



Lebanese University
Faculty of Information

Branch II

Dr. Mohamad Nagi

PhD in Data Mining - Computer Science

Social Network Analysis

Semester 6



Social Network Analysis

- › A social network is a set of entities called actors and the links connecting them.
 - › Ex: students enrolled in same courses, people and likes, etc
 - › A social network is mostly represented as a graph called sociogram
- › Social Network Analysis (SNA) is powerful because it has foundations in math/graph theory



Social Network Analysis

› Social Network for Sports





Social Network Analysis

- › SNA provides a set of tools to empirically extend our theoretical intuition of the patterns that compose a social structure.
- › SNA provides a set of relational methods for systematically understanding and identifying connections among actors.
- › SNA embodies a range of theories relating types of observable social spaces and their relation to individual and group behavior.



Social Network Analysis

- › A social network model is a set of actors and the interactions between them.
- › It is successfully represented as a graph
 - › Actors are nodes and interactions are links
 - › Actors might be connected by different kinds of ties and relations



Social Network Analysis

› A social relation is anything that links two actors.

› Examples include:

- | | |
|-----------------|---------------|
| › Kinship | Co-membership |
| › Friendship | Talking with |
| › Love | Hate |
| › Exchange | Trust |
| › Coauthor ship | Fighting |



Social Network Analysis

- › Social network analysis is a method by which one can analyze the connections across individuals or groups or institutions. That is, it allows us to examine how political actors or institutions are interrelated.
- › Network analysis allows us to examine how the configuration of networks influences how individuals and groups, organizations, or systems function.
- › It can be applied across disciplines—there are social networks, political networks, electrical networks, transportation networks, and so on.



Social Network Analysis

History (based on freeman, 2000)

- › 17th century: Spinoza developed first model
- › 1937: J.L. Moreno introduced sociometry; he also invented the sociogram
- › 1948: A. Bavelas founded the group networks laboratory at MIT; he also specified centrality



Social Network Analysis

History (based on freeman, 2000)

- › 1949: A. Rapaport developed a probability based model of information flow
- › 50s and 60s: Distinct research by individual researchers
- › 70s: Field of social network analysis emerged.
 - › New features in graph theory – more general structural models
 - › Better computer power – analysis of complex relational data sets



Social Network Analysis

History (based on freeman, 2000)

- › Strength of ties:
 - › Nominal
 - › Signed
 - › Ordinal
 - › Valued



Social Network Analysis

- › Our main argument is that a social network does not need to involve only humans as actors.
- › It is a general model that consists of actors and the links between them.
- › Actors could be any kind of entities that exist and could form groups based on some factors either internal to the model or external to the model
 - › factors could be also due to hybrid effect which is a combination of internal and external.
- › The way we decide on the links and their semantics is the main criteria for differentiating an abstract graph from a social network graph which is full of semantics.



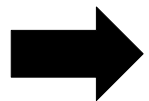
Social Network Analysis

› Social Network for Facebook



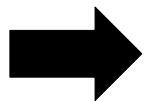


Social Network Analysis

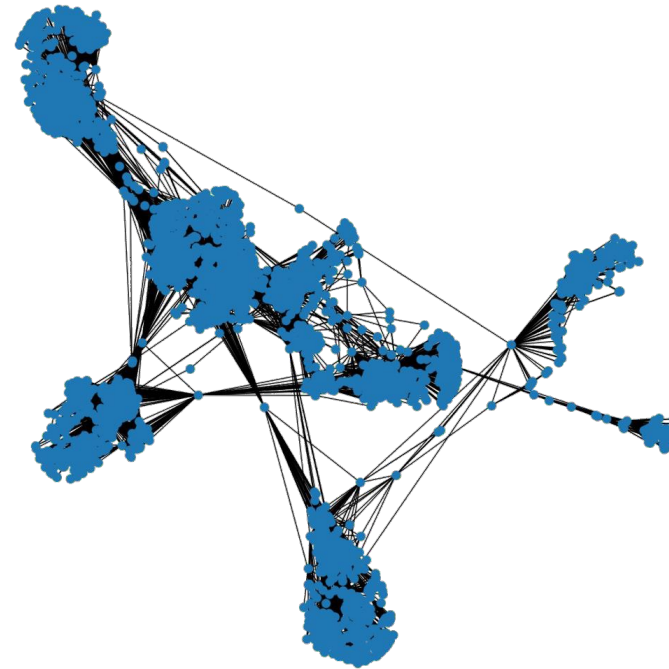


Dataset

104	113
104	122
104	123
104	128
104	142
104	169
104	186
104	188
104	200
104	203
104	212
104	239
104	252
104	271



Social
Network





Social Network Analysis

› Degree

- › Sum of connections (sum of the weights of connections in case of weighted graphs) from or to an actor

› Closeness centrality

- › Distance of one actor to all others in the network

› Betweenness centrality

- › The number of shortest paths that passes through an actor

› Eigen-vector

- › Measures how importance of an actor

Centrality Measures (Example)

Example 1

The red nodes have the highest degree centrality
The blue node has the highest Closeness and betweenness centrality

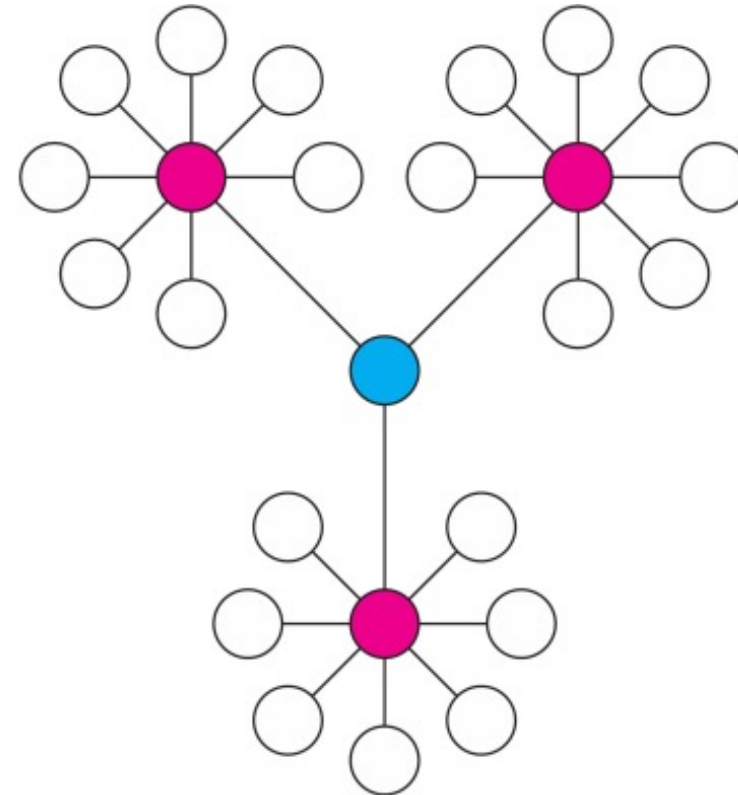


Image Reference:
<http://www.biomedcentral.com/>

Centrality Measures (Example)

Example 2

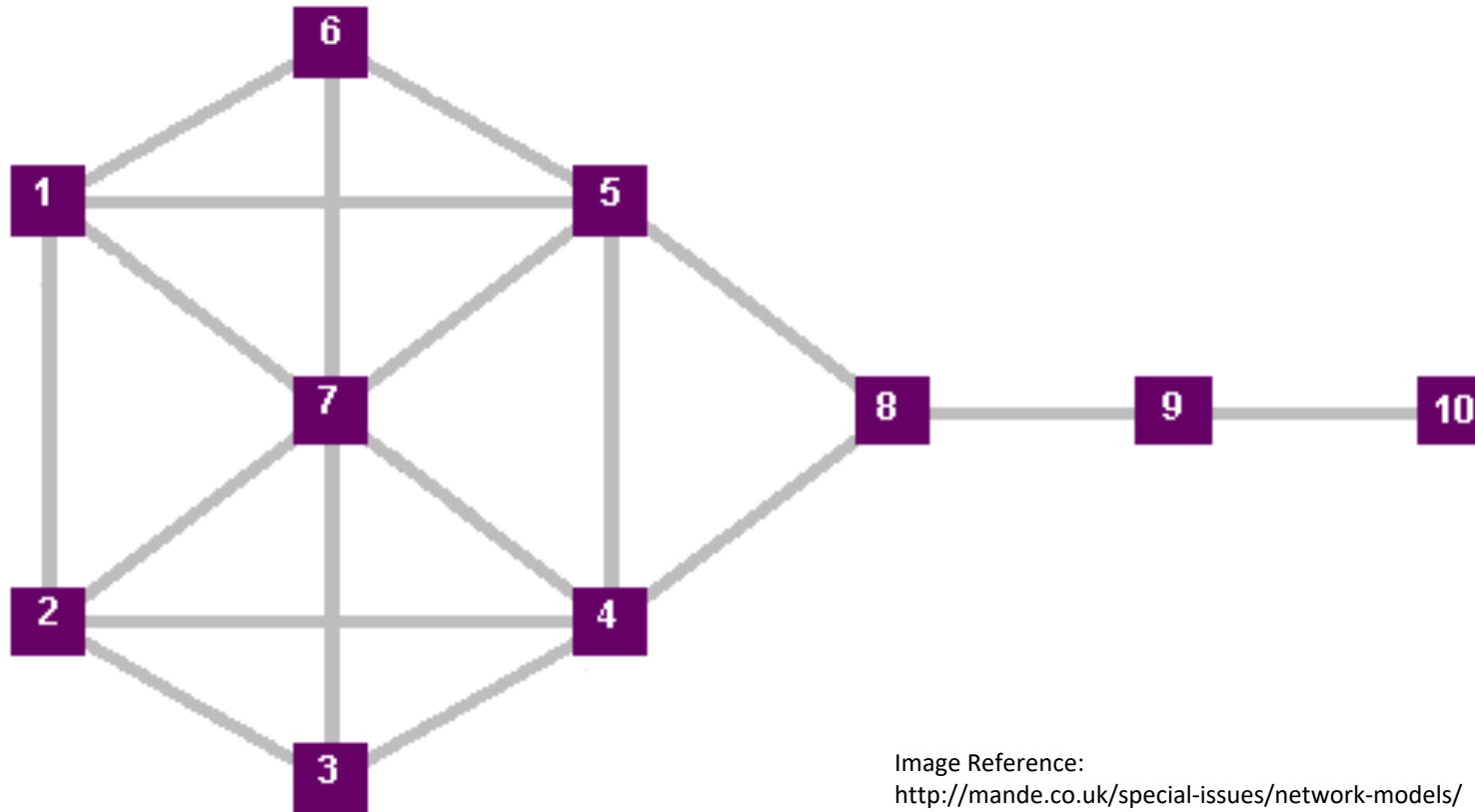
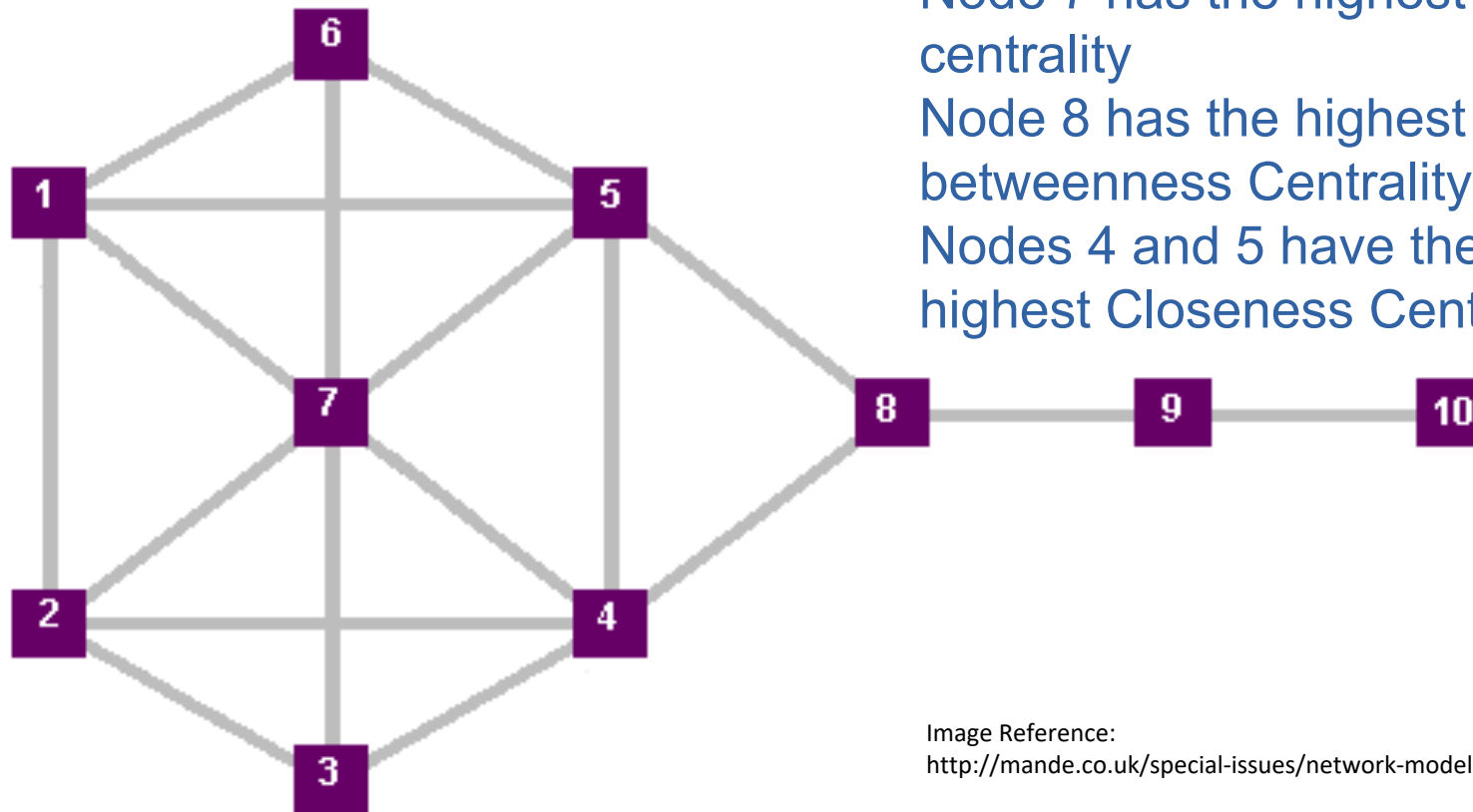


