Introduction to big data analytics

Network Data Analytics

Recap from Topic 7: Textual Data Analytics

❖Textual Data Analytics Application

Recap from Topic 7: Types of Textual Analytics

- **Characteristics of textual data:**
  - Unstructured
  - Building blocks are words
    - Words are not independent
  - Each text segment (e.g. sentence) encapsulates semantics behind

Syntactical Analysis: transform unstructured text to structured representation

Semantic Analysis

Recap from Topic 7: Syntactical Analysis

- Transform a textual data into a multi-dimensional vector
- **Approaches:**
  - Feature Engineering: hand-craft the features (e.g. TF-IDF)
  - Representation learning: auto-learn the features (e.g. neural embedding)
- **Applications**
  - Information retrieval: search relevant documents given a query
  - Classification: categorize a text into a pre-defined label
    - e.g. spam email detection, Gmail tabbed categories
  - ...

Recap from Topic 7: Sentiment Analysis

- Computational study of opinions, sentiments, etc., expressed in text.
  - E.g. extract from text how people feel about different products (Reviews, blogs, discussions, news, comments, feedback, …)
- Techniques:
  - Classification approach: compute labels from vector representation
  - Lexicon approach: build a dictionary of sentiment words and average their scores

Modules and Topics

01 **Data preparation and Pre-processing**

TOPIC 01 – Introduction to Big Data Analysis

TOPIC 02 – Data Preparation and Pre-processing

02 **Data Analysis and Interpretation**

TOPIC 03 – Exploratory Data Analysis

TOPIC 04 – Statistical Data Analysis

03 **Data Visualisation / Analysis of Special Types of Data**

TOPIC 05 – Visualisation and Tools

TOPIC 06 – Analysis of Time Series Data

04 **Analysis of Special Types of Data (Cont'd)**

TOPIC 07 – Analysis of Textual Data

TOPIC 08 – Analysis of Network Data

05 **Analysis with Big Data Infrastructures**

TOPIC 09 – Cloud Computing

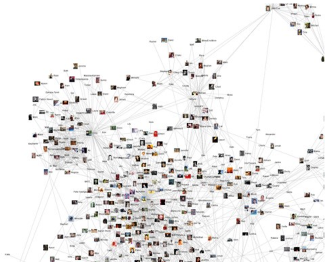
TOPIC 10 – Distributed Big Data Analysis

Learning Outcomes

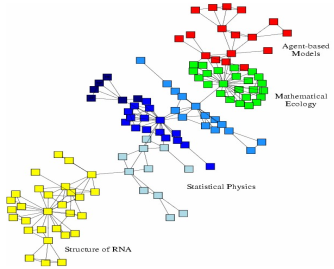
- At the end of this lecture you will be able to know:
  - Graph Representation
  - Graph Applications
  - Graph Analysis Techniques

Graph Representation

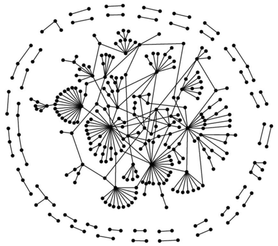
Many Data are Graphs



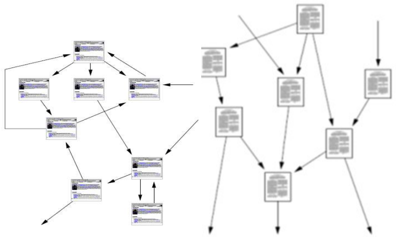
Social networks



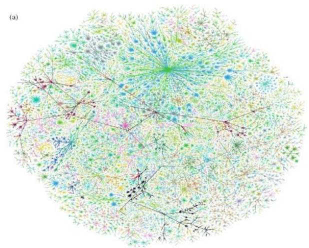
Economic networks



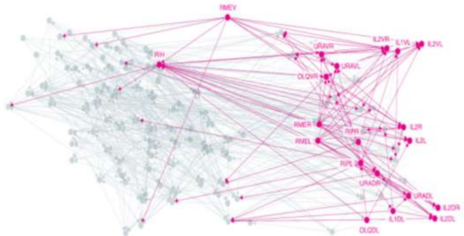
Biomedical networks



Information networks:
Web & citations



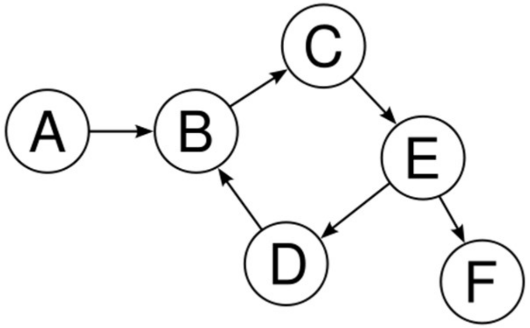
Internet



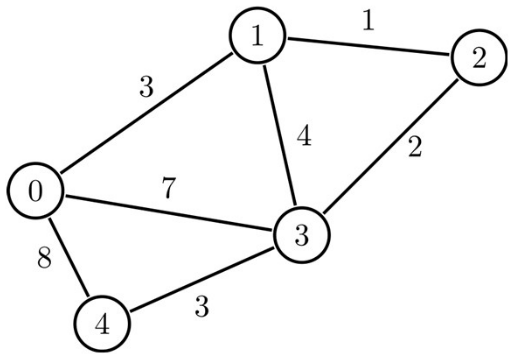
Networks of neurons

Graph Definition

- Generic Graph
  - A graph G=(V,E) is composed of two sets: a set of vertices V and a set of edges E
- Directed Graph
  - Each edge is an ordered pair of vertices
- Weighted Graph
  - Each edge has a numeric weight w



