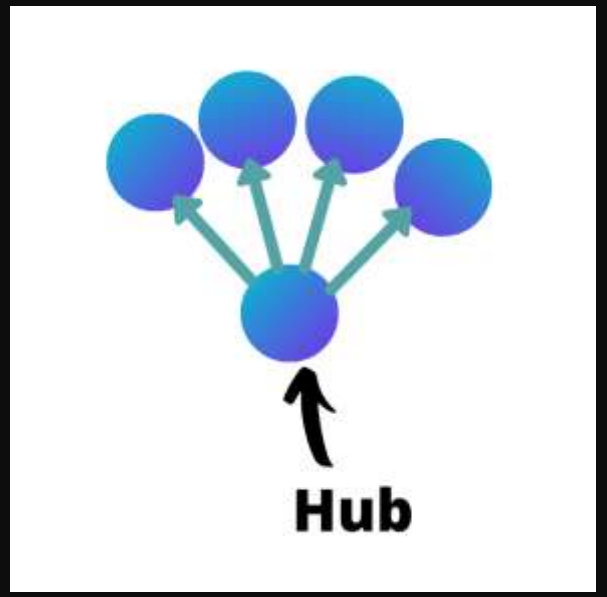
- Cliques / Gank
  - Sub-set of actors
    - More closely tied to each other than to actors who are not part of the sub-set
      - (A lot of work on “trawling” for communities in the web-graph)
      - Often, you first find the clique (or a densely connected subgraph) and then try to interpret what the clique is about
  - Clique is a pairing of three or more nodes
  - Dyad is a pairing of two nodes

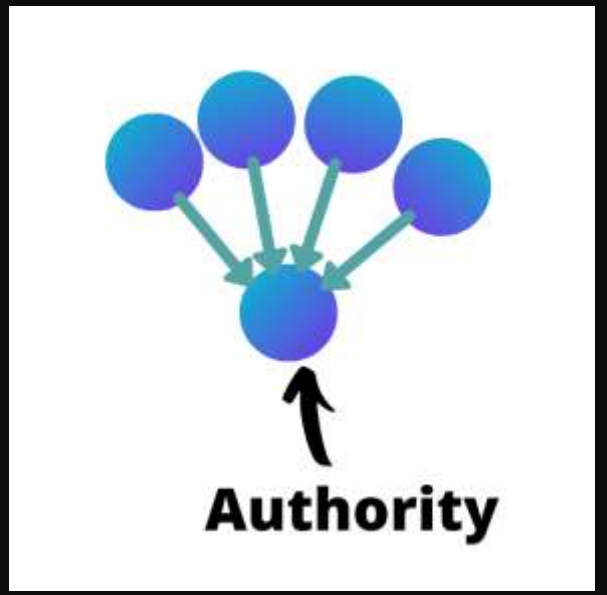
Social roles

- Defined by regularities in the patterns of relations among actors

Hubs and Authority

- Hubs and Authorities are node classifications used in directed networks
- hub is a node that has many edges pointing out of it.
- authority, on the other hand, is a node that has many edges pointing to it