Image Reference:
<http://mande.co.uk/special-issues/network-models/>

Centrality Measures (Example)



Example 2

Node 7 has the highest degree centrality

Node 8 has the highest betweenness Centrality

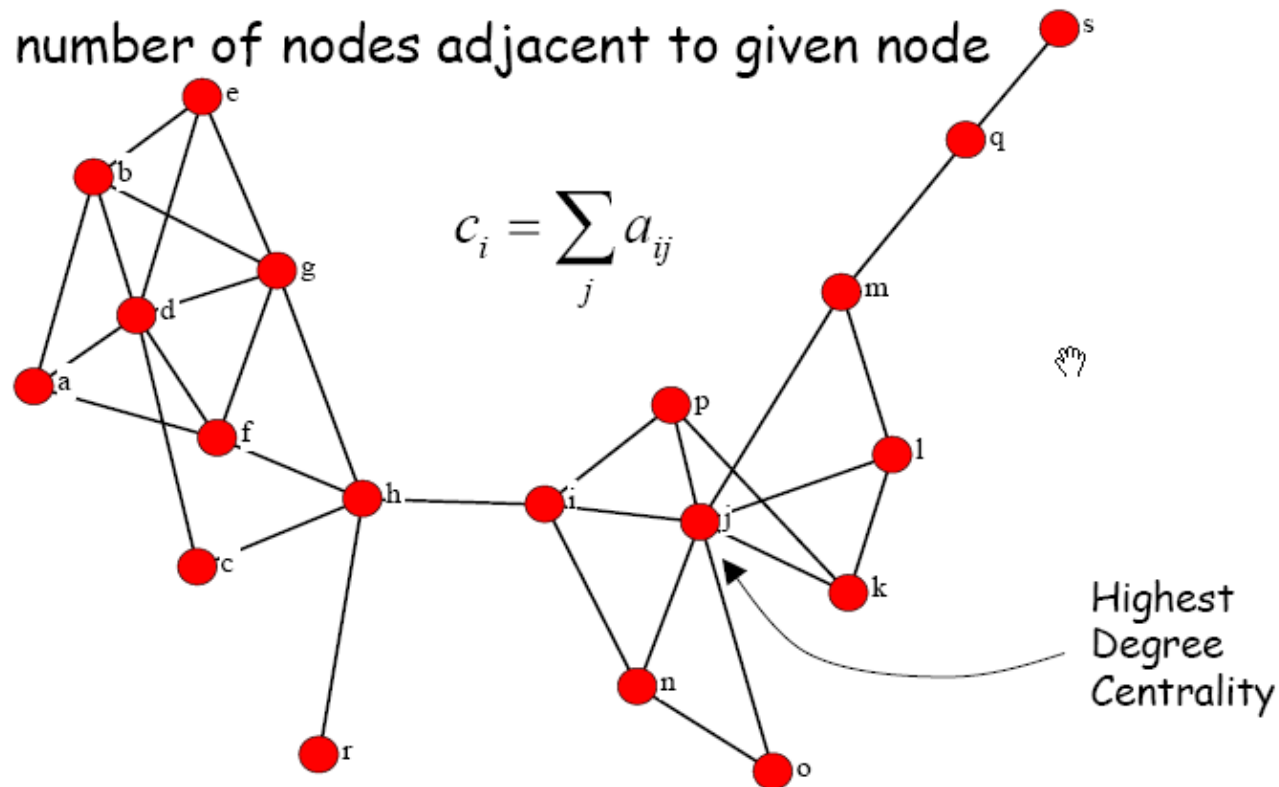
Nodes 4 and 5 have the highest Closeness Centrality

Image Reference:
<http://mande.co.uk/special-issues/network-models/>

Centrality Measures (Example)

Degree Centrality

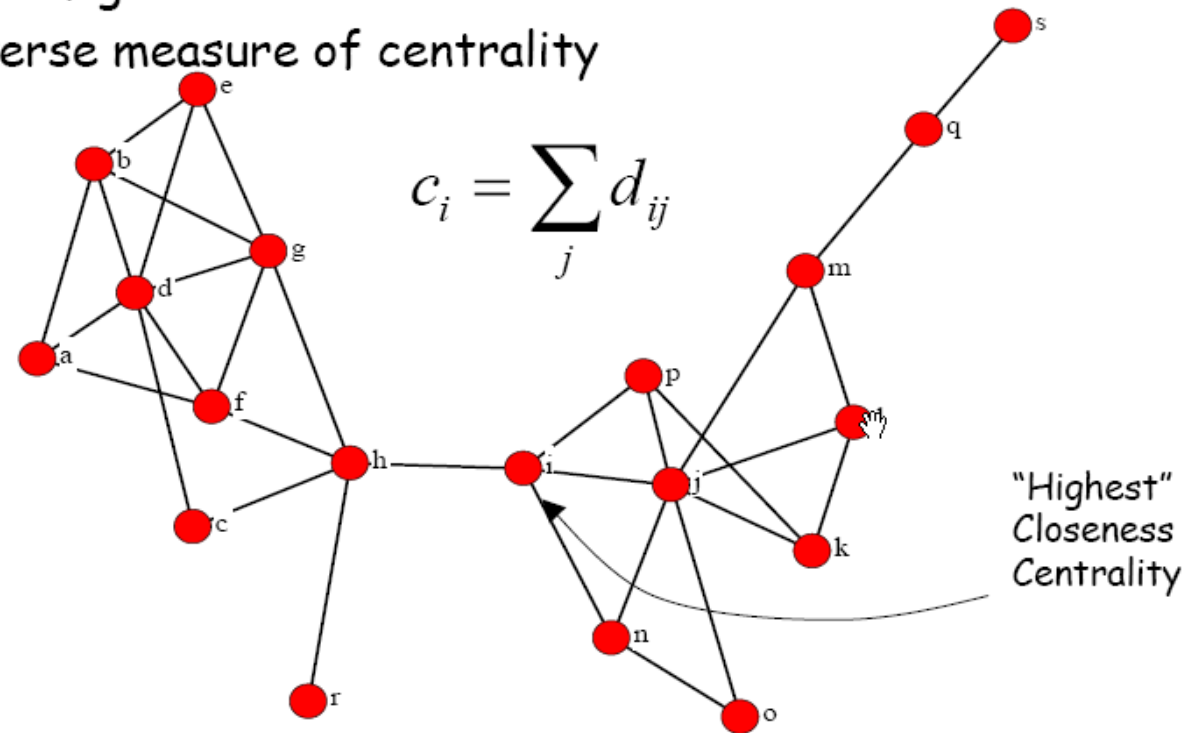
- The number of nodes adjacent to given node



Centrality Measures (Example)

Closeness Centrality

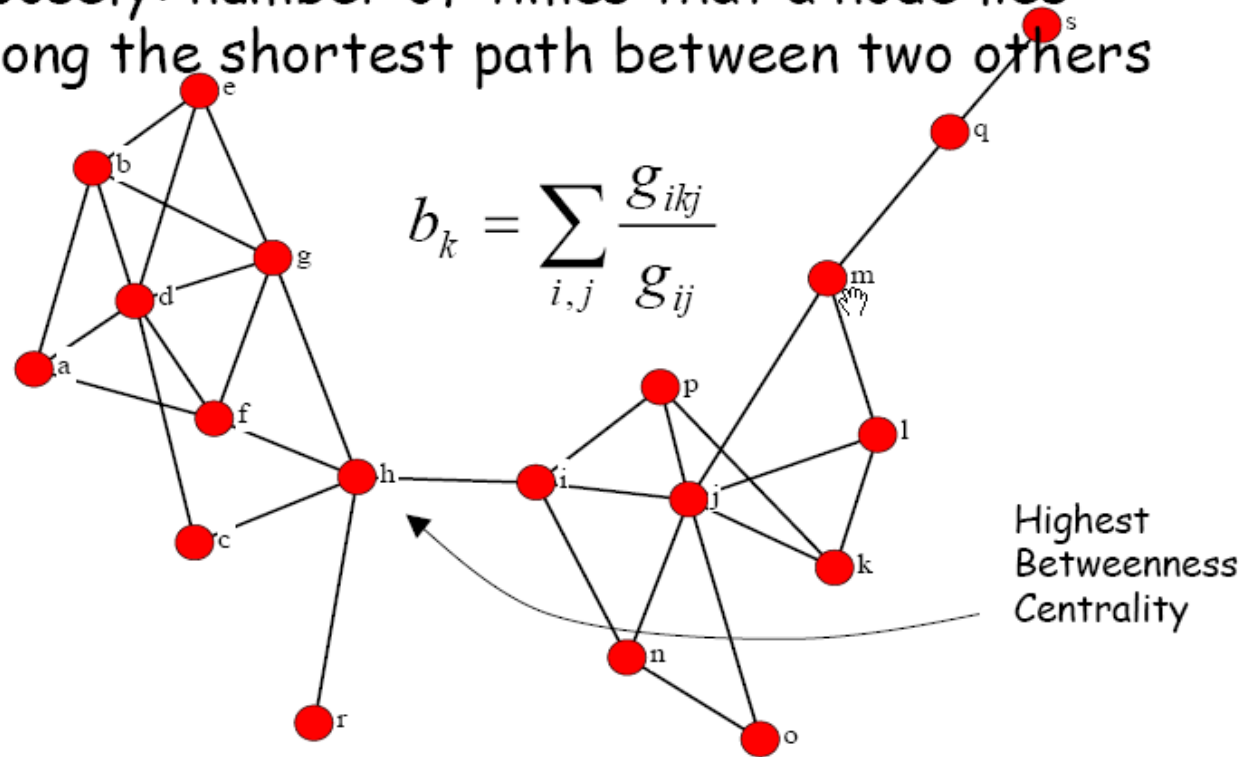
- Sum of geodesic distances to all other nodes
- Inverse measure of centrality



Centrality Measures (Example)

Betweenness Centrality

- Loosely: number of times that a node lies along the shortest path between two others

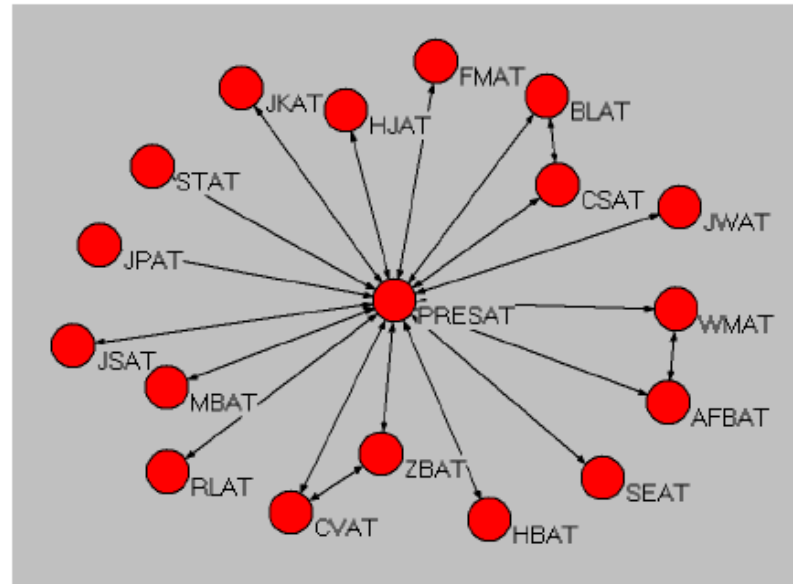


Centrality Measures (Example)