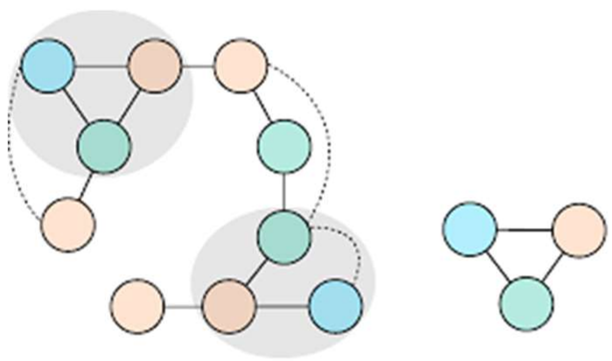
https://computersciencewiki.org/index.php/The_web_as_a_directed_graph



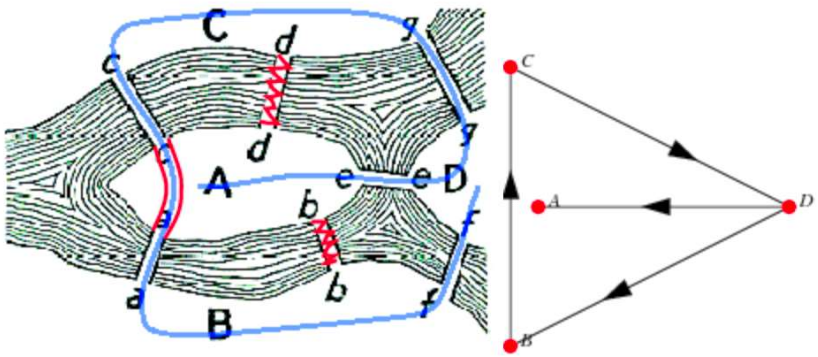
https://hyperskill.org/learn/step/5645

Why Analyze Graphs?

- Extraction of insightful and actionable knowledge



Frequent subgraph



Konigsberg bridge problem as Hamilton cycle

https://github.com/ehab-abdelhamid/GraMi

https://www.analyticsvidhya.com/blog/2018/04/introduction-to-graph-theory-network-analysis-python-codes/

Graph Applications

Applications

- Web Graph
- Social Network Graph
- Cybersecurity Graph
- Healthcare Graph
- Entertainment Graph
- And many more

Web Graph

- Node
  - Web Pages
- Edge
  - Hyperlinks
- Application
  - Identify authorities and hubs
  - Provide more accurate search services

http://farrall.org/papers/webgraph_as_content.html

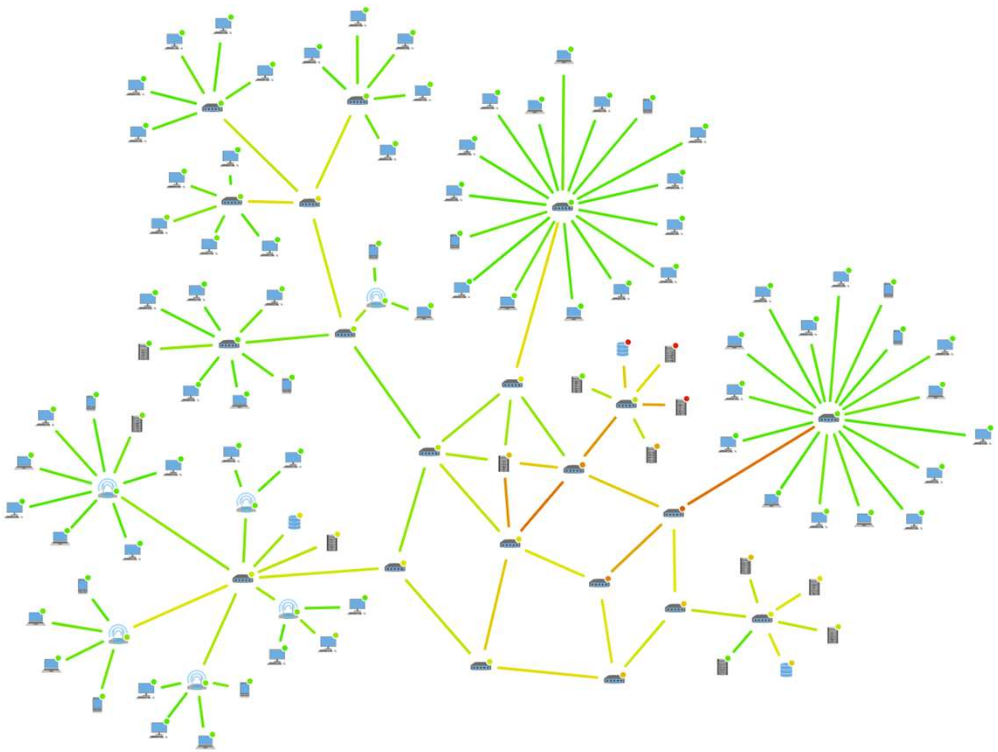
Social Network Graph

- Node
  - People or accounts
- Edge
  - Friendship or followership
- Application
  - Identify the most influential people
  - Recommend friends
  - Conduct marketing campaigns