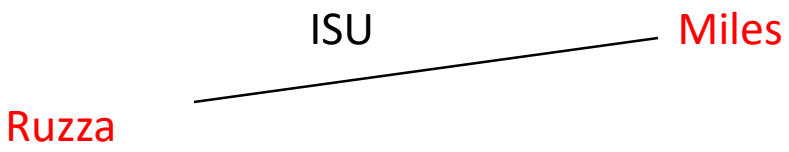




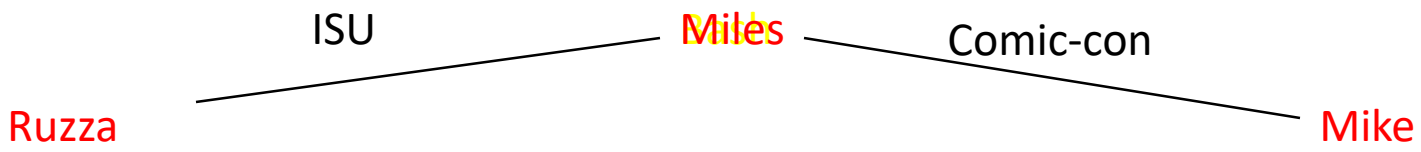
Trying to make friends

Ruzza

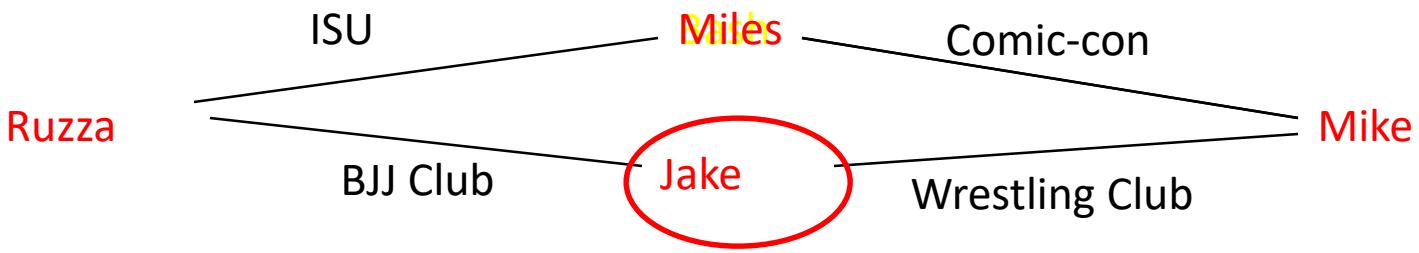
Trying to make friends



Trying to make friends

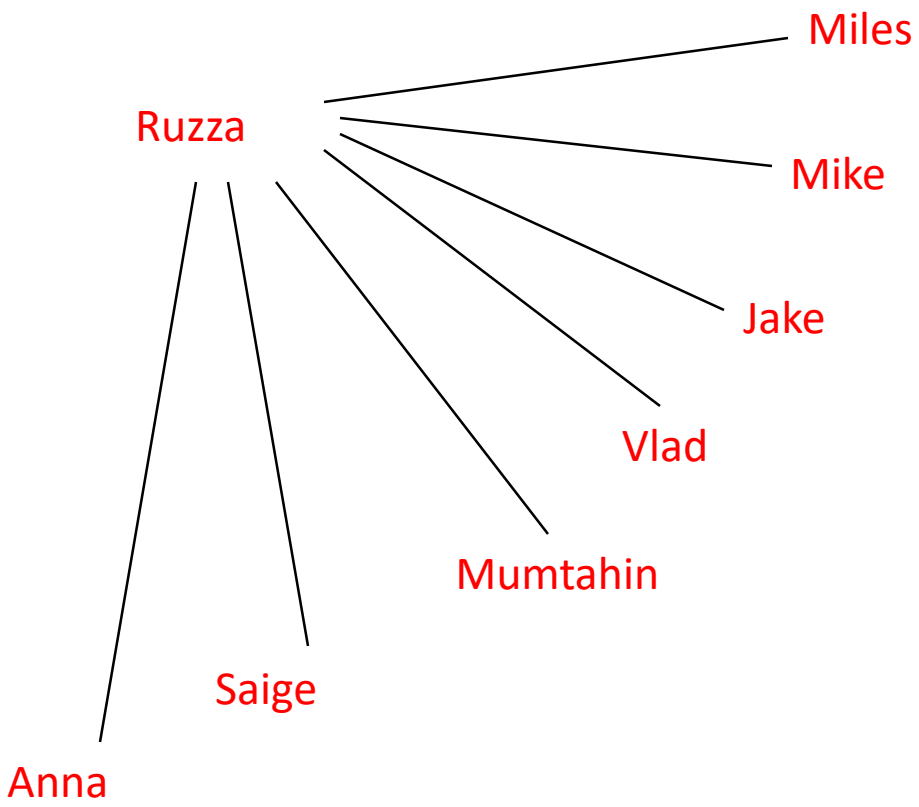


Trying to make friends

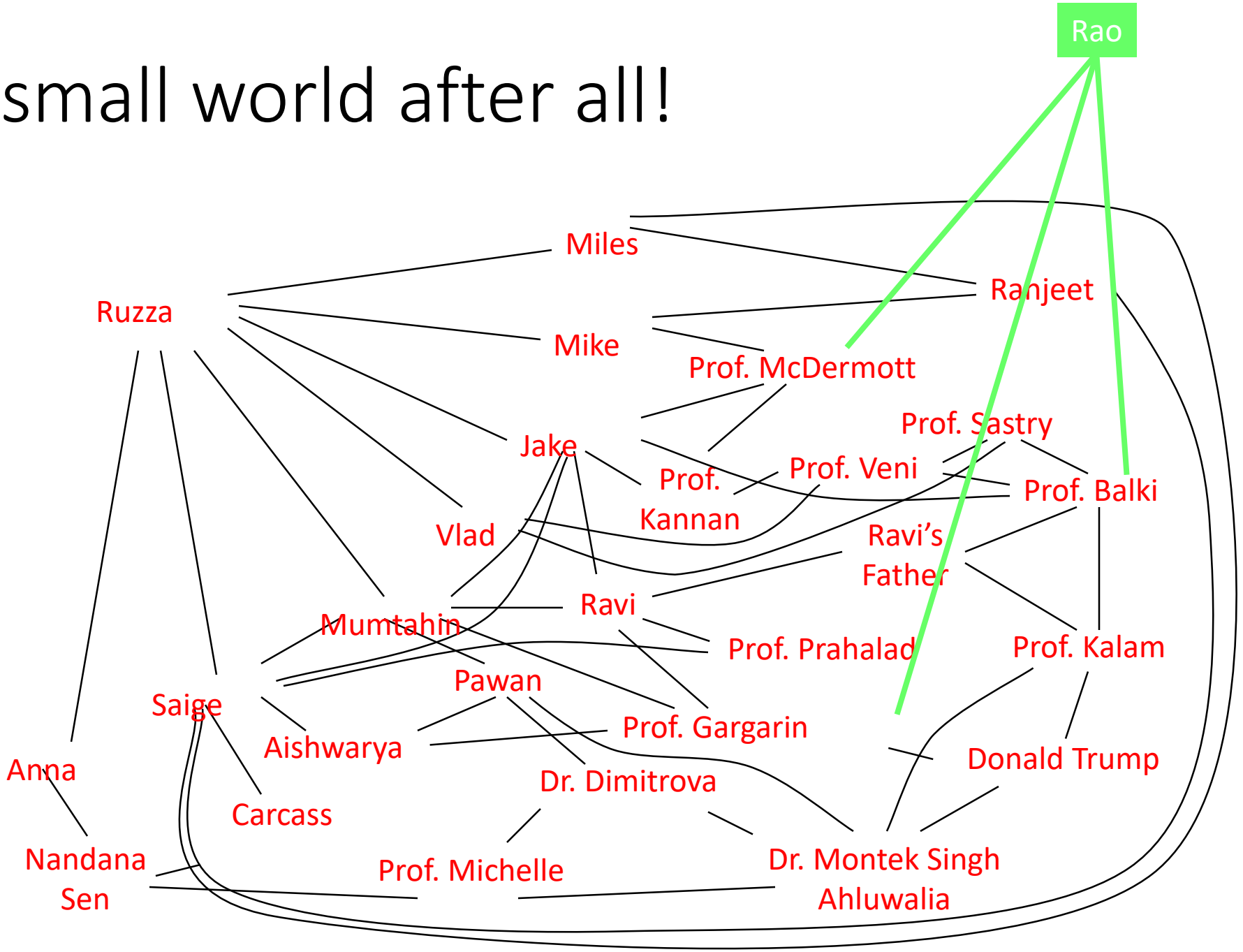


Mike and I already had a friend in common!

I didn’t have to worry…



It’s a small world after all!



Six Degrees of Separation

Milgram (1967)

The experiment:

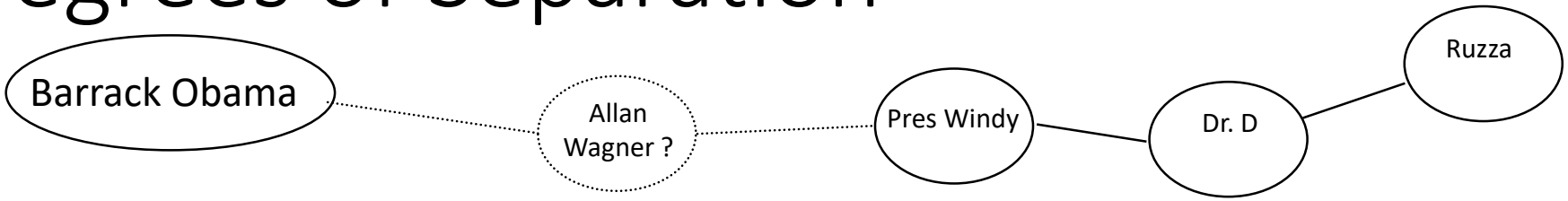
- Random people from Nebraska were to send a letter (via intermediaries) to a stock broker in Boston.

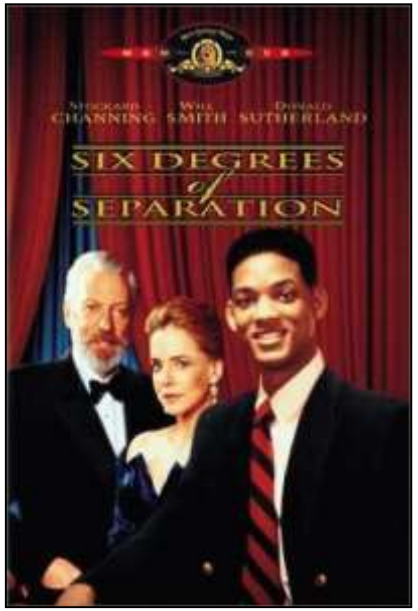
- Could only send to someone with whom they were on a first-name basis.

Among the letters that found the target, the average number of links was six.

Stanley Milgram (1933-1984)

Six Degrees of Separation



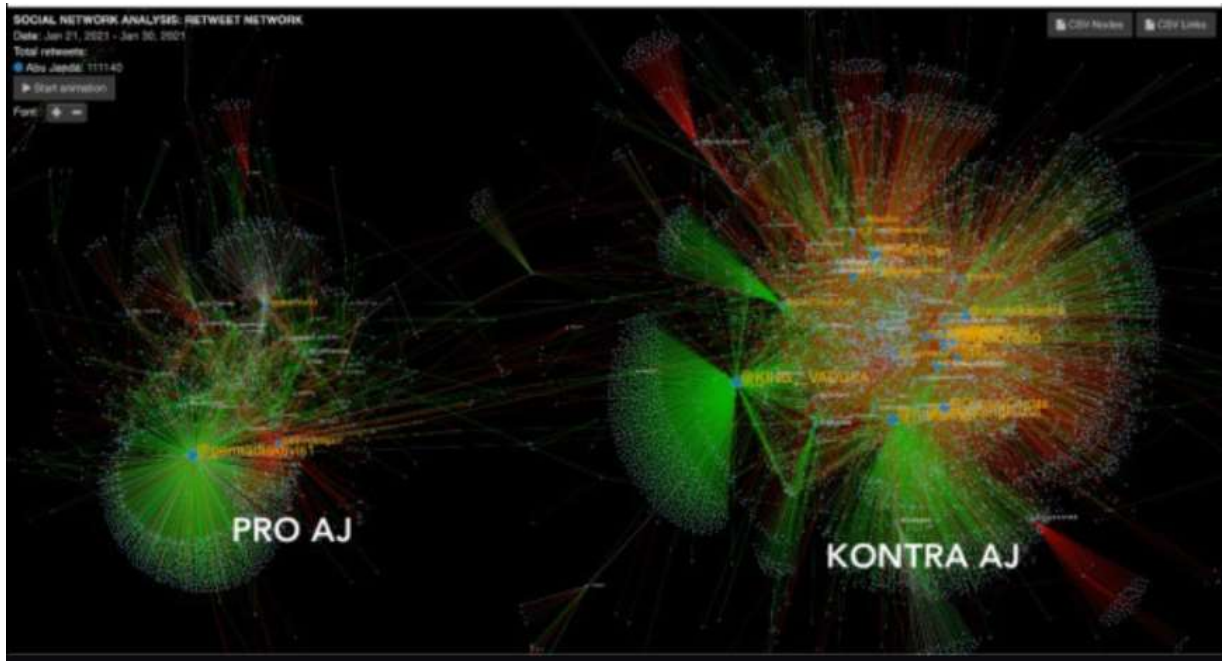


John Guare wrote a play called *Six Degrees of Separation,* based on this concept.

"*Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice… It's not just the big names. It's anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am bound to everyone on this planet by a trail of six people…*"

Twitter Community (Buzzers?) Detection

- Twitter has many communities (Kpop, Anime Weebs, Rachel Vennya Enthusiasts, Pak Antono, etc.)
- Sometimes, important issues and discourse is clouded by these communities, using trending hashtags, etc
- These communities can also sometime act as paid buzzers.



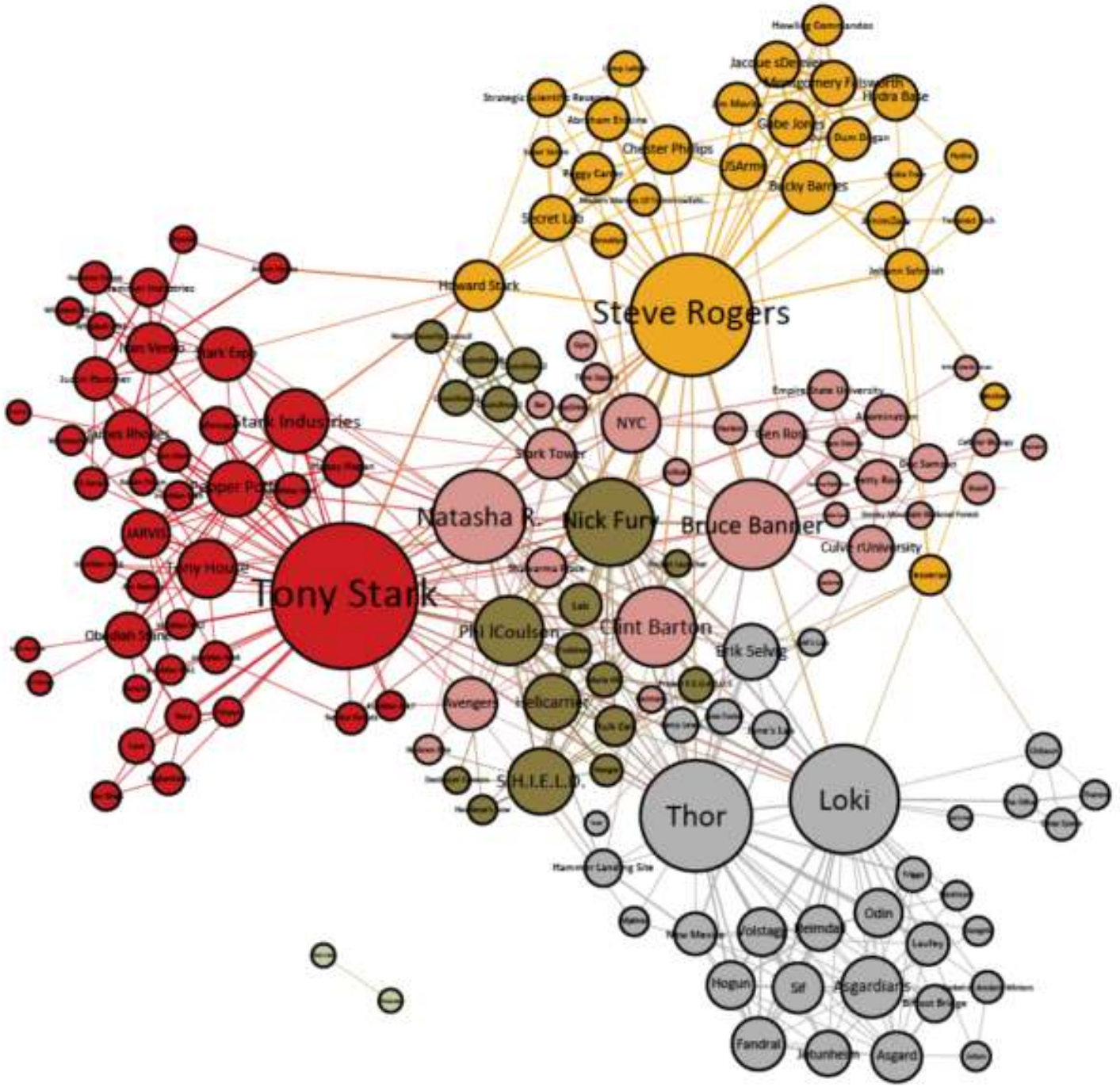
Example: SNA of Marvel Heroes Universe

- Data from Kaggle
- How can we interpret the networks?
- By which measures?
- What does it mean?
- What can we learn from it?

Marvel Heroes

- Who knows them all?
- Who is “the man” of them all?
- Who is the most popular among heroes?

What can you infer from this?

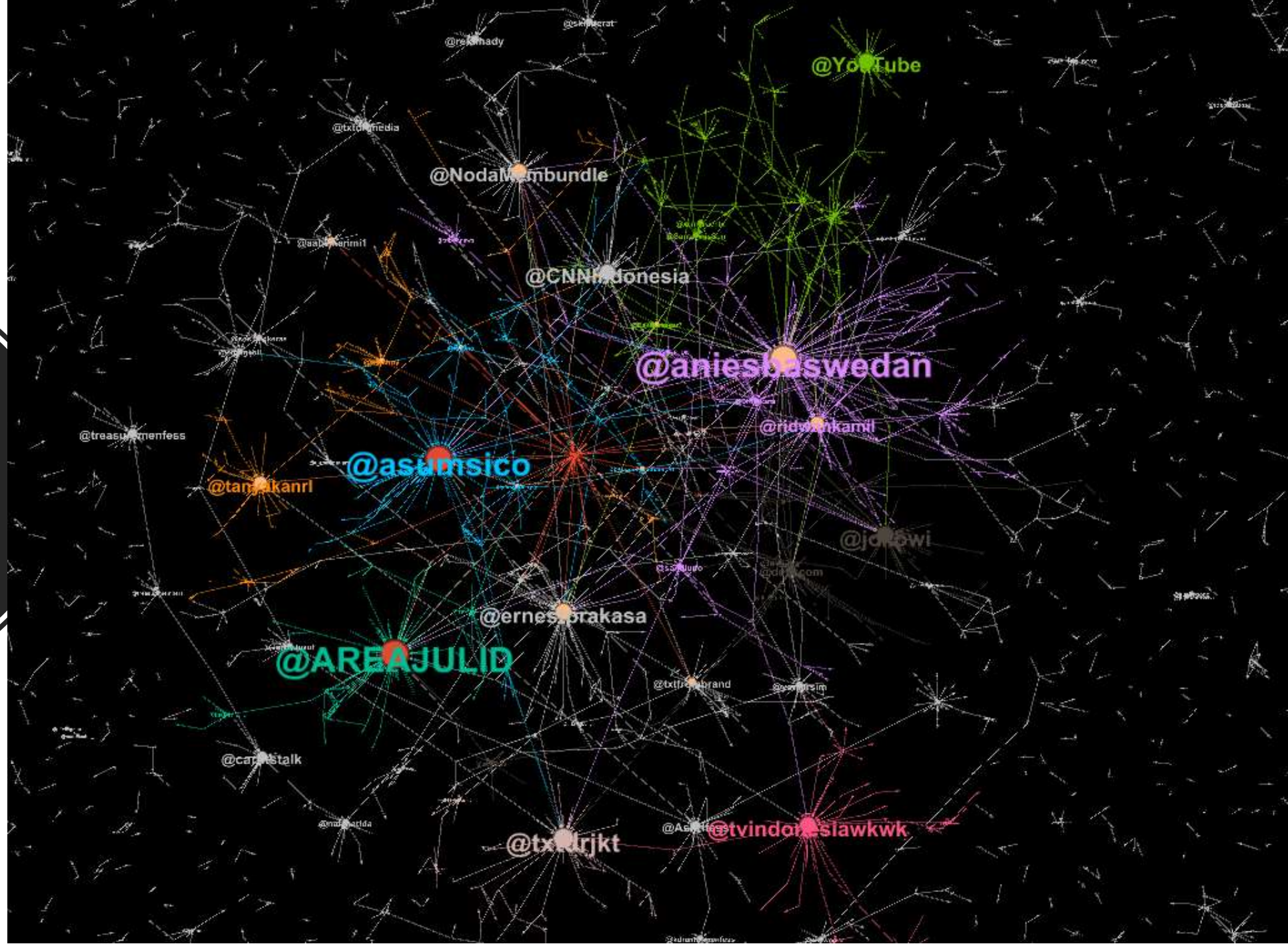


You can also use networkX function

```
cent_df = pd.DataFrame(index=list(marvel_net.nodes()))

# pagerank
cent_ = nx.pagerank(marvel_net, weight='weight')
cent_df['w_pagerank_cent'] = pd.Series(index=[k for k, v in cent_.items()], data=[float(v) for k, v
in cent_.items()])

# eigenvalue centrality
cent_ = nx.eigenvector_centrality(marvel_net, weight='weight')
cent_df['w_eigenvector_cent'] = pd.Series(index=[k for k, v in cent_.items()], data=[float(v) for k,
v in cent_.items()])

# degree centrality
cent_ = {h:0.0 for h in marvel_net.nodes()}
for u, v, d in marvel_net.edges(data=True):
    cent_[u]+=d['weight']; cent_[v]+=d['weight'];
cent_df['w_degree_cent'] = pd.Series(index=[k for k, v in cent_.items()], data=[float(v) for k, v in
cent_.items()])

# closeness centrality
temp_net = marvel_net.copy()
for u,v,d in temp_net.edges(data=True):
    if 'distance' not in d:
        d['distance'] = 1.0/d['weight']
cent_ = nx.closeness_centrality(temp_net, distance='distance')
cent_df['w_closeness_cent'] = pd.Series(index=[k for k, v in cent_.items()], data=[float(v) for k, v
in cent_.items()])

# betweenness centrality
cent_ = nx.betweenness_centrality(marvel_net, weight='weight')
cent_df['w_betweenness_cent'] = pd.Series(index=[k for k, v in cent_.items()], data=[float(v) for k,
v in cent_.items()])

display(cent_df)
cent_df = cent_df.drop(columns=['w_betweenness_cent'])
```

	w_pagerank_cent	w_eigenvector_cent	w_degree_cent	w_closeness_cent	w_betweenness_cent
SPIDER-MAN	0.050127	0.145800	2874.0	95.334005	0.000000
CAPTAIN AMERICA	0.066699	0.312066	4409.0	141.338880	0.000000
IRON MAN	0.054943	0.271015	3600.0	118.305766	0.000000
THING	0.057688	0.318561	3828.0	104.086414	0.000000
THOR	0.046146	0.231617	2969.0	108.986186	0.000000
HUMAN TORCH	0.056935	0.316594	3776.0	102.769509	0.000000
MR. FANTASTIC	0.055603	0.313944	3689.0	99.228022	0.000000
HULK	0.029614	0.112058	1674.0	79.893904	0.000000
WOLVERINE	0.033988	0.099205	1929.0	71.713747	0.000000
INVISIBLE WOMAN	0.052618	0.301620	3478.0	96.400542	0.000000
SCARLET WITCH	0.050598	0.250119	3297.0	115.518830	0.000000
BEAST	0.038460	0.142554	2303.0	95.134670	0.000000
DR. STRANGE	0.022348	0.074948	1159.0	57.281086	0.000000

Citayam fashion week Interaction Network



Id	Label	Timestamp	Modularity Class	Eigenvector Centrality	PageRank
@AREAJULID	@AREAJULID		1248	0.582564	0.00411
@aniesbaswedan	@aniesbaswedan		985	1.0	0.004105
@asumsico	@asumsico		321	0.663069	0.003825
@txtdrjkt	@txtdrjkt		326	0.427007	0.003032
@jokowi	@jokowi		45	0.357684	0.00261
@tvindonesiawkwk	@tvindonesiawkwk		1291	0.339744	0.002534
@ernestprakasa	@ernestprakasa		250	0.483773	0.002519
@CNNIndonesia	@CNNIndonesia		237	0.364013	0.002458
@YouTube	@YouTube		105	0.271563	0.002444
@NodaMembundle	@NodaMembundle		1284	0.344752	0.002309
@tanyakanrl	@tanyakanrl		61	0.260508	0.002054
@ridwankamil	@ridwankamil		985	0.485788	0.002014
@convomf	@convomf		15	0.206014	0.00185
@caratstalk	@caratstalk		926	0.181272	0.001686
@convomfs	@convomfs		327	0.177451	0.001529
@Askrlfess	@Askrlfess		721	0.198091	0.001454
@treasuremenfess	@treasuremenfess		926	0.157501	0.001413
@detikcom	@detikcom		45	0.258847	0.001406
@nctzenbase	@nctzenbase		1291	0.142625	0.001343
@FWBESS	@FWBESS		273	0.156529	0.001307

Any Questions?

Tugas Kelompok

- Kumpulkan minimal 15 artikel ilmiah yang menggunakan social network analysis sebagai metode utama dan membahas mengenai topik atau isu yang sebidang
- Buat literature review untuk Kumpulan artikel tersebut
- Kumpulkan dalam PDF

- **Struktur Literatur Review**:
- **Pendahuluan**: Berikan pengantar mengenai pentingnya *social media mining* dan bagaimana *text clustering* dan *SNA* digunakan untuk menganalisis data dari media sosial.
- **Text Clustering**:
  - Jelaskan dasar teori *text clustering*, termasuk algoritma yang sering digunakan
  - Jelaskan metodologi umum yang digunakan untuk *text preprocessing*
  - Uraikan tantangan yang sering dihadapi dalam *text clustering* pada data media sosial (misalnya, data teks yang tidak terstruktur, ambiguitas bahasa).
  - Berikan contoh kasus aplikasi *text clustering* yang dijelaskan dalam literatur yang Anda temukan.
- **Social Network Analysis (SNA)**:
  - Uraikan konsep dasar *social network analysis*
  - Jelaskan teknik visualisasi jaringan sosial dan alat yang sering digunakan, seperti Gephi atau NetworkX.
  - Bahas tantangan dalam melakukan *SNA* pada data media sosial (misalnya, kompleksitas jaringan yang besar, dinamika hubungan antar pengguna).
  - uraikan contoh-contoh aplikasi *SNA* yang ditemukan dalam literatur yang Anda tinjau.
- **Perbandingan dan Keterkaitan**:
  - Jelaskan bagaimana *text clustering* dan *SNA* dapat saling melengkapi dalam analisis data sosial media.
  - Berikan insight mengenai kombinasi dari kedua metode ini dalam literatur yang Anda tinjau.
- **Kesimpulan**:
  - Buat kesimpulan dari literatur yang Anda tinjau, termasuk tren terkini dalam penelitian dan aplikasi dari kedua topik ini. Berikan juga rekomendasi atau saran untuk penelitian lanjutan.