Centralization

- Degree to which network revolves around a single node



Carter admin.
Year 1

Centrality Measures

Centrality Vs centralization

Centrality Vs centralization

- › Centrality is a node level metric; each node has a centrality score representing its position within the network
- › Centralization is a group level metric indicating the extent to which the network is dominated by one or a few nodes

Clustering

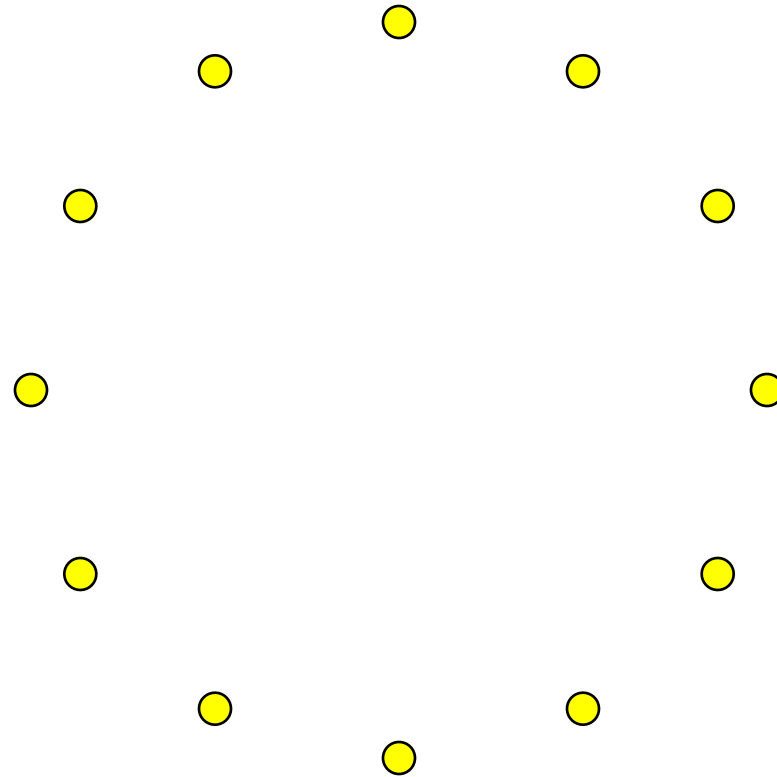
- › Many social networks contain subsets of nodes that are highly connected within the subset and have relatively few connections to nodes outside the subset .
- › The nodes in such subsets are likely to share some attributes and form their own communities.
- › Clustering coefficient: to measure the degrees of nodes to decide which nodes in a graph tend to be clustered together.

Modeling and Visualization

- › Visualizing Online Social Networks
- › A Taxonomy of Visualizations
- › Graph Representation
 - › Centrality- Clustering
- › Node-Edge Diagrams
- › Visualizing Social Networks with Matrix Based Representations
- › Node-Link Diagrams
- › Hybrid Representations
- › Modelling and aggregating social network data
- › Random Walks and their Applications
- › Use of Hadoop and Map Reduce
- › Ontological representation of social individuals and relationships

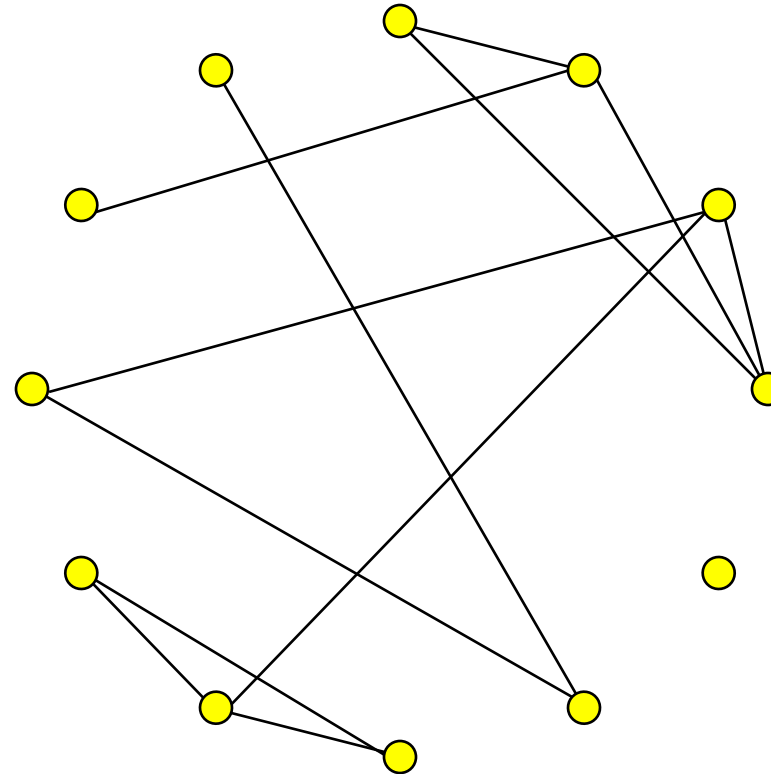
Society as a Graph

- › People are represented as nodes.



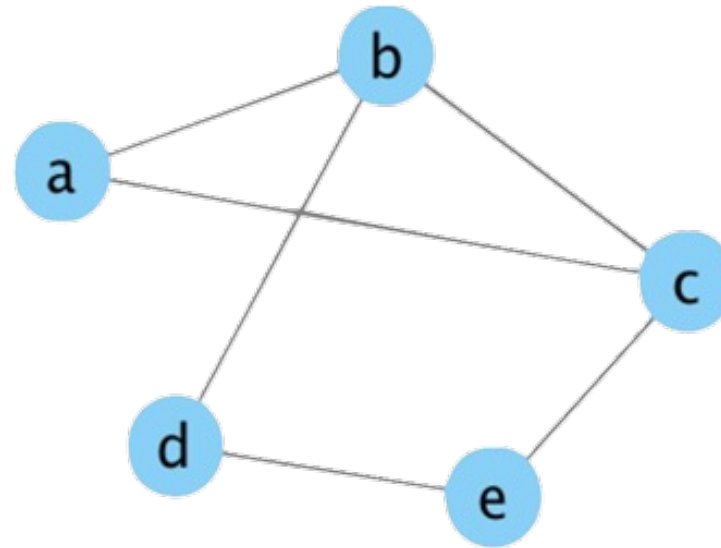
Society as a Graph

- › People are represented as nodes.
- › Relationships are represented as edges.
- › (Relationships may be acquaintanceship, friendship, co-authorship, etc.)



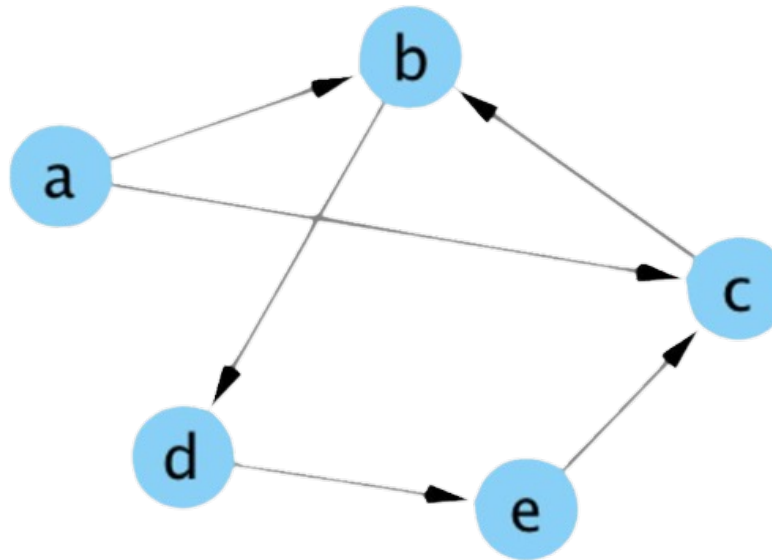
Society as a Graph

› Undirected Graph



Society as a Graph

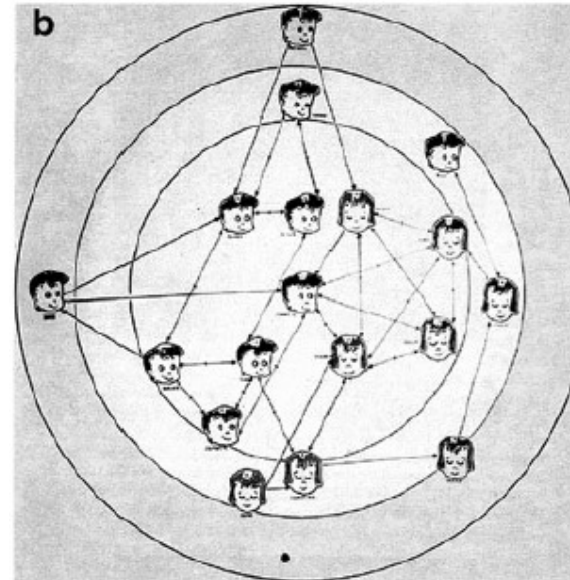
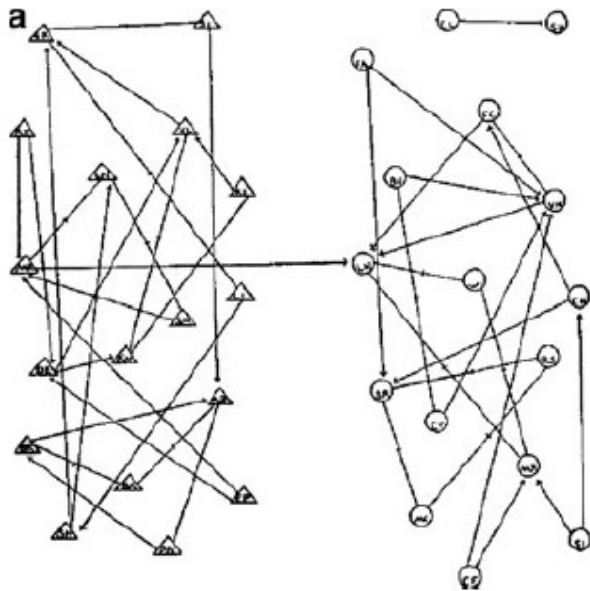
Directed Graph



Modeling and Visualization

- › Visualization of Social Network using Hand-drawn Images
- › Friendship choices among
- › fourth grade children

The Sociogram of a first grade class



Modeling and Visualization

Visualizing Online Social Networks

- › Social network visualizations try to reveal the structure of a network, its individuals and the relations between them
- › Some visual representations are
 - › node-edge diagrams
 - › matrix representations

Modeling and Visualization

Visualizing Online Social Networks

- › Visualization of online social networks can be categorized into three types by their social relationships:
 - › user-centric visualization
 - › content-centric visualization
 - › hybrid visualization.

Modeling and Visualization

1. User-centric visualization

- › Present various characteristics of actors and helps explore different subjects and relationships of interests.
- › For example, to discover individuals and communities that meet the following expectations:
 - › actors or groups with similar/complementary features
 - › key actors or those with high social impacts
 - › actors with popular interpersonal relationships or active social interactions.
- › user-centric visualizations are widely utilized to help people access their social networks and discover the social networks of their interests.

Modeling and Visualization

2.Content-centric visualization

- › Various kinds of contents can be properly presented to facilitate people analyzing social networks.
- › For example, at least the following three sorts of social network contents can displayed with content-centric visualization:
 - › the distribution of user opinions, including key opinions and controversial comments
 - › user opinions with high impacts toward their social communities, especially the effective-impact period of time
 - › the relations among different content groups.

Modeling and Visualization

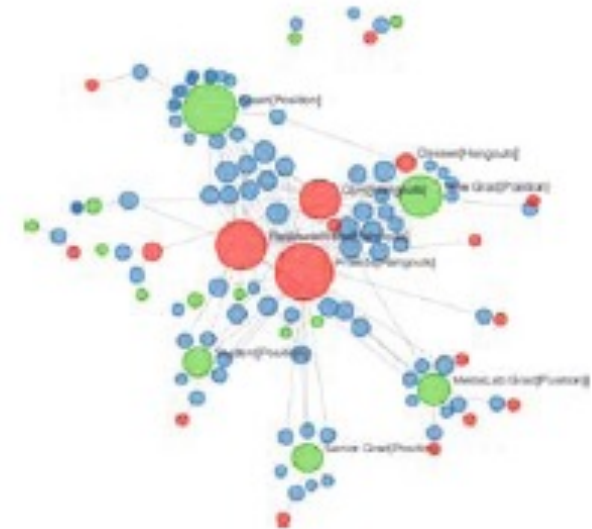
3. Hybrid visualization

- › Hybrid visualization is to visualize social networks from different aspects of attributes, e.g. people and contents.
- › Particularly, online social activities, such as email and dating services, usually include such elements.

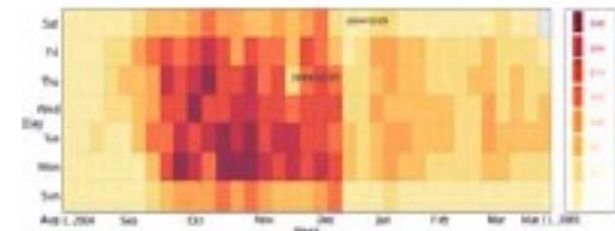
Social Network Analysis SNA

Dr. Mohamad NAGI

- 35



(b) Semantic



(d) Temporal

Modeling and Visualization

- › Visualizations of online social networks were developed according to the attributes of network sociality to present their network structure.
 - › Web communities
 - › Email groups
 - › Digital libraries
 - › Web 2.0 services

Modeling and Visualization

1. Web communities

- › Different social network services were created on the Web to help people maintain their social relationships.

Modeling and Visualization

2.Email groups

- › For analyzing the social structures of the daily email activities, visualization techniques are employed to explore different patterns.
- › Examples
 - › Soylent visualization
 - › Social Network Fragments (SNF) and PostHistory
 - › Themail

Modeling and Visualization

3.Digital libraries

- › In digital libraries, social networks can be mainly analyzed from two aspects: authors and writings.
- › Co-Authorship Networks
 - › Some characteristics, such as clustering coefficient and average path length, can be analyzed
- › Co-Citation Relations
 - › co-citation social networks can be formed through the continuously accumulated publications.
 - › With proper visualization of co-citation networks, documents with high impacts or similar citation patterns can be immediately identified

Modeling and Visualization

4. Web 2.0 services

- › Many Web 2.0 applications are popularly accessed by users to connect their social networks, such as Twitter and Facebook.
- › For example, Twitter provides users convenient functionalities to share the up-to-date status with their followers.
- › Nexus4 is a visualization application on Facebook communities to illustrate their large network graphs.



Association Rules Mining

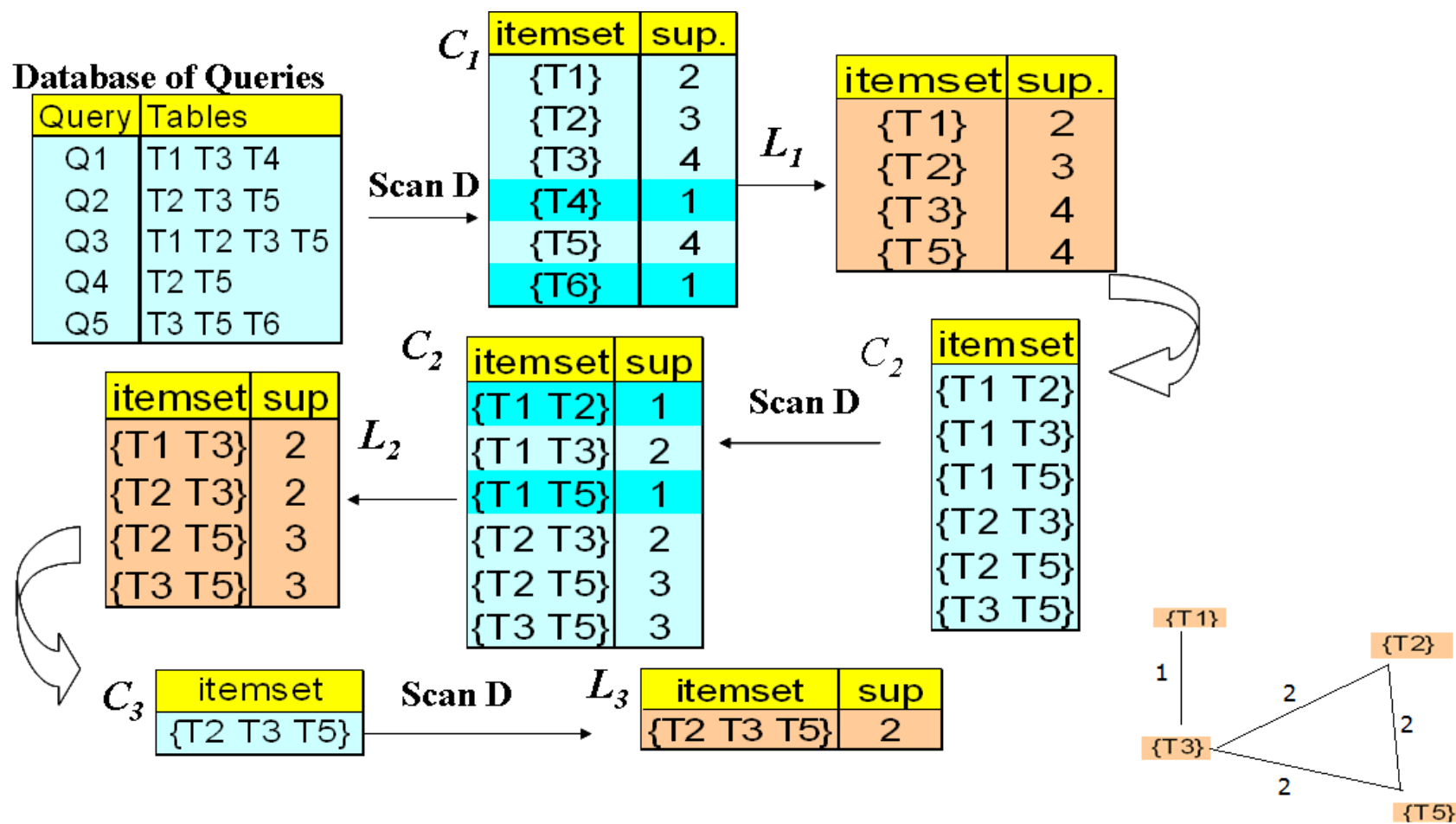
- › Proposed by Agrawal in 1993 to analyze market baskets and find in a supermarket items that are mostly sold together
- › It is a general model that fits any domain where
 - › There are two sets of entities, one corresponds to transactions and one corresponds to items
 - › The entities are connected by many-to-many relationship, they form like a bipartite graph
- › Example
 - › documents versus words
 - › Users versus web pages
 - › Web pages linked to web pages
 - › Drugs versus disease
 - › Drugs versus proteins

Association Rules Mining

Two Steps

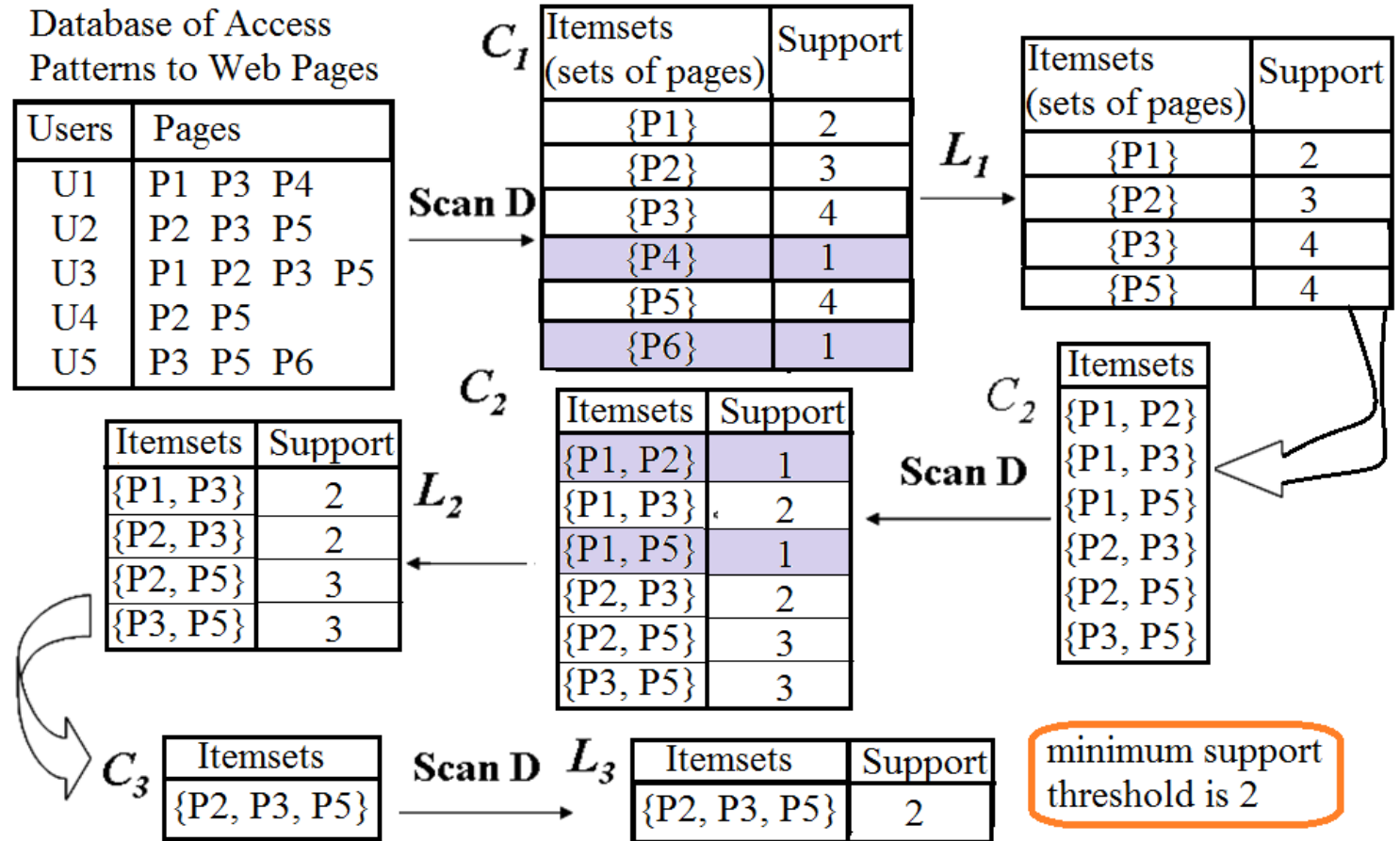
- › In general association rule mining can be reduced to the following two steps:
 - › Find all frequent itemsets
 - › Each itemset will occur at least as frequently as a minimum support count
 - › Generate strong association rules from the frequent itemsets
 - › These rules will satisfy minimum support and confidence measures
- › We only use the first step in our algorithm

Association Rules Mining



From FPM to Network Construction

- › Find frequent itemsets
 - › set of items that appear together in at least k transactions
 - › Any subset of a frequent itemset must be frequent
 - › Apriori pruning principle: If there is any itemset which is infrequent, its superset should not be generated/tested!



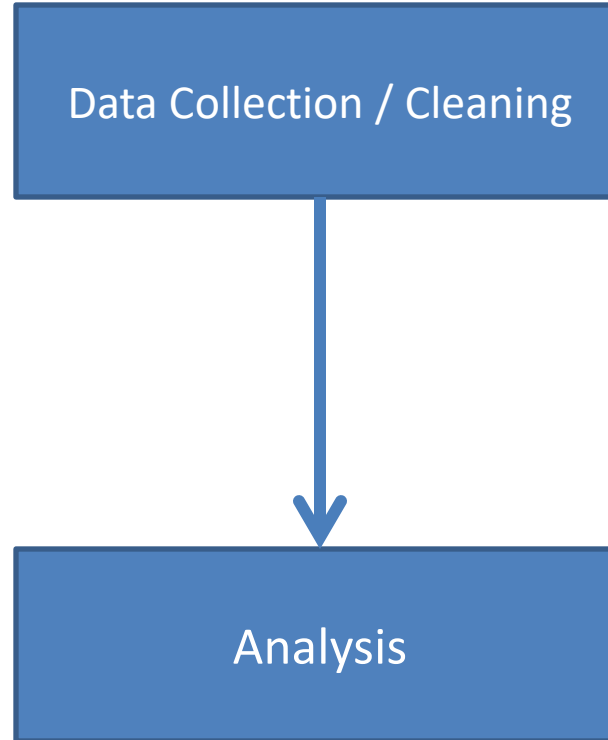


Social Network Analysis Common Tasks

- › Measuring “importance”
 - › Centrality, prestige
- › Link prediction
- › Diffusion modeling
 - › Epidemiological
- › Clustering
 - › Blockmodeling, Girvan-Newman
- › Structure analysis
 - › Motifs, Isomorphisms, etc.
- › Visualization/Privacy/etc.

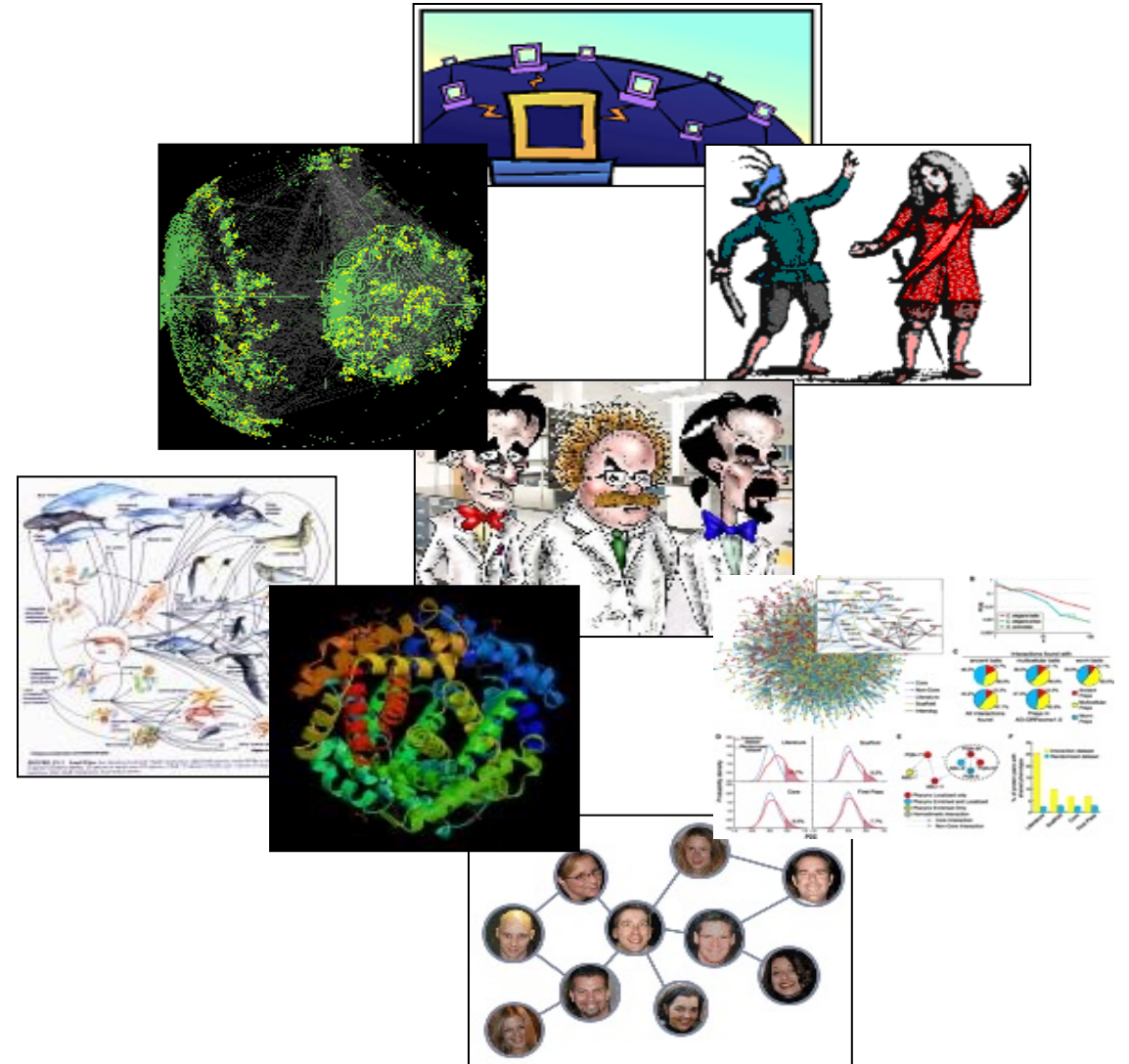


Social Network Analysis Common Tasks



Applications of Network Theory

- › World Wide Web and hyperlink structure
- › The Internet and router connectivity
- › Collaborations among...
 - › Movie actors
 - › Scientists and mathematicians
- › Cellular networks in biology
- › Food webs in ecology
- › Phone call patterns
- › Word co-occurrence in text
- › Neural network connectivity of flatworms
- › Conformational states in protein folding



Web Applications of Social Networks

- › Analyzing page importance
 - › Authorities/Hubs
 - › Page Rank
 - › Related to recursive in-degree computation
- › Discovering Communities
 - › Finding near-cliques
- › Analyzing Trust
 - › Propagating Trust
 - › Using propagated trust to fight spam
 - › In Email
 - › In Web page ranking

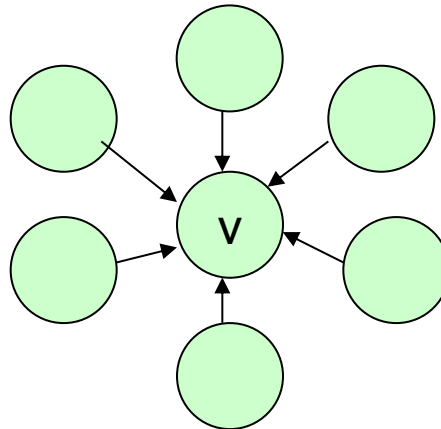
Hub & Authorities

- › HITS is a ranking algorithm which ranks “hubs” and “authorities”.

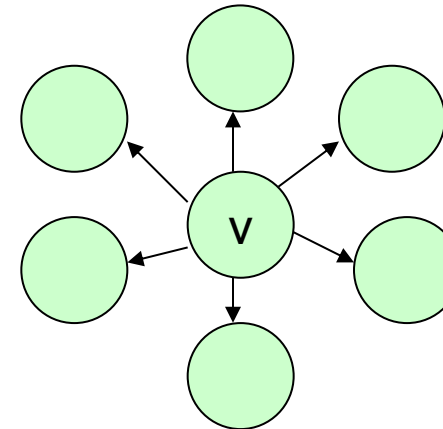
Each page has
two weights

1. Authority weight $a(v)$
2. Hub weight $h(v)$

■ Authority



■ Hub



Hub & Authorities

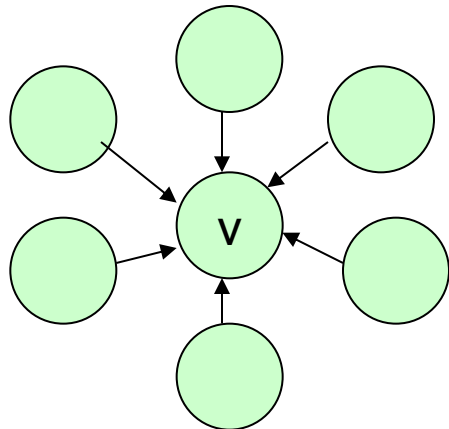
- › Each vertex has two weights
 - › Authority weight
 - › Hub weight

A good hub has many outgoing edges to good authorities

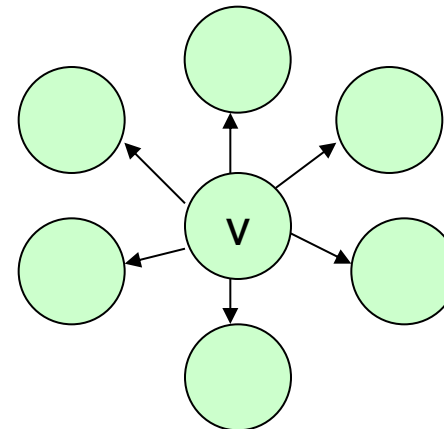
A good authority has many edges from good hubs

Authority Weight

Hub Weight



$$a(v) = \sum_{u \rightarrow v} h(u)$$



$$h(v) = \sum_{v \rightarrow u} a(u)$$



Hub & Authorities

- › HITS involves two major steps.
 - › Step 1: Sampling Step
 - › Step 2: Iteration Step



Hub & Authorities

Step 1 – Sampling Step

- › Given a user query with several terms
 - › Collect a set of pages that are very relevant – called the base set
- › How to find base set?
 - › We retrieve all webpages that contain the query terms. The set of webpages is called the root set.
 - › Next, find the link pages, which are either pages with a hyperlink to some page in the root set or some page in the root set has hyperlink to these pages
 - › All pages found form the base set.



Hub & Authorities

Step 2 – Iteration Step

$$h(N) = a(N) + a(MS) + a(A)$$

$$h(MS) = a(A)$$

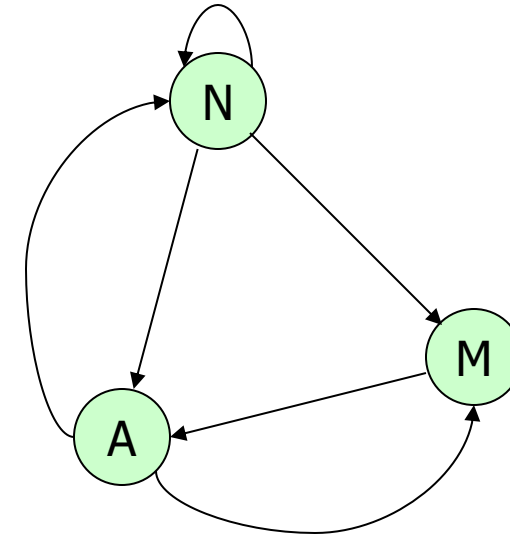
$$h(A) = a(N) + a(MS)$$

$$\begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix}$$

$$\vec{h} = M\vec{a}$$

N: Netscape
MS: Microsoft
A: Amazon.com

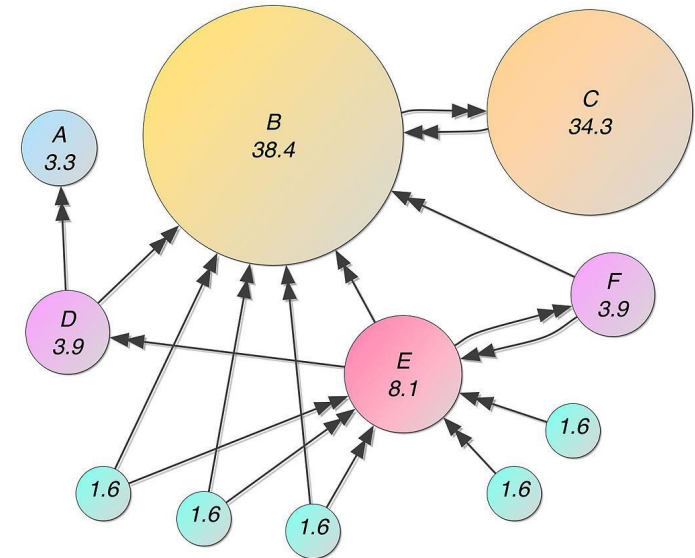
$$= \begin{matrix} & \begin{matrix} N & MS & A \end{matrix} \\ \begin{matrix} N \\ MS \\ A \end{matrix} & \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \end{matrix}$$



$$\vec{h} = \begin{pmatrix} h(N) \\ h(MS) \\ h(A) \end{pmatrix} \quad \vec{a} = \begin{pmatrix} a(N) \\ a(MS) \\ a(A) \end{pmatrix}$$

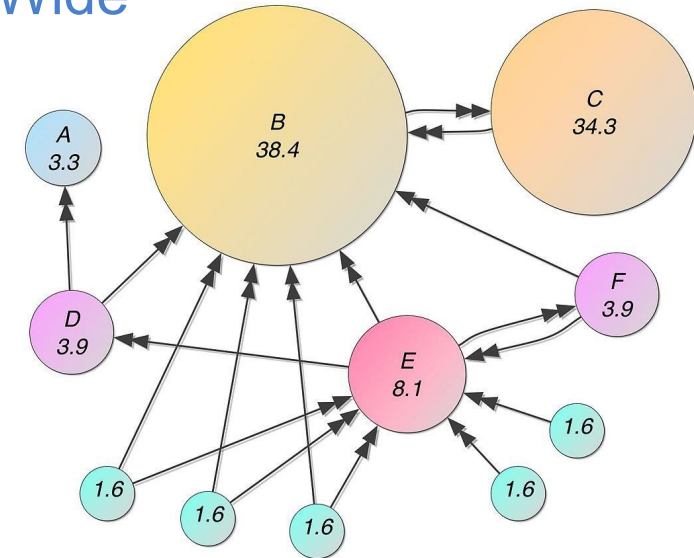
Web Applications of Social Networks

- › PageRank (PR) is an algorithm used by Google Search to rank web pages in their search engine results. PageRank is a way of measuring the importance of website pages.
- › According to Google: PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.



Page Rank

- › Why is Page Importance Rating important?
 - › New challenges for information retrieval on the World Wide Web.
 - › Huge number of web pages:
 - › 150 million by 1998
 - › 1000 billion by 2008
 - › Diversity of web pages:
 - › different topics, different quality, etc.
- › What is PageRank?
 - › A method for rating the importance of web pages objectively and mechanically using the link structure of the web.

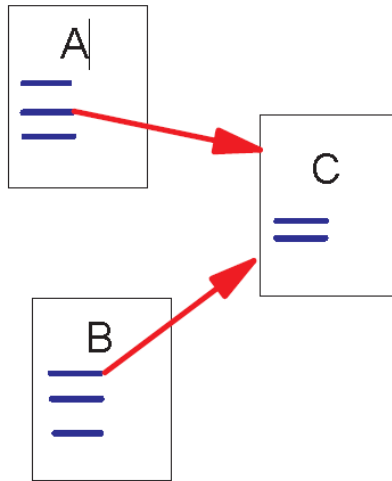


Page Rank

- › PageRank was developed by Larry Page (hence the name Page-Rank) and Sergey Brin.
- › It is first as part of a research project about a new kind of search engine. That project started in 1995 and led to a functional prototype in 1998.

Page Rank

› 150 million web pages → 1.7 billion links



- › Backlinks and Forward links:
 - › A and B are C's backlinks
 - › C is A and B's forward link

Intuitively, a webpage is important if it has a lot of backlinks.

Page Rank

› A Simple Version of PageRank

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

- › u: a web page
- › B_u : the set of u's backlinks
- › N_v : the number of forward links of page v
- › c: the normalization factor to make $\|R\|_1 = 1$ ($\|R\|_1 = |R_1| + \dots + |R_n|$)

Page Rank

- › PageRank is a global ranking of all web pages based on their locations in the web graph structure
- › PageRank uses information which is external to the web pages – backlinks
- › Backlinks from important pages are more significant than backlinks from average pages
- › The structure of the web graph is very useful for information retrieval tasks.

PageRank vs HITS

- › PageRank advantage over HITS
 - › Query-time cost is low
 - › HITS: computes an eigenvector for every query
 - › Less susceptible to localized link-spam
- › HITS advantage over PageRank
 - › HITS ranking is sensitive to query
 - › HITS has notion of hubs and authorities
- › Topic-sensitive PageRanking [Haveliwala WWW11]
 - › Attempt to make PageRanking query sensitive

Cliques

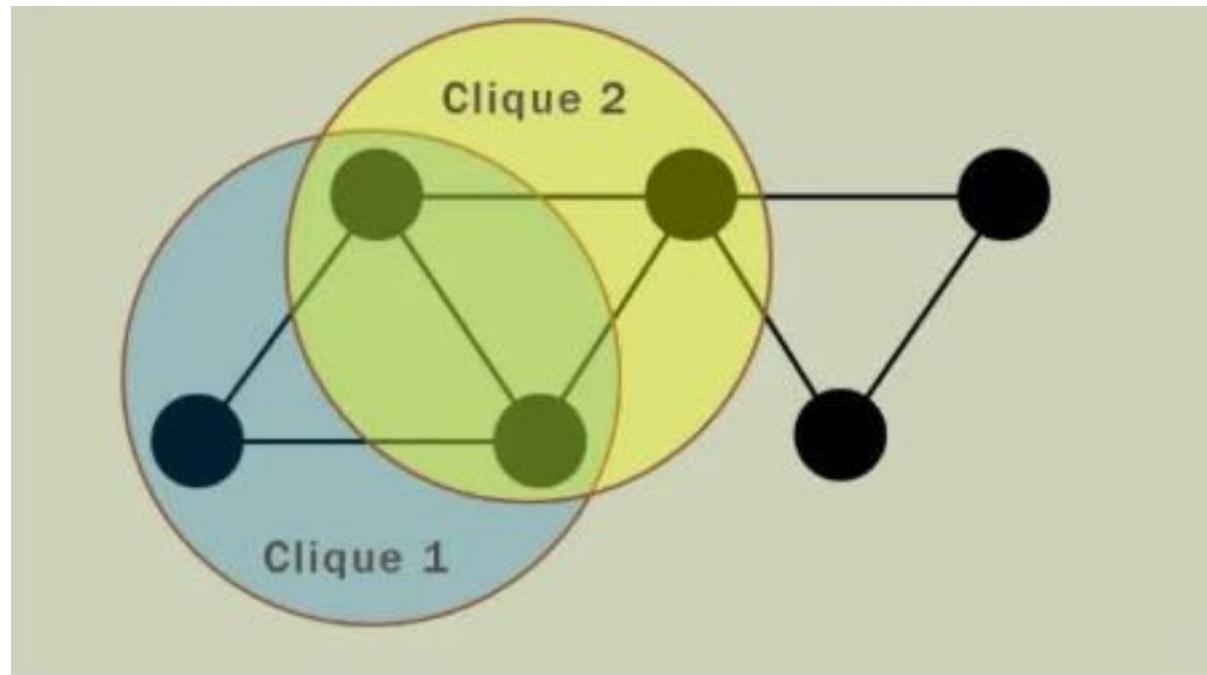
- › Direct Discovery of Community
- › Cliques : a small group of people with the same interests who do not want others to join their group.
- › Cliques in graph : A clique is a collection of vertices in an undirected graph G such that every two different vertices in the clique are nearby, implying that the induced subgraph is complete.

Cliques

- › Direct Discovery of Community using Clique Percolation Method (CPM)
 - › The clique percolation method builds up the communities from k -cliques, which correspond to complete (fully connected) sub-graphs of k nodes. (E.g., a k -clique at $k=3$ is equivalent to a triangle).
 - › Two k -cliques are considered adjacent if they share $k-1$ nodes.
 - › A community is defined as the maximal union of k -cliques that can be reached from each other through a series of adjacent k -cliques

Cliques

- › Direct Discovery of Community using Clique Percolation Method (CPM)





Social Network Analysis Connections

- › 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

Social Network Analysis 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

SEO

- › SEO (Search Engine Optimisation) involves altering website code, content, and presence in order to increase rankings in search engines
- › At least half a billion searches are made worldwide on a daily basis
 - › If you are ranked at the top of a search engine, you are essentially running a free marketing campaign

SEO

- › How a search engine works?
- › Search engines use programs called spiders, or bots, to search the internet and document their findings.
- › Spiders firstly go to a webpage and analyse it
 - › Goes through the HTML code and categorises the most important words (keywords) in the document.
- › It then moves on to any pages that linked from the current page.
 - › This is one way how search engines find new pages.
 - › It also visits pages that may not be linked by other pages, but were submitted by search engines.

SEO

Search Engine Optimization:

- › SEO consists of activities / techniques applied towards making a website rank at the top in search engine.
- › The prime aim of SEO exercise for a marketer is to help generate awareness for its brand website by gaining attention through content relevance and keyword prominence rather than paying for it through SEM campaigns.
- › The key aim on any search engine is to help users find a specific piece of information from the billion pieces of content lying on the internet.
- › Search Engines execute following major functions which form the core of their businesses.



SEO

Crawling:

- › The purpose of the crawl is to identify relevant pages for indexing and assess whether they have changed.
- › Crawling is performed by robots (bots) which are also known as spiders. These access web pages and retrieve a reference URL of the page for later analysis.

SEO

Indexing:

- › An index is created to enable the search engine to rapidly find the most relevant pages containing the query typed by the searcher.
- › Rather than searching each page for a query phrase, a search engine 'inverts' the index to produce a lookup table of the documents containing particular words.

SEO

Ranking & Scoring:

- › The indexing process has produced a lookup of all the pages that contain particular words in a query, but they are not sorted in terms of relevance.
- › Ranking of the document to assess the most relevant set of documents to return in the Search engine result pages (SERPs) occurs in real time for the search query entered.
- › First, relevant documents will be retrieved from a runtime version of the index at a particular data center, then a rank in the SERPs for each.

> SEO

Query request and Results Serving:

- › The familiar search engine interface accepts the searcher's query.
- › The user's location is assessed through their IP address and the query is then passed to a relevant data center for processing.
- › Ranking then occurs in real time for a particular query to return a sorted list of relevant documents, and these are then displayed on the Search Results Page.

SEO

Improving Search Engine Ranking Through SEO

- › The challenge for SEOs is that there are many ranking factors and they are constantly changing!
- › Google references over 200 main ranking factors which mean hundreds of detailed changes to the way pages are evaluated and represented in the search results places happening each year.
- › Fortunately there are 4 common factors that influence search engine rankings, that can be managed as part of an SEO strategy:

SEO

1. High-Quality, Relevant Content

- › Content is the most critical factor, often referred to as "Helpful Content." Google prioritizes content that fulfills user intent, offers unique insights, and demonstrates authority.
 - › How to Manage: Develop topical authority rather than just targeting keywords. Create comprehensive guides that answer all potential questions a user might have on a topic, and keep it updated.
 - › Focus on E-E-A-T: Ensure content demonstrates Experience, Expertise, Authoritativeness, and Trustworthiness.

SEO

2. Authoritative Backlinks (Off-Page SEO)

- › Backlinks act as trust signals or "votes of confidence" from other websites. The quality and relevance of the linking domain are far more important than the sheer quantity of links.
 - › How to Manage: Focus on earning natural links through high-quality content, digital PR, or guest posting on reputable, topically relevant sites.
 - › Avoid Spam: Steer clear of link exchanges or buying links, as these can lead to penalties.

> SEO

3. Technical SEO (Site Performance)

- › Technical SEO ensures that search engines can easily crawl, index, and understand your website. This includes mobile-friendliness, site speed, and secure browsing (HTTPS).
 - › How to Manage: Improve page loading speed by compressing images and reducing unnecessary code. Use mobile-first design, ensuring your site is fully functional on smartphones.
 - › Structure: Use a clear XML sitemap and a logical site structure to assist crawlers.

> SEO

4. User Experience (UX & Core Web Vitals)

- › Search engines favor websites that provide a seamless, enjoyable experience. Google uses Core Web Vitals to measure page loading performance, interactivity, and visual stability.
 - › How to Manage: Optimize for Core Web Vitals (LCP, INP, CLS) to ensure pages load quickly and are stable while loading.
 - › Engagement: Improve usability with intuitive navigation, engaging content, and low bounce rates.

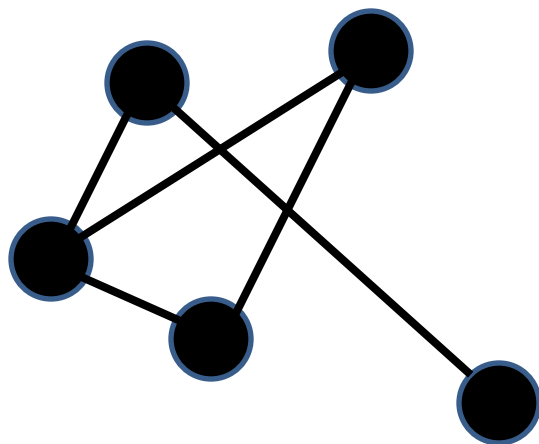
SEO

Important SEO Tips You Need to Know:

- › Remove anything that slows down your site.
- › Link to other websites with relevant content.
- › Write for humans first, search engines second
- › Encourage other trustworthy sites to link to you.
- › Write unique and relevant Meta descriptions for every page.
- › Use readable and meaningful URLs only.
- › Use the right keywords in your images.
- › Create & publish unique content consistently to improve your rankings.
- › Build momentum with social signals.

Link Prediction Introduction

› Natural examples of social networks:



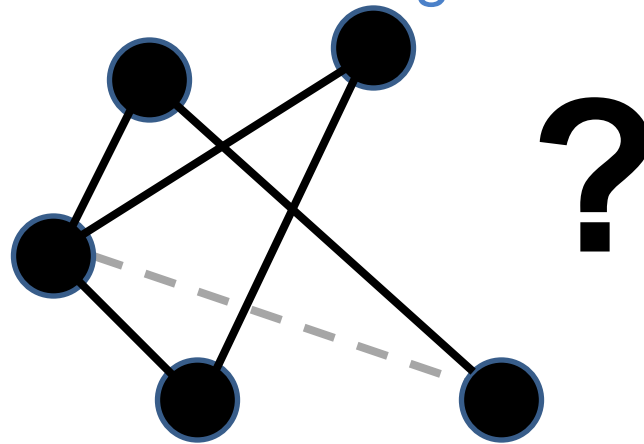
Nodes = people/entities

Edges = interaction/ collaboration

Nodes	Edges
Scientists in a discipline	Co-authors of a paper
Employees in a large company	Working on a project
Business Leaders	Serve together on a board

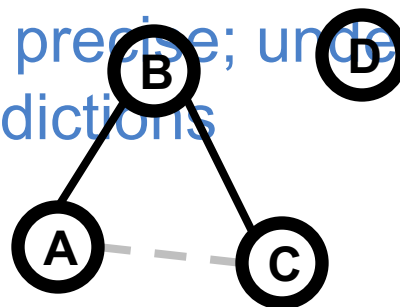
Link Prediction Motivation

- › Understanding how social networks evolve
- › The link prediction problem
- › Given a snapshot of a social network at time t , we seek to accurately predict the edges that will be added to the network during the interval (t, t')



Link Prediction Motivation

- › Co-authorship network for scientists
- › Scientists who are “close” in the network will have common colleagues & circles – likely to collaborate
- › Caveat: Scientists who have never collaborated might in future - hard to predict
- › Goal: make that intuitive notion precise; understand which measures of “proximity” lead to accurate predictions



Link Prediction Why ?

- › To suggest interactions or collaborations that haven't yet been utilized within an organization
- › To monitor terrorist networks - to deduce possible interaction between terrorists (without direct evidence)
- › Used in Facebook and Linked In to suggest friends

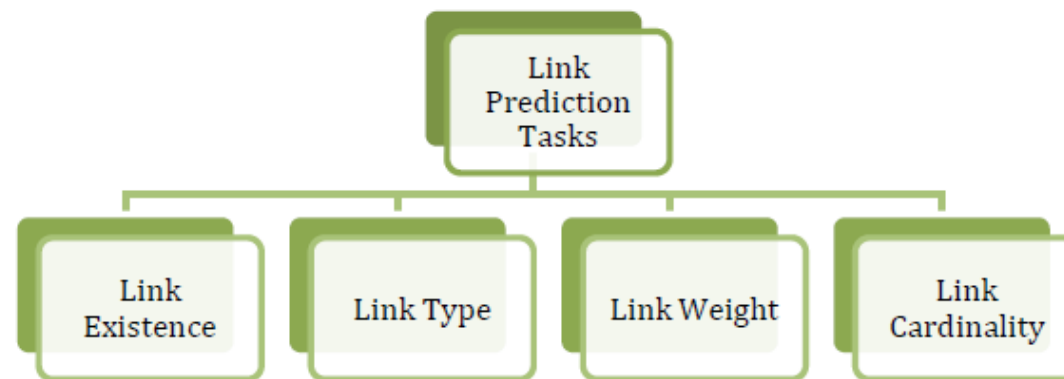
Link Prediction Problem

- › The “Link Prediction Problem”
- › Given a snapshot of a social network, can we infer which new interactions among its members are likely to occur in the near future?
- › Based on “proximity” of nodes in a network

Link Prediction Variation

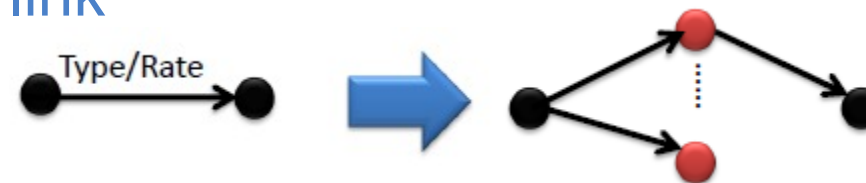
- › Link completion
 - › Example: When a user buys five books online and the name of one book is corrupted in transfer. A link completion algorithm could infer the name of the missing book based on the user's name and the other books she bought
 - › Alice, Bob, and a third person attended a meeting. Given people's previous co-occurrences, who is the third person?
- › Link discovery

Link Prediction Definition



› Generalized Definition

- › Existence (i.e., does a link exist?)
- › Classification (i.e., what type of the relationship?)
- › Regression (i.e., how does the user rate the item?)
- › Cardinality (i.e., is there more than one link between same pair of nodes?)



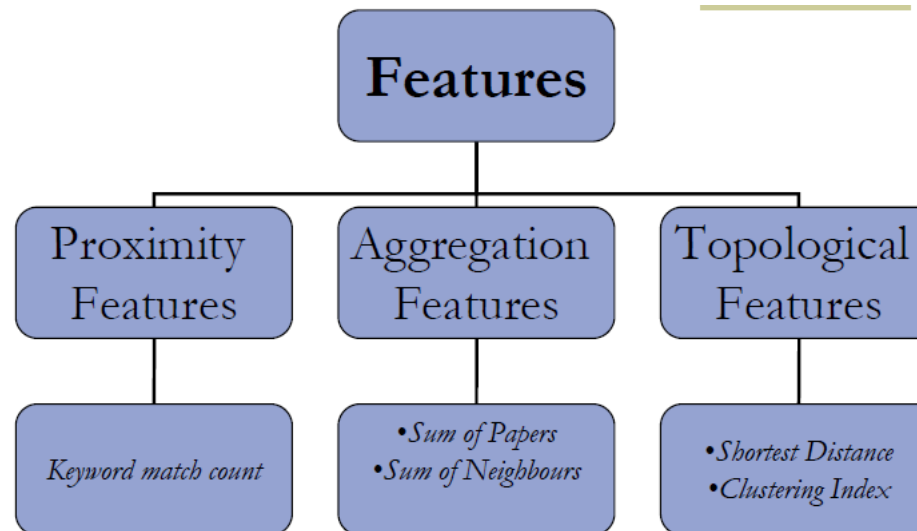
Link Prediction Challenges

- › Data sparse
 - › Missing data/relationship
- › Imbalance
 - › So many possibilities, so few choices
 - › Referred as inaccurate problem due to low accuracy in practice
- › Accuracy vs. Scalability
 - › Modeling (unobserved/unknown factors)
 - › Tasks of approximation/optimization

Link Prediction Feature based

- › model the link prediction problem as a supervised classification task
- › each data point corresponds to a pair of vertices in the social network graph
- › To train the learning model, we can use the link information from the training interval $[t_0, t'_0]$.
- › From this model, predictions of future links in the test interval $[t_1, t'_1]$ can be made
- › Any of the popular supervised classification tools, such as naive Bayes, neural networks, SVM and k nearest neighbors, can be used.

Link Prediction Supervised Learning



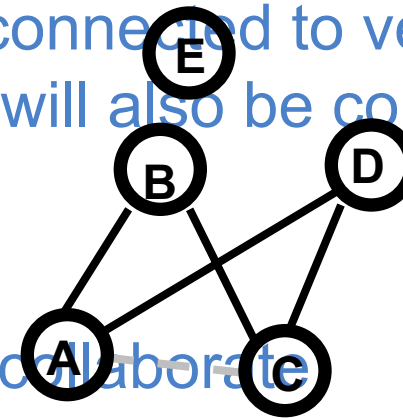


Link Prediction

Node Neighborhood based Features

Common Neighbours

- if vertex x is connected to vertex z and vertex y is connected to vertex z , then there is a heightened probability that vertex x will also be connected to vertex y .



- A and C have 2 common neighbors $|\Gamma(x) \cap \Gamma(y)|$ are likely to collaborate
- As the number of common neighbors grows higher, the chance that x and y will have a link between them increases.



Link Prediction

Node Neighborhood based Features

› Jaccard Coefficient

- › normalizes the size of common neighbors as below

$$\frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

- › it defines the probability that a common neighbor of a pair of vertices x and y would be selected if the selection is made randomly from the union of the neighbor-sets of x and y .
- › For high number of common neighbors, the score would be higher.

Link Prediction

Node Neighborhood based Features

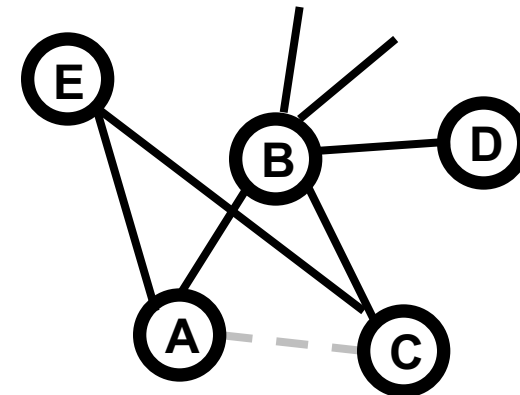
› Adamic / Adar

- › a metric of similarity between two web pages.
- › For a set of features z , it is defined as below.

$$\sum_{z : \text{feature shared by } x, y} \frac{1}{\log(\text{frequency}(z))}$$

- › For link prediction: weighting rarer neighbors more heavily

$$\sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|}$$

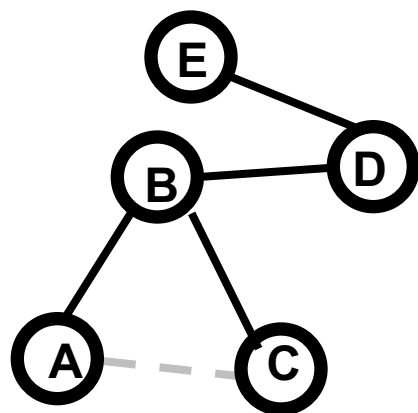


Link Prediction

Path based Features

› Shortest Path Distance

- › (Negated) length of shortest path between x and y



(A, C)	-2
(C, D)	-2
(A, E)	-3

- › The shorter the distance, the higher the formation of a link between them

Link Prediction

Path based Features

- › Hitting time: expected number of steps for a random walk starting at x to reach y
$$-(H_{x,y} + H_{y,x})$$
 - › Shorter hitting time denotes that the nodes are similar to each other, so they have a higher chance of linking in the future.
- › Commute time:
 - › it is easy to compute by performing some trial random walks.
 - › On the downside, its value can have high variance; hence, prediction by this feature can be poor
- › Normalized-hitting-time(x,y): If y has a large stationary probability, $H_{x,y}$ is small. To counterbalance, we can normalize

$$\text{score}(x, y) := -(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$$

Link Prediction

Path based Features

- › PageRank: to cut down on long random walks, walk can return to x with a probability α at every step y
- › Rooted PageRank
 - › similarity score between two vertices x and y can be measured as the stationary probability of y in a random walk that returns to x with probability $1 - \beta$ in each step, moving to a random neighbor with probability β .
 - › The rooted pagerank between all node pairs (represented as RPR) can be derived as follows.
 - › Let D be a diagonal degree matrix with $D[i, i] = \sum_j A[i, j]$
 - › Let, $N = D^{-1}A$ be the adjacency matrix with row sums normalized to 1. Then,

$$RPR = (1 - \beta)(I - \beta N)^{-1}$$



Online Social Networks (OSN)

- › Online Web services connecting people to socialize
 - › Find friends, make friends
 - › Follow celebrities and companies
 - › Join groups
 - › Share photos, videos
 - › Communicate via messages, live chats
 - › Etc.



Social networks vs social networking

- › Social networks are links of actors and their relationships usually represented as a graph or network
- › Social networking is the actual implementation of social networks in the digital world or media
 - › A social network service focuses on building and reflecting of social networks or social relations among people, e.g., who share interests and/or activities. A social network service essentially consists of a representation of each user (often a profile), his/her social links, and a variety of additional services. Most social network services are web based and provide means for users to interact over the internet, such as e-mail and instant messaging.



Facebook vs Google

Great Wall of Facebook: The Social Network's Plan to Dominate the Internet — and Keep Google Out

By Fred Vogelstein  06.22.09



Web 2.0

- › A perceived second generation of web development and design, that aims to facilitate communication, secure information sharing, interoperability, and collaboration on the World Wide Web.
- › Web 2.0 concepts have led to the development and evolution of web-based communities, hosted services, and applications such as social-networking sites, video-sharing sites, wikis, blogs, and folksonomies.



Social Media

- › Information content created by people using highly accessible and scalable publishing technologies that is intended to facilitate communications, influence and interaction with peers and with public audiences, typically via the Internet and mobile communications networks.
- › The term most often refers to activities that integrate technology, telecommunications and social interaction, and the construction of words, pictures, videos and audio.
- › Businesses also refer to social media as user-generated content (UGC) or consumer-generated media (CGM).

Social Media on Web 2.0

- › Multimedia
 - › Photo-sharing: Flickr
 - › Video-sharing: YouTube
 - › Audio-sharing: imeem
- › Entertainment
 - › Virtual Worlds: Second Life
 - › Online Gaming: World of Warcraft
- › News/Opinion
 - › Social news: Digg, Reddit
 - › Reviews: Yelp, epinions
- › Communication
 - › Microblogs: Twitter, Pownce
 - › Events: Evite
- › Social Networking Services:
 - › Facebook, LinkedIn, MySpace





Top 10 Websites - 2009

Rank	Property	Unique Visitors (000) ✕
	Total Internet : Total Audience	195,538
1	Google Sites	158,820
2	Yahoo! Sites	156,506
3	Microsoft Sites	129,339
4	AOL LLC	104,759
5	FACEBOOK.COM	87,748
6	Fox Interactive Media	80,895
7	Ask Network	79,326
8	eBay	73,089
9	Amazon Sites	66,545
10	Wikimedia Foundation Sites	62,075



Top 10 websites - 2026

Position	Website	Change	Visits	Pages / Visit	Bounce rate
1	 google.com	–	88,460,065,425	5.53	37.02%
2	 youtube.com	–	45,068,713,468	9.43	30.28%
3	 facebook.com	–	8,223,710,674	8.0	45.86%
4	 instagram.com	–	5,320,502,486	9.23	49.06%
5	 chatgpt.com	–	5,194,915,553	3.96	37.88%
6	 reddit.com	–	4,451,904,136	3.51	59.09%
7	 wikipedia.org	–	3,772,950,164	3.26	58.36%
8	 whatsapp.com	–	3,342,914,099	7.59	23.88%
9	 x.com	–	3,284,472,016	9.45	46.25%
10	 whatsapp.com	–	2,486,282,964	5.15	57.33%



Top 10 Social Media - 2012

Top 10 Social Media Websites

Week ending June 9, 2012

Source: Experian Hitwise US

Websites	Total Visits	Visits Share ▼	Rank 06/02	Rank 05/26	Rank 05/19
Facebook	1,658,902,073	62.74%	1	1	1
YouTube	538,676,741	20.37%	2	2	2
Twitter	51,896,672	1.96%	3	3	3
Yahoo! Ans...	24,168,660	0.91%	4	4	4
Google+	22,837,815	0.86%	7	8	8
Pinterest	22,038,629	0.83%	5	5	5
Tagged	20,881,205	0.79%	6	7	6
Linkedin	20,684,282	0.78%	8	6	7
MySpace	11,178,414	0.42%	9	9	9
Yelp	9,304,977	0.35%	10	11	11

© 2012 Hitwise Pty Ltd. All rights reserved. [Disclaimer](#)  [Full Report](#)

Google+ ranks 5th position in Top 10 US social media websites (Source: Hitwise)



Top Social Media - 2026

Platform	Monthly active users (MAU)	Top content formats	Best for
Facebook	3.07B	Text, images, video, reels, stories, groups	General brand presence
Instagram	3B	Reels, stories, posts, lives	Visual storytelling
WhatsApp	3B	Messaging, voice notes, product catalogues	Customer support and updates
YouTube	2.6B	Long-form video, shorts, live	Video content and SEO
TikTok	1.9B	Short-form video, duets, lives	Short-form video and trends
WeChat	1.4B	Messaging, mini apps, payment	Chinese market reach
Telegram	1B	Channels, chats, bots	Broadcast and private messaging
Snapchat	932M	Snap, stories, AR filters, Spotlight	Direct engagement with Gen Z
Reddit	765M	Text, links, comments	Community insights and trends



Social Network Service

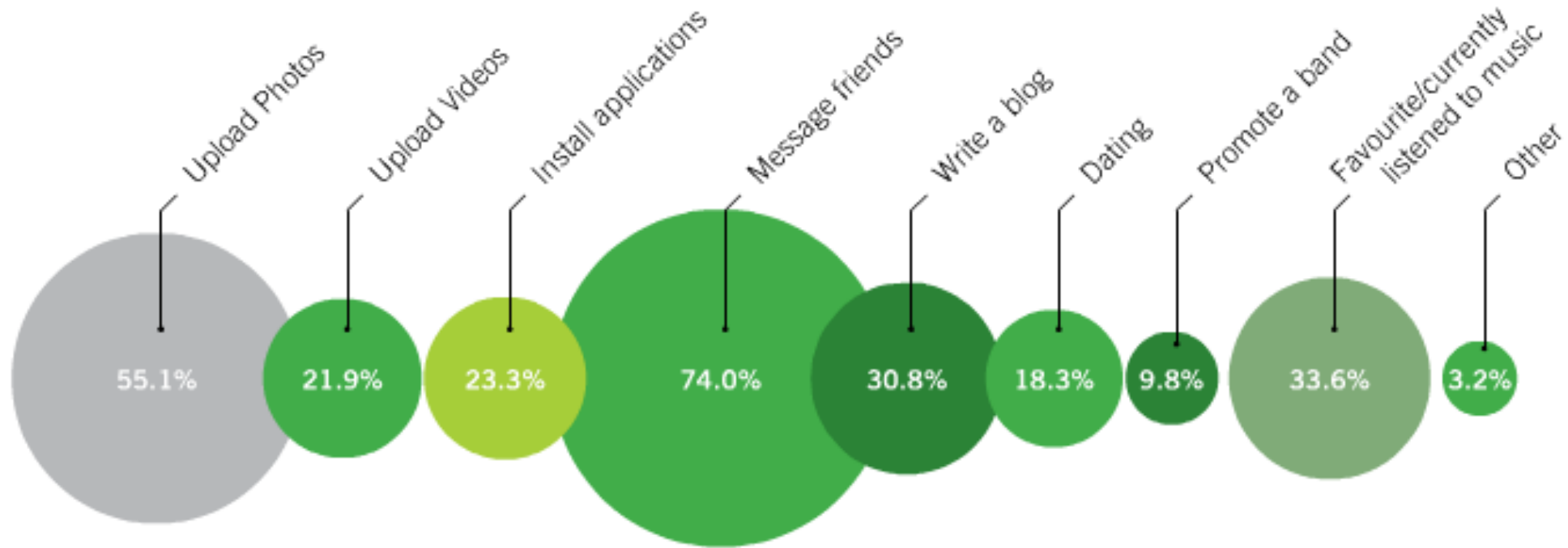
- › A social network service focuses on building online communities of people who share interests and/or activities, or who are interested in exploring the interests and activities of others.
- › Most social network services are web based and provide a variety of ways for users to interact, such as e-mail and instant messaging services.



Usage of Social Network

Content posted on social network

"What do you do with your social networking profile?" Active Internet Universe



% Added to Social Network Page (Social Network Users)



Usage of Social Network

- › Profile-based
- › Content-based
- › White-label
- › Multi-user Virtual environments
- › Mobile
- › Micro-blogging
- › Social Search

Profile-based

- › Primarily organized around members' profile pages – pages that mainly consist of information about an individual member, including the person's picture and details of interests, likes and dislikes.
- › [Bebo](#), [Facebook](#) and [MySpace](#) are all good examples of profile-based services.

Content-based

- › In these services, the user's profile remains an important way of organizing connections, but plays a secondary role to the posting of content.
- › Examples of content-based communities:
 - › YouTube.com for video sharing
 - › Last.fm, in which the content is arranged by software that monitors and represents the music that users listen to.
 - › Photo-sharing site like Flickr.com.

White-label

- › Offer some group-building functionality, which allows users to form mini-communities within sites.
- › Example sites:
 - › PeopleAggregator
 - › Ning
 - › Colexn
- › Seem to be losing popularity

Multi-user Virtual Environment

- › allow users to interact with each other's avatars
- › users have profile cards, their functional profiles are the characters they customize or build and control. Friends lists are usually private and not publicly shared or displayed.
- › Example Sites:
 - › Second Life
 - › World of Warcraft



Mobile Social Networking

- › Many social networking sites, for example [MySpace](#) and [Twitter](#), offer mobile phone versions of their services, allowing members to interact with their friends via their phones.
- › mobile-led and mobile-only communities include profiles and media-sharing just as with web-based social networking services.
- › [MYUBO](#), for example, allows users to share and view video over mobile networks
- › [MobiSNA](#), mobile video social networking

Micro-blogging

- › Allows you to publish short (140 characters, including spaces) messages publicly or within contact groups.
- › These services are designed to work as mobile services, but are popularly used on the web as well.
- › Example Services
 - › [Twitter](#)
 - › [Jaiku](#)

Social Search

- › Social search engines are an important web development which utilize the popularity of social networking services.
- › There are various kinds of social search engine, but sites like [Wink](#) and [Spokeo](#) generate results by searching across the public profiles of multiple social networking sites, allowing the creation of web-based dossiers on individuals.
- › This type of people search cuts across the traditional boundaries of social networking site membership, although any data retrieved should already be in the public domain.

Things you can do in a Social Network

- › Communicating with existing networks, making and developing friendships/contacts
- › Represent themselves online, create and develop an online presence
- › Viewing content/finding information
- › Creating and customizing profiles
- › Authoring and uploading content
- › Adding and sharing content
- › Posting messages – public & private
- › Collaborating with other people



Social Media Mining





Social Media Mining

What is data mining?

- › “Data mining is the efficient discovery of valuable, non-obvious information from a large collection of data.”

Joseph P. Bigus



Social Media Mining

Social Media

- › Social media is defined as a group of Internet-based applications that allow the creation and exchanges of user-generated content.
- › Social media gives users an easy-to-use way to communicate and network with each other on an unprecedented scale.
- › Facebook, the social networking site, recorded more than 845 million active users as of December 2011.



Social Media Mining

- › Social media mining is extracting information from social media.
- › Primary objectives of the data mining process are to effectively handle large-scale data, extract actionable patterns, and gain insightful knowledge.
- › Users on Twitter generate over 400 million Tweets everyday



Social Media Mining

- › Data mining of social media can expand researchers' capability of understanding new phenomena to provide better services and develop innovative opportunities.
- › Mining social media is a growing multidisciplinary area where researchers of different backgrounds can make important contributions that matter for social media research and development.



Social Media Mining

- › The reasons for growth of social media mining
- › Social media growth is driven by these:
 - › How can a user be heard?
 - › Which source of information should a user use?
 - › How can user experience be improved?

Answers to these questions are hidden in the social media data.



Social Media Mining

The amount of data!

- › For example, Facebook and Twitter report Web data from approximately 149 million and 90 million unique U.S. visitors per month, respectively.
- › According to the video sharing site YouTube, 5 more than 4 billion videos are viewed per day, and 60 hours of videos are uploaded every minute.
- › The picture sharing site Flickr, as of August 2011, hosts more than 6 billion photo images.
- › Web-based, collaborative, and multilingual Wikipedia hosts over 20 million articles attracting over 365 million readers.

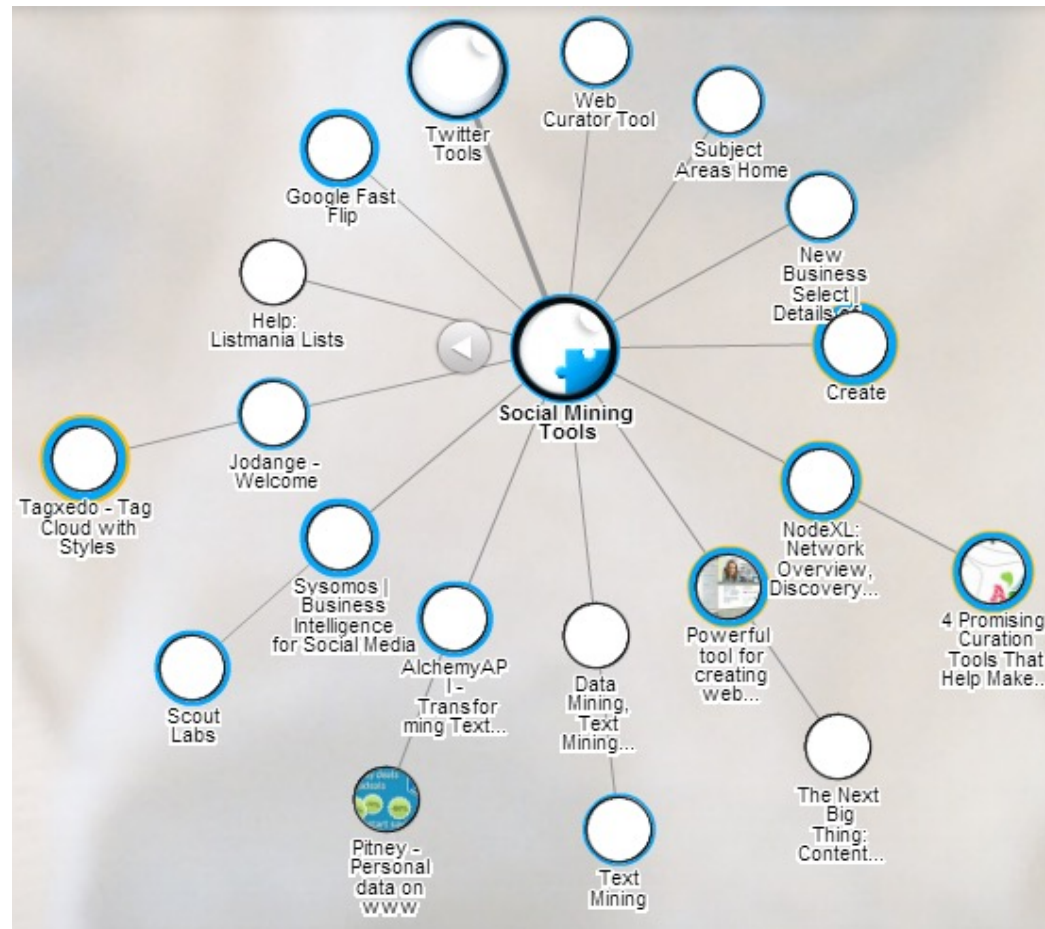


Social Media Mining

- › Challenges:
 - › Social media data are
 - › vast,
 - › noisy,
 - › distributed,
 - › unstructured,
 - › dynamic.
-
- › These characteristics pose challenges to data mining tasks to invent new efficient techniques and algorithms.



Social Media Mining





Social Media Mining

List of common tools:

- › Twitter tools
- › Cloud4Trends
- › TweetTracker



Social Media Mining

What is use of data mining in social media mining?

- › Social Media data is everywhere.
- › There is an overload.
- › Information Overload (blogs, photos, videos, bookmarks)
- › Interaction Overload (friends, taggers, followers, commenters)
- › How to extract data from this chaos?



Social Media Mining

What is use of data mining in social media mining?

- › Social media captures 'pulse of humanity'.
- › Can directly study opinions and behaviors of millions of users to gain insight into:
 - › Human behavior
 - › Market analytics
 - › Product sentiments



Social Media Mining

Applications of social media mining:

- › Personalization
- › Targeted marketing
- › Suggesting choices

- › Community Analysis
- › Sentiment Analysis and Opinion Mining
- › Social Recommendation
- › Influence Modeling



Lebanese University
Faculty of Information

Brunch I

Dr. Mohamad **Nagi**

PhD in Data Mining - Computer Science

Thank you for Listening