https://associationsnow.com/2018/04/study-ceos-diverse-social-networks-see-success/

Cybersecurity Graph

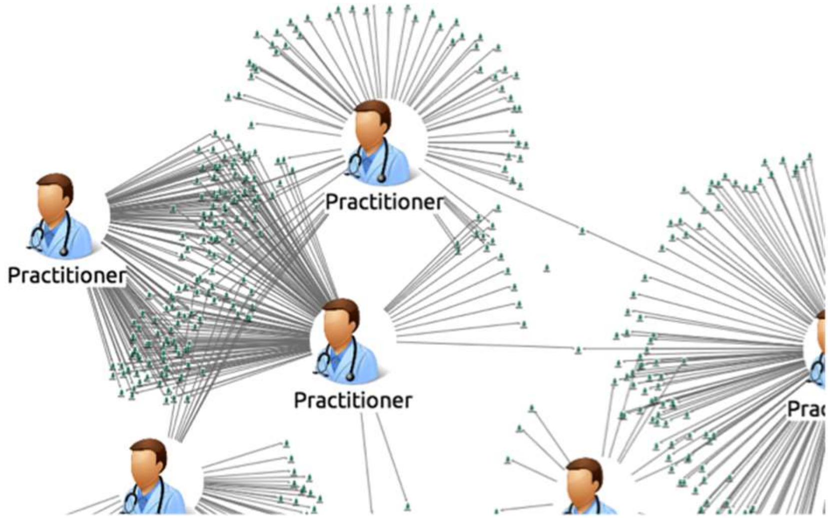
- Node
  - Computers
- Edge
  - Message traffic
- Application
  - Provide knowledge of computer viruses propagation
  - Identify intruder machines
  - Predict computers without proper authorization



https://www.yworks.com/pages/network-monitoring-visualization

Healthcare Graph

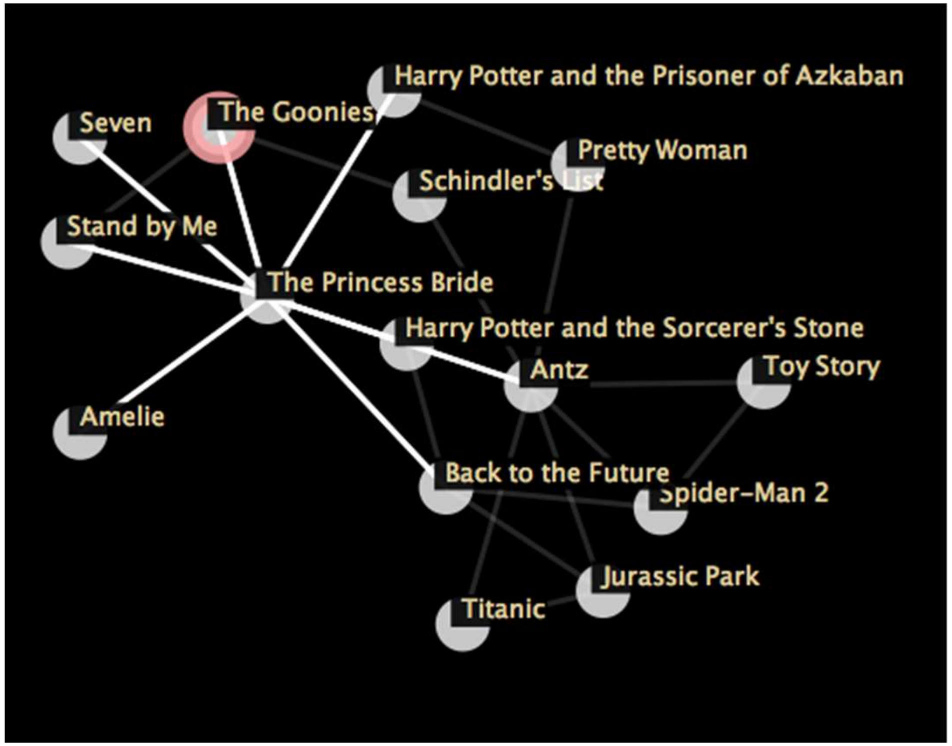
- Node
    - People (lawyers, customers, doctors, etc.)
- Edge
    - Names being present together in a claim
- Application
    - Detect groups of people collaborating to submit fraudulent claims



https://cambridge-intelligence.com/detecting-healthcare-fraud-graph-visualization

Entertainment Graph

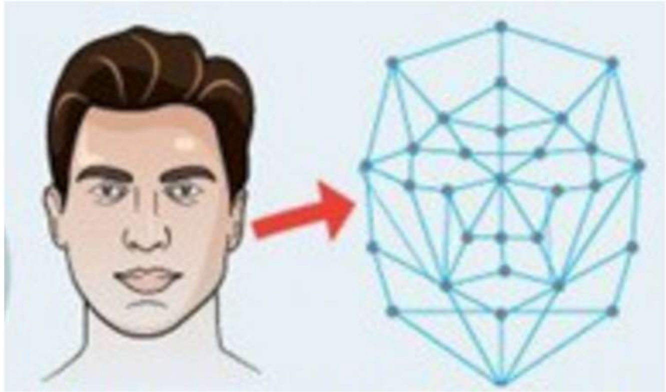
- Node
  - Movies
- Edge
  - Movies share same audience
- Application
  - Predict of upcoming movie popularity
  - Distinguish popular movies from poorly ranked movies
  - Discover key factors in determining whether a movie will be nominated for awards

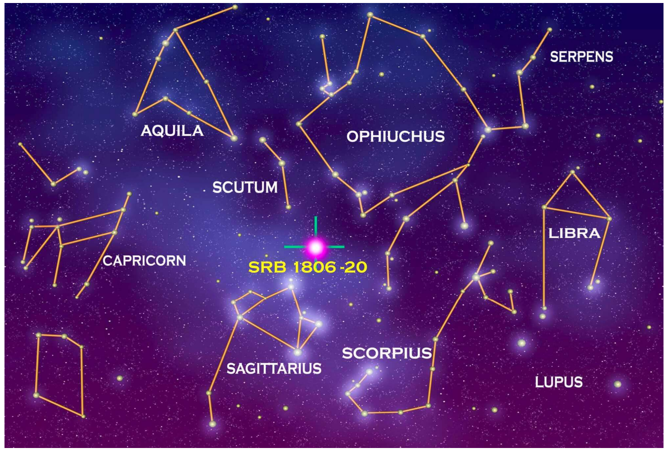


http://khreda.com/vis/graphlix/

And Many More

- Facial Graph:
  - Divide a face into multiple sections
  - Each fiducial point is a node
  - Edge connects two sections
- Star Constellation





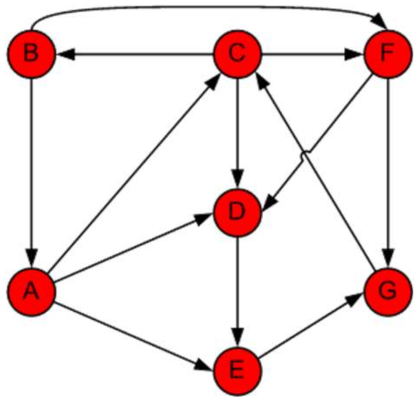
Graph Analysis Techniques

Graph Analysis Techniques

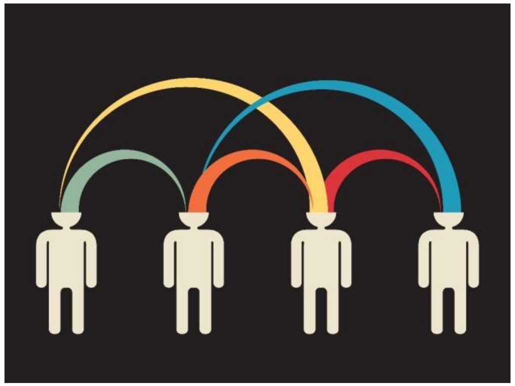
1. Centrality
2. Link Prediction
3. Network alignment
4. Network Classification
5. Node Classification

1. Centrality

- What is centrality?
  - Identify the central figures (influential individuals) in the network



- Why centrality? a measure of influence
  - The act or power of producing an effect without apparent exertion of force or direct exercise of command



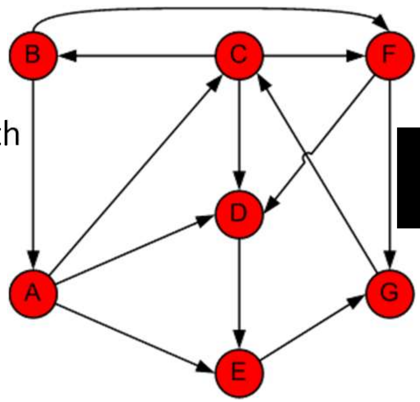
Degree Centrality

- **Question:** who is the most important?

- **Degree centrality (DC):** ranks nodes with more connections higher in terms of centrality

$$C_d(v_i) = d_i$$

  - where d_i is the number of neighbors (count both incoming and outgoing edges)



Rank

Node	DC	Rank
A	4	2
B	3	3
C	5	1
D	4	2
E	3	3
F	4	2
G	3	3

- **Shortcoming:** having more friends does not guarantee that someone is more important?

Eigenvector Centrality

- **Principle:** More important if neighbors are important

$$C_e(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{j,i} \times C_e(v_i)$$

where λ is a normalization factor to avoid numerical overflow

- **Shortcoming:** In directed graphs, once a node has a high centrality, it passes all its centrality along all of its out-links.
  - A recommendation letter written by important person who is easy to write for everyone vs. by slightly less important person but picky

PageRank Centrality

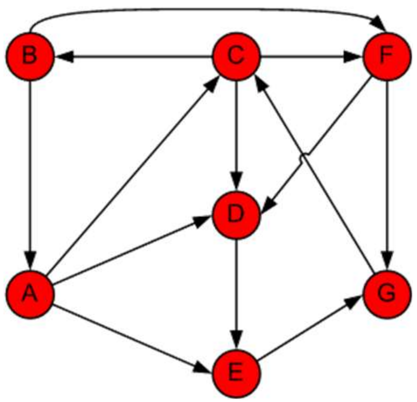
- **Principle:** you are important if your neighbors are also important (and vice-versa)

- **Technical:**
  - Divide the value of passed centrality by the number of outgoing links
  - Each connected neighbor only gets a fraction of the source node's centrality

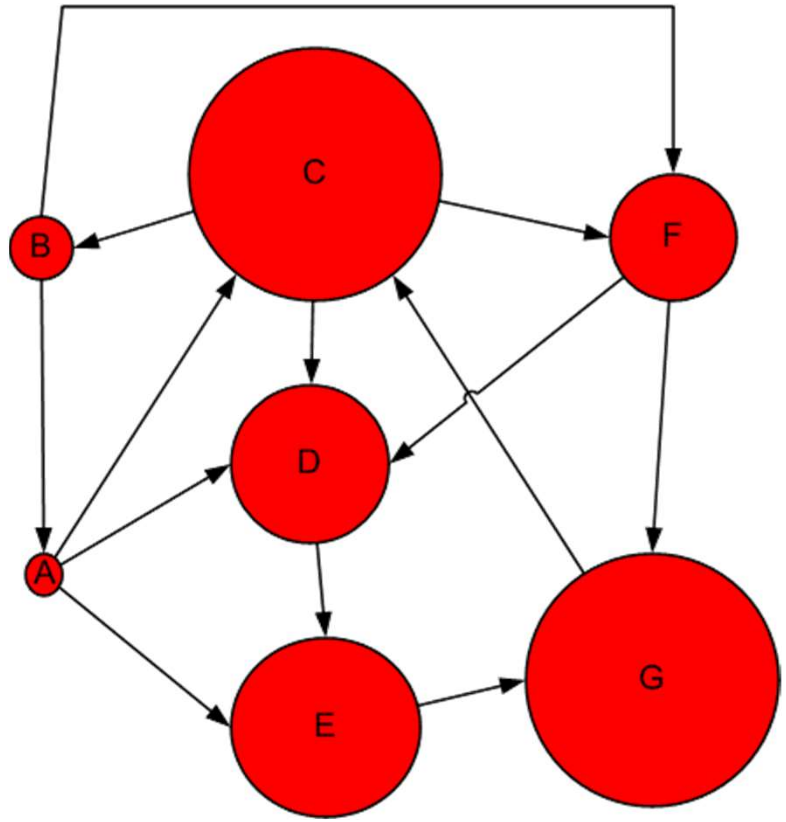
$$C_p(v_i) = \frac{1}{\lambda}\sum_{j=1}^{n} A_{j,i} \times \frac{C_p(v_j)}{d_j^{out}}$$

    - where λ is a normalization factor to avoid numerical overflow

PageRank Centrality: Example



Node	Rank
A	7
B	6
C	1
D	4
E	3
F	5
G	2



Other types of centrality

1. Centrality in terms of those who you are connected to

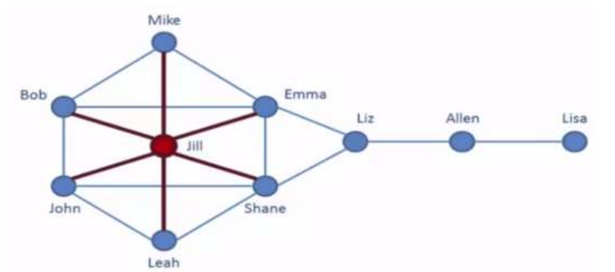
    - e.g. degree centrality, eigenvector centrality, Pagerank centrality

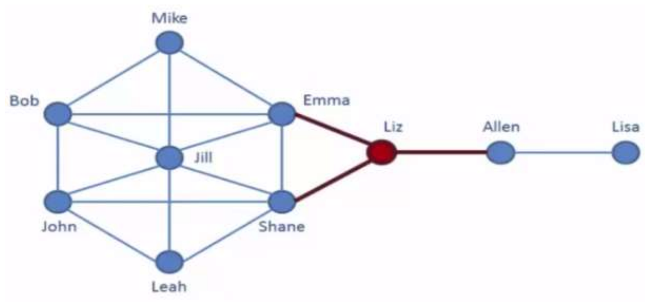
2. Centrality in terms of how you connect others

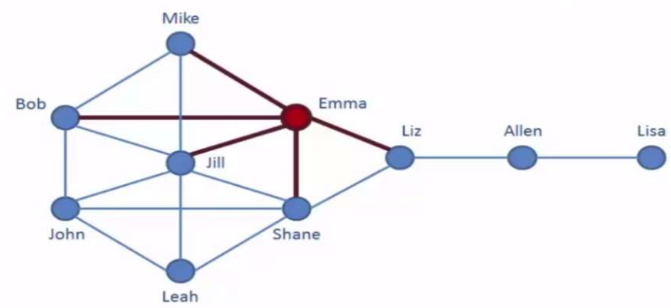
    - e.g. betweenness centrality

3. Centrality in terms of how fast you can reach others

    - e.g. closeness centrality

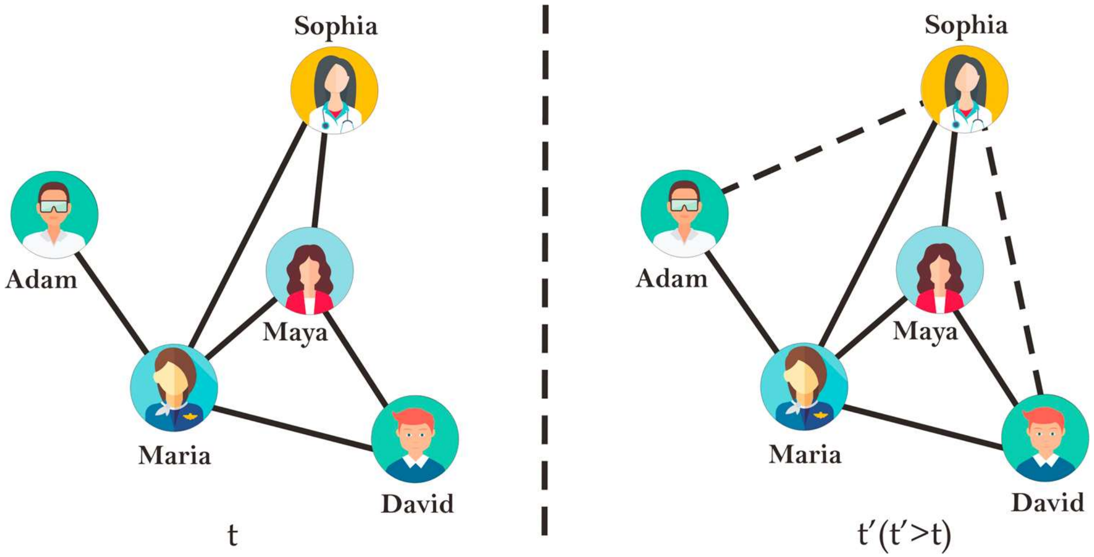






2. Link Prediction

- Predict the edges that will be added in the future
- Applications:
  - Friend suggestion
  - Collaboration prediction
  - Recommender systems



Predict friendship in the future for social networks

https://www.nature.com/articles/s41598-019-57304-y

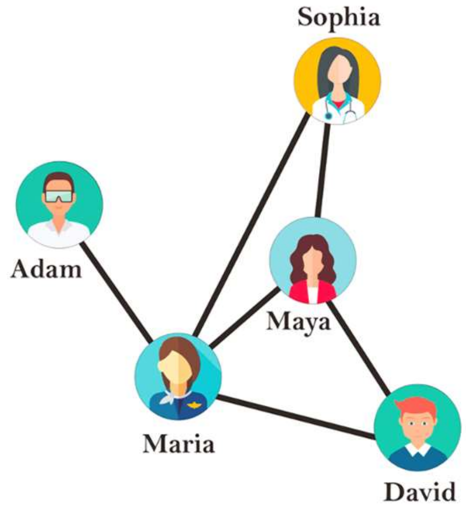
Link prediction: Preferential Attachment

Principle: the greater number of neighbors two nodes have, the more likely they will be connected in the future.

Example: two popular people are likely to meet each other

Formula:

$$PA(u, v) = |N(u)| \times |N(v)|$$



- PA (Adam, Maya) = 1x3
- PA (Adam, David) = 1x2

Conclusion: *Adam* and *Maya* is more likely to have a future interaction than *Adam and David* because Maya is more popular

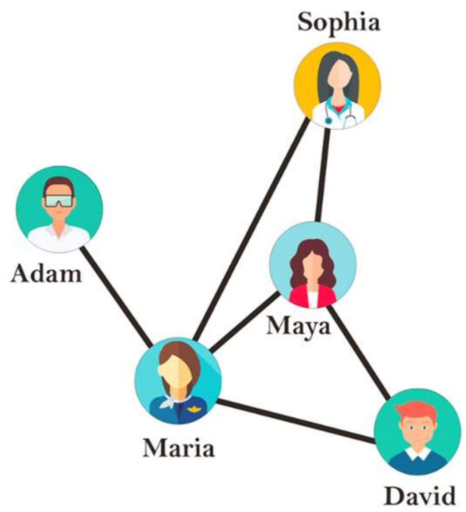
Link prediction: Common Neighbors

Principle: the more common neighbours two nodes have, the more likely they will be connected in the future.

Example: two people have the same friends are likely to be introduced to each other

Formula:

$$CN(u,v) = |N(u) \cap N(v)|$$

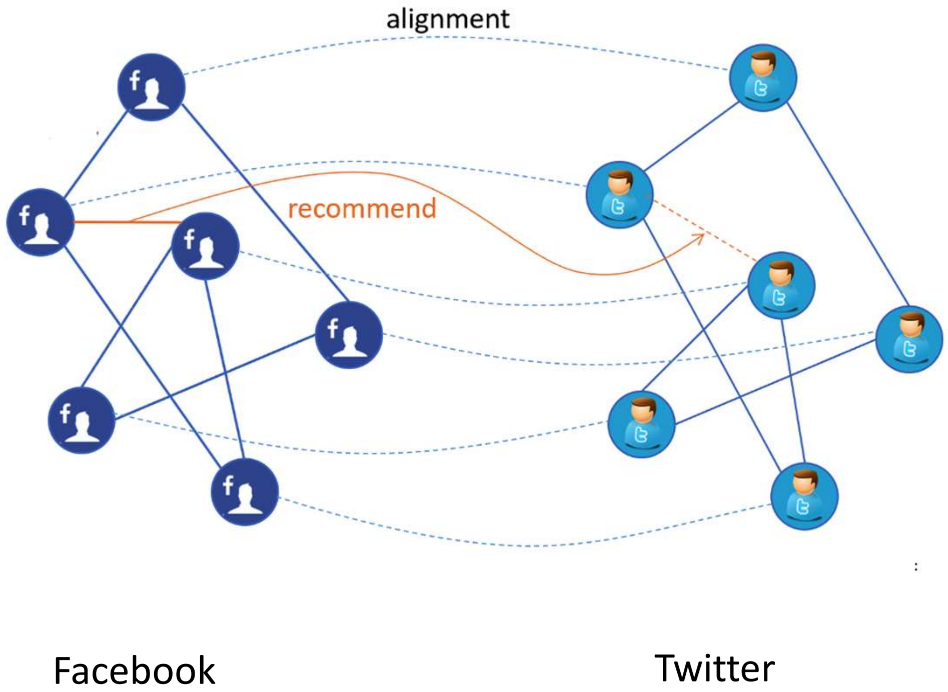


- CN (*Adam, Maya*) = 1 as they have one common neighbor {Maria}
- CN (*Adam, David*) = 1 as they have one common neighbor {Maria}

Conclusion: *Adam* and *Maya* has the same likelihood of becoming friends as *Adam* and *David.*

3. Network Alignment

- **Definition:** the task of recognizing node correspondence across different networks.
- **Applications:**
  - Friend Suggestion: if two users are friends on Facebook, suggest them to become friends on Twitter too
  - Find common groups



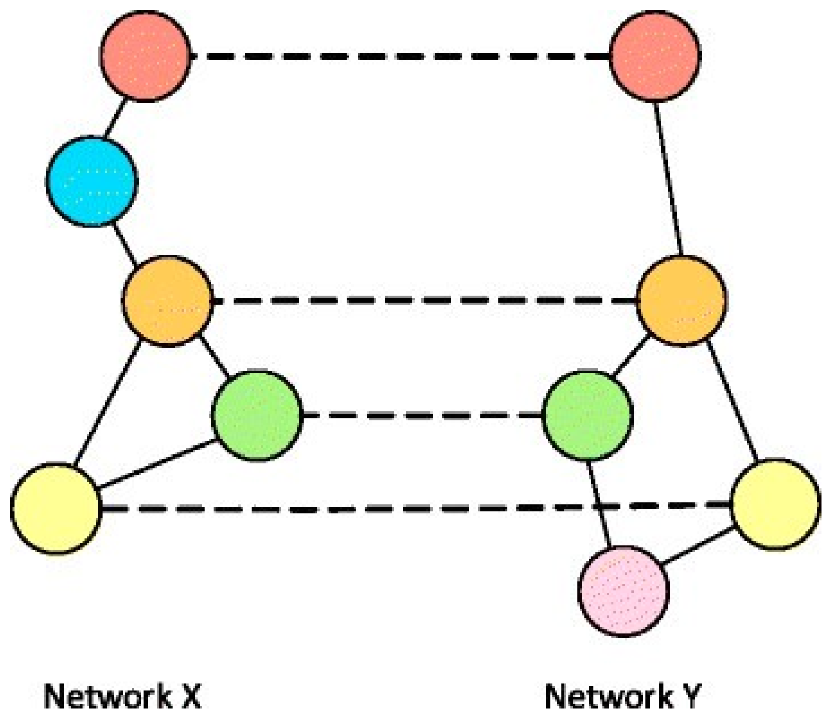
Examples: Align two accounts of the same user in social media platforms.

Network Alignment: Degree Method

- Principle: two nodes are aligned if they have similar degrees

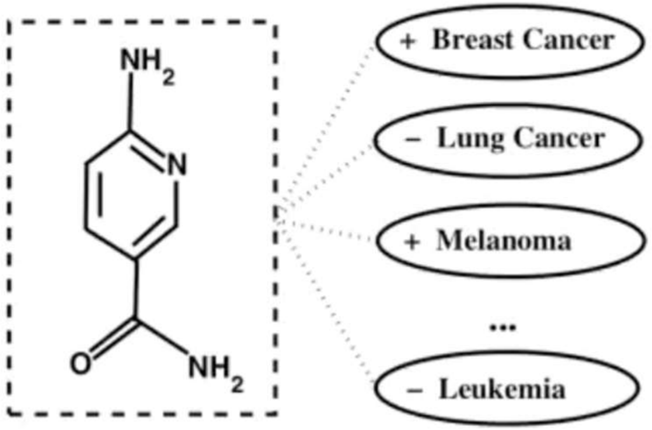
- Formula:

$$s(u,v) = |N(u) - N(v)|$$



4. Network Classification

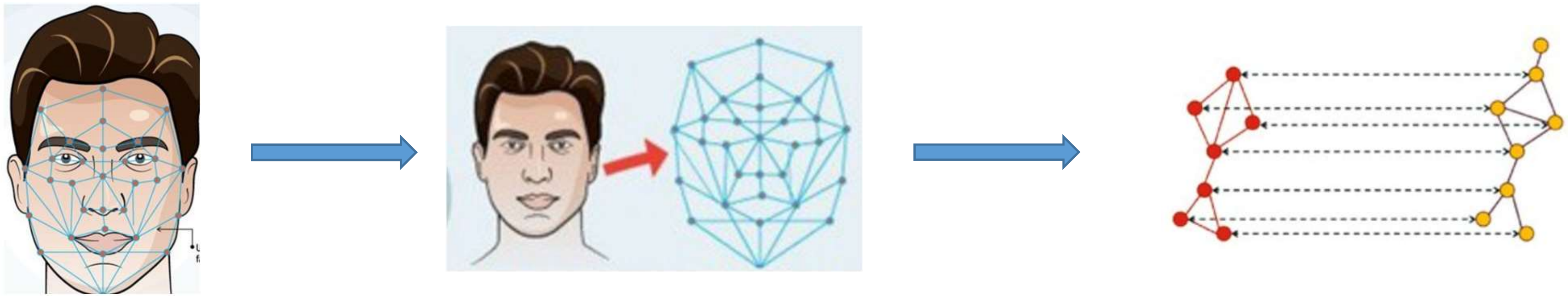
- Aka Graph Classification
  - Compute a single or multiple categories for a graph



[Chemical molecule classification]

Network Classification: Applications

- Face Recognition



- Constellation Recognition

https://medium.com/@fenjiro/face-id-deep-learning-for-face-recognition-324b50d916d1

https://towardsdatascience.com/beyond-graph-convolution-networks-8f22c403955a

Network Classification: k-NN method

Input: a set of graphs D=(G_1,G_2,…), a query graph G

Output: a label for G

1. Compute similarity between G and every $G_i \in D$:

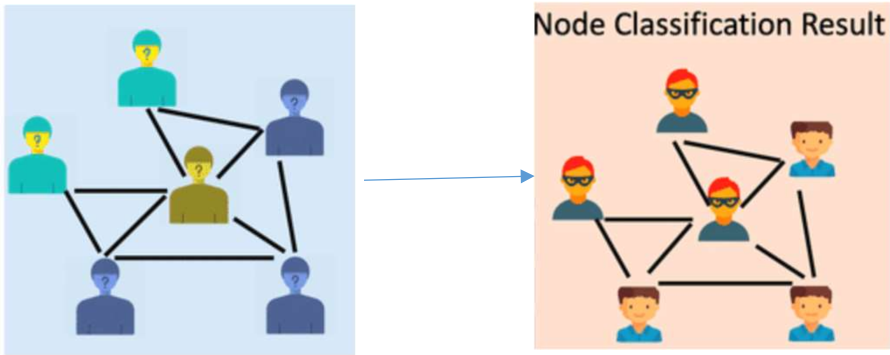
   $sim\ (G, G_i)$ = the number of alignments between G and G_i

2. Get top-k most similar graphs K for G

3. Compute the label for G by majority voting over K

5. Node Classification

The classification of individual nodes within a graph



[Scammer detection]

Node Classification: Feature Engineering

- Compute a feature vector for each node v=(f1,f2,f3,…)
  - f1: Node degree
  - f2: Centrality
  - f3: Degrees of neighbors
  - Etc.
- Use classification method: e.g. KNN
  - Compute similarity between nodes (e.g. cosine similarity)
  - Take the majority voting form k-most similar nodes

https://www.slideshare.net/jleskovec/graph-representation-learning

Summary

- In this lecture, you learnt about:
  - Graph Representation
  - Graph Applications
  - Graph Analysis Techniques

Griffith
UNIVERSITY

Queensland, Australia