CentraleSupélec

DATA STREAM MANAGEMENT SYSTEM
SOCIAL NETWORK ANALYSIS AND MINING

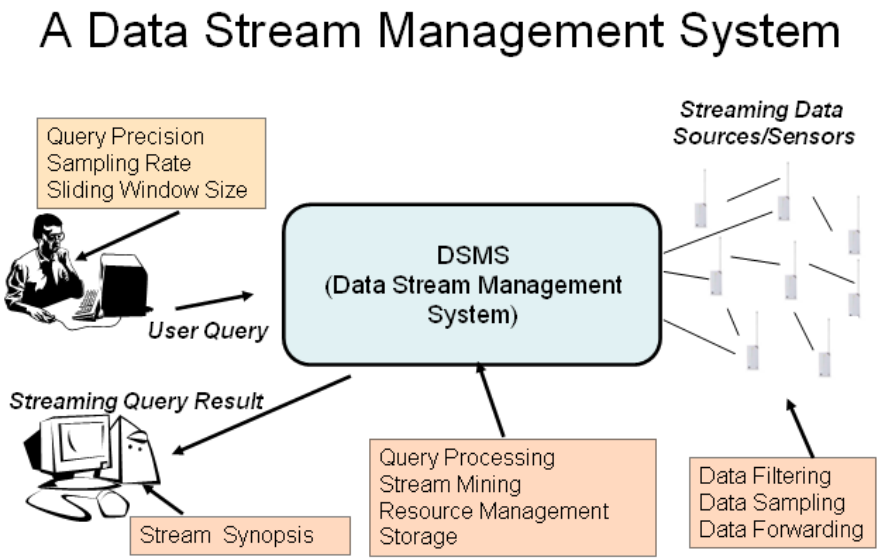
MARIO CATALDI

WWW.AI.UNIV-PARIS8.FR/~CATALDI/

M.CATALDI@IUT.UNIV-PARIS8.FR

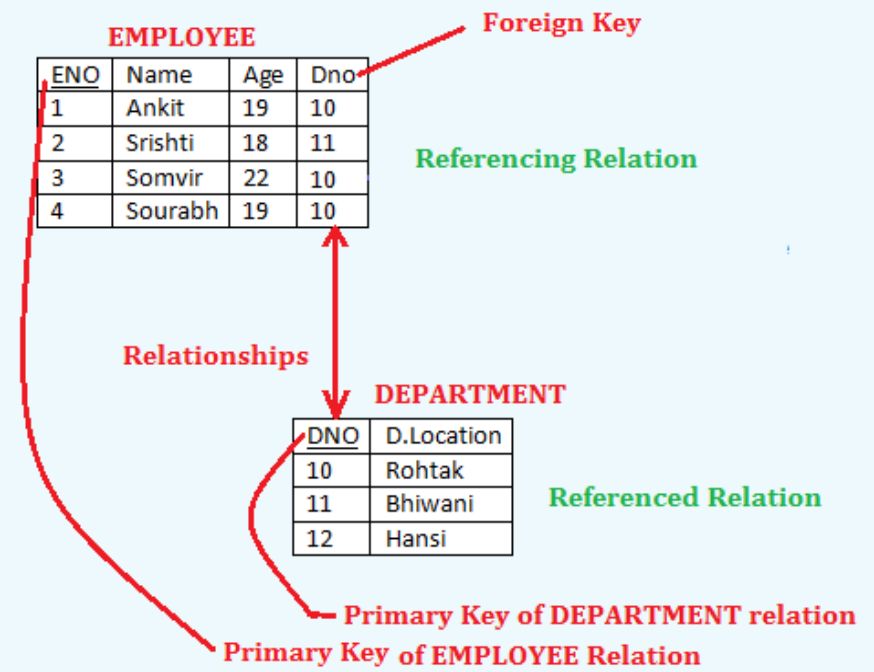
A data stream management system (**DSMS**) is a computer software system to manage continuous data streams.

It is similar to a database management system (DBMS), which is, however, designed for static data in conventional **databases.**



DATA STREAM MANAGEMENT SYSTEM

DBMS



DSMS

DATA STREAM MANAGEMENT SYSTEM

Data is *dynamic*!

- No longer possible to keep a static db
- Emerging need of novel techniques for updating our db in real time
- Social networks as source of information

DATA STREAM: SOCIAL NETWORKS

SOCIAL NETWORKS

What is a social network?

In the social sciences, a social network is a theoretical construct useful to study social relationships.

- ***Facebook***: over 800 million active users.
- ***Twitter:*** over 300 million users generating over 300 million tweets and handling over 1.6 billion search queries per day.
- ***LinkedIn:*** 21.4 million monthly unique U.S. visitors and 47.6 million globally
- ***Google+:*** 90 million users
- ***DBLP:*** info about isted more than 1.8 million articles on computer science

Why social networks?

- Huge amount of information about real data and persons!
- Interest from Research Communities, Public Institutions, Companies, etc.

...but, there is an emerging need of novel techniques to analyze this amount of data!

DATA STREAM SOCIAL NETWORKS

SOCIAL NETWORKS: BIG OPPORTUNITY

- Rich and big data: Billions users, billions contents
- Textual, Multimedia (image, videos, etc.)
- Billions of connections
- Behaviors, preferences, trends…

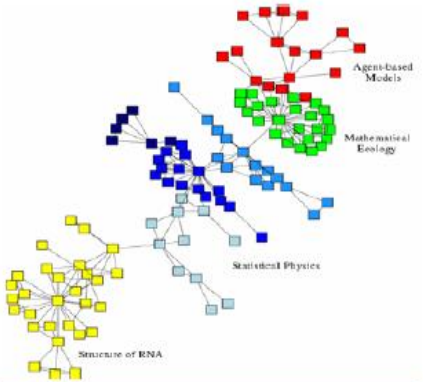
- Data is open: It’s easy to get data from Social Media
- Datasets
- Developers APIs
- Spidering the Web

REAL APPLICATIONS

- Consumer Brand Analytics: What are people saying about our brand?
- Marketing Communications
- Brand analytics helps to determine whether such campaigns are effective
- Product reviews: automatically mine product reviews for information on product features, new requests, … Easy to use, Comfortable chair, Light weight, Sturdy, Good price

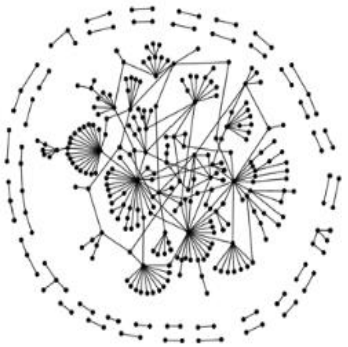
Online social networks

facebook LinkedIn

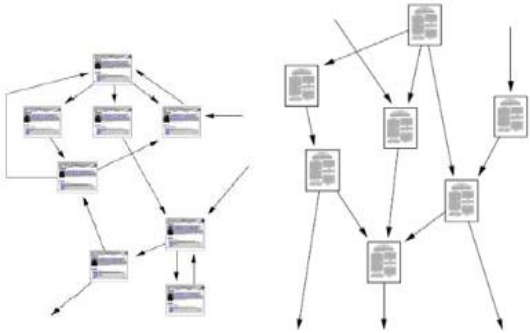


Collaboration networks

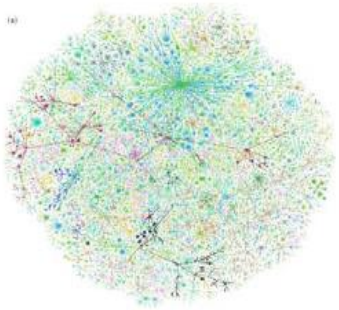
arXiv DBLP



Systems biology networks



Web graph & citation networks



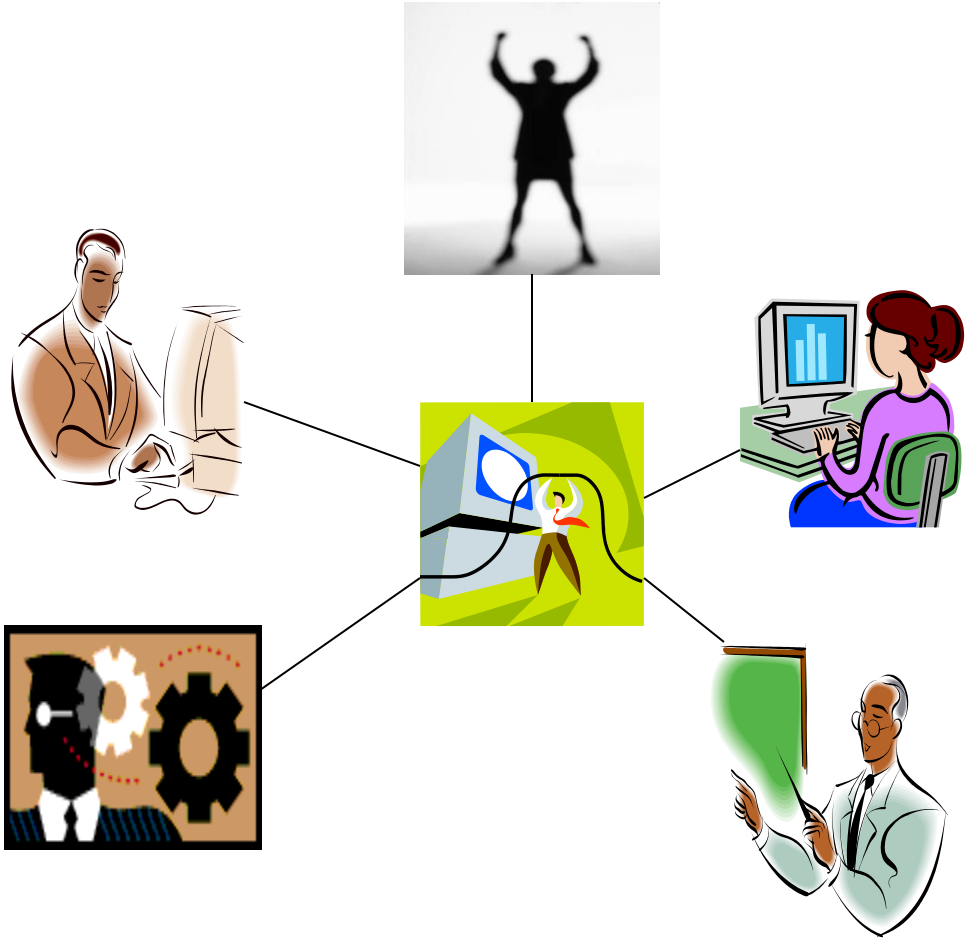
Internet

Communication networks

DATA STREAM & SOCIAL NETWORKS

SOCIAL NETWORK

A **social network** is a description of the social structure between actors, mostly individuals or organizations. It indicates the ways in which they are connected through various social familiarities ranging from casual acquaintance to close familiar bonds.



DATA STREAM MANAGEMENT SYSTEM

A Data Stream Management System needs:

1. A data stream extraction system to handle, ***in real time***, a flow of information
2. A data store system to properly store the relevant information into a limited amount of space
3. A data formalization model for studying the information and the structure of the extrated data.

DATA STREAM MANAGEMENT SYSTEM

A Data Stream Management System needs:

1. A data stream extraction system to handle, ***in real time***, a flow of information
2. A data store system to properly store the relevant information into a limited amount of space
3. A data formalization model for studying the information and the structure of the extrated data.

DATA STREAM MANAGEMENT SYSTEM

One of the biggest challenges for a DSMS is to handle **potentially infinite** data streams using a fixed amount of memory and no random access to the data.

Two main classes of approaches:

1. *compression techniques that try to summarize the*
2. *window techniques that try to portion the data into (finite) parts*.

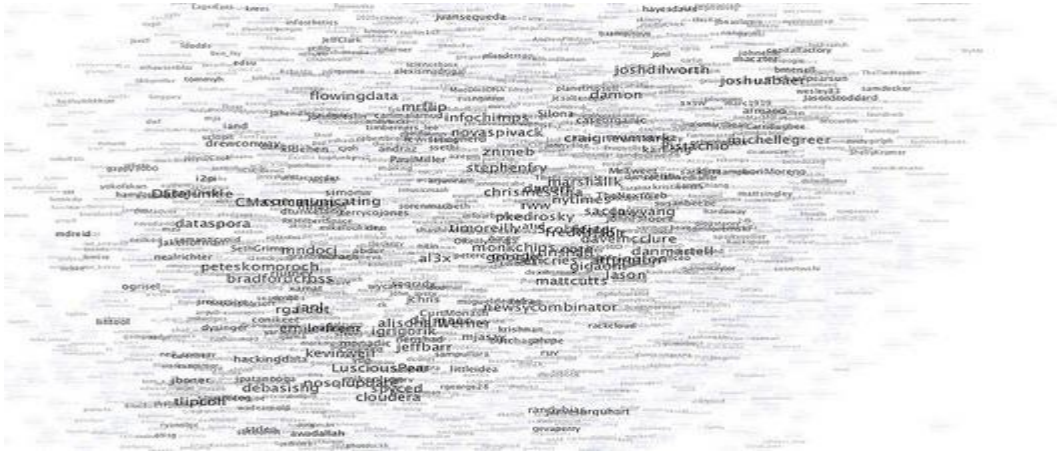
A data extraction platform tries to extract **meaningful information** from an unlimited flow of unstructured text data.

Text mining is an extension of data mining to textual data.
A social network contains a lot of data in the nodes of various forms.

1 – TEXT DATA EXTRACTION

PROBLEM:

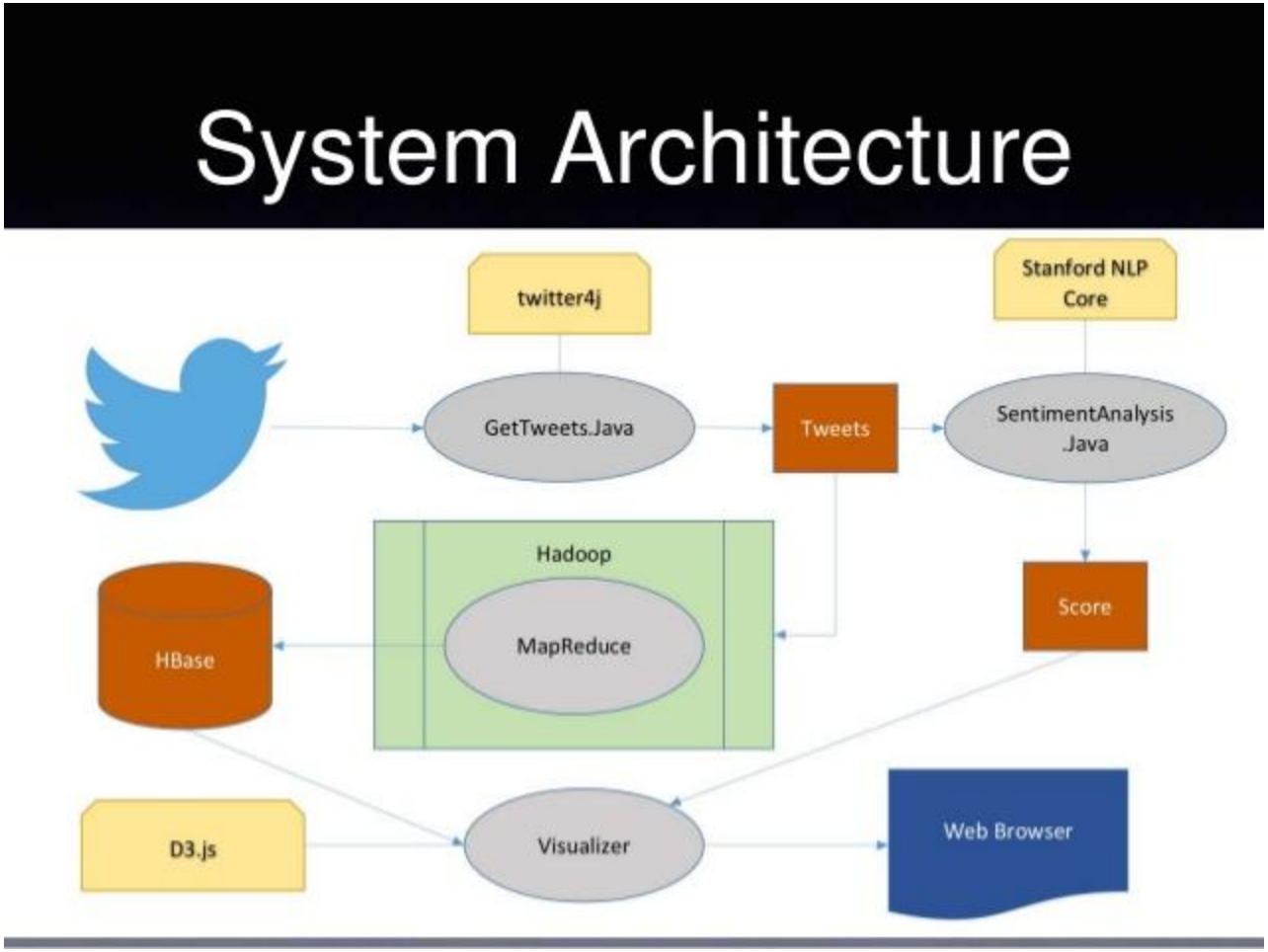
too much information.



SOLUTIONS:

- Threads (parallel computing)
- Volatile solutions (not everything needs to be threated and/or stored)
- Target the extraction before extracting the data

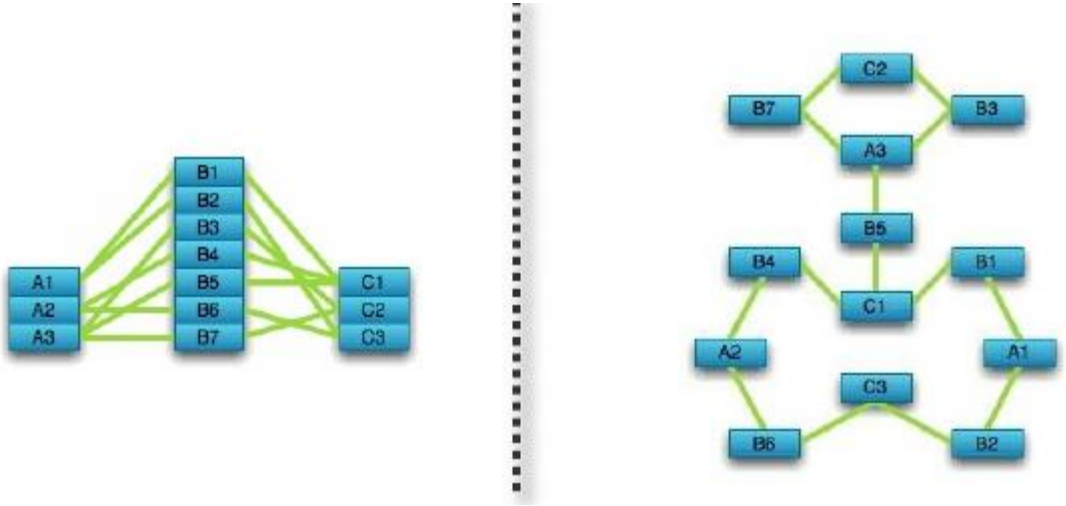
1 – TEXT DATA EXTRACTION



1 – TEXT DATA EXTRACTION

A DSMS obviously need a certain (finite) amount of space for storing the relevant data.

Many possible solutions: DB (relational, multimedia, graphs), Files, Memory.



2 – DATA STORE

It is necessary to reduce as much as possible the stored data.

Many techniques:

- simplification
- filtering
- data replacement (data is deleted after a considered amount of time and/or in presence of some predefined condition). The DB therefore contains only the latest relevant information

2 – DATA STORE

TIME FRAME SEGMENTATION: the overall data extraction time is divided in limited time frames, more manageable in terms of data size.

2 – DATA STORE

Social network: a social structure consists of nodes and ties

- Nodes are the individual actors within the networks
  - May be different kinds
  - May have attributes, labels or classes
- Ties are the relationships between the actors
  - May be different kinds
  - Links may have attributes, directed or undirected

3 – MODEL FORMALIZATION

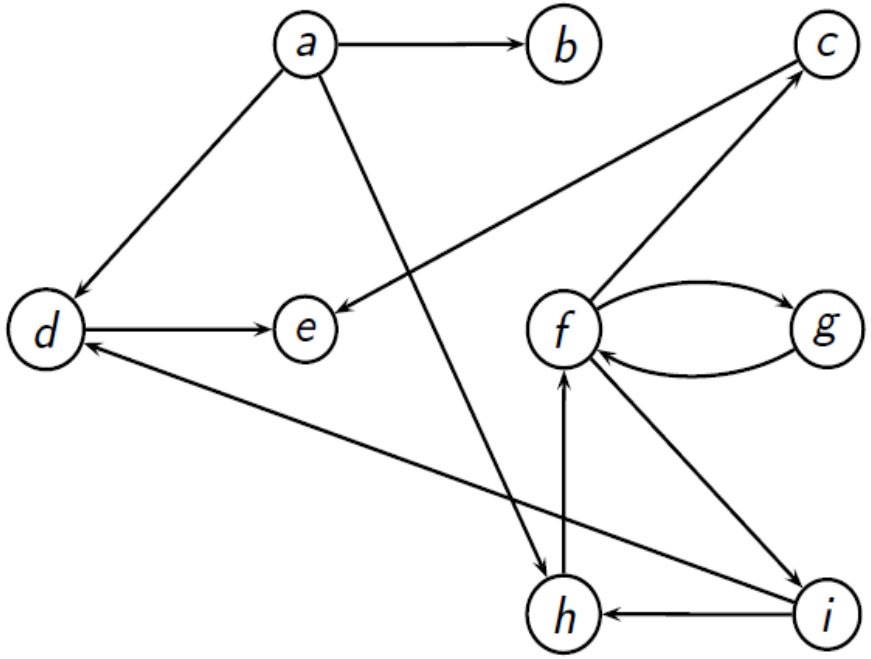
Homogeneous networks

- Single object type and single link type
- Single model social networks (e.g., friends)
- WWW: a collection of linked Web pages

Heterogeneous networks

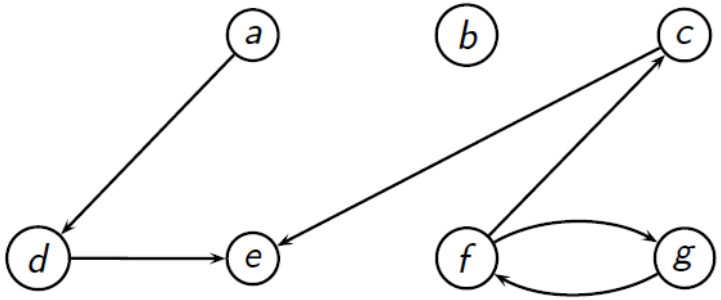
- Multiple object and link types
- Medical network: patients, doctors, disease, contacts, treatments
- Bibliographic network: publications, authors, venues

3 – MODEL FORMALIZATION

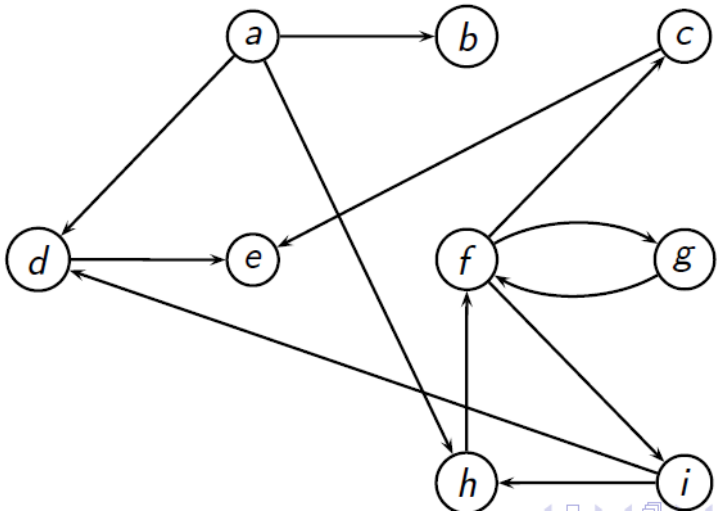


- $N = \{a, b, c, d, e, f, g, h, i\}$ (by default, node labels also are used as node names, when all labels are distinct)
- Edges are unlabelled.
- **example** : a and i are the predecessors of d, while e is a successor of d

WHAT IS A GRAPH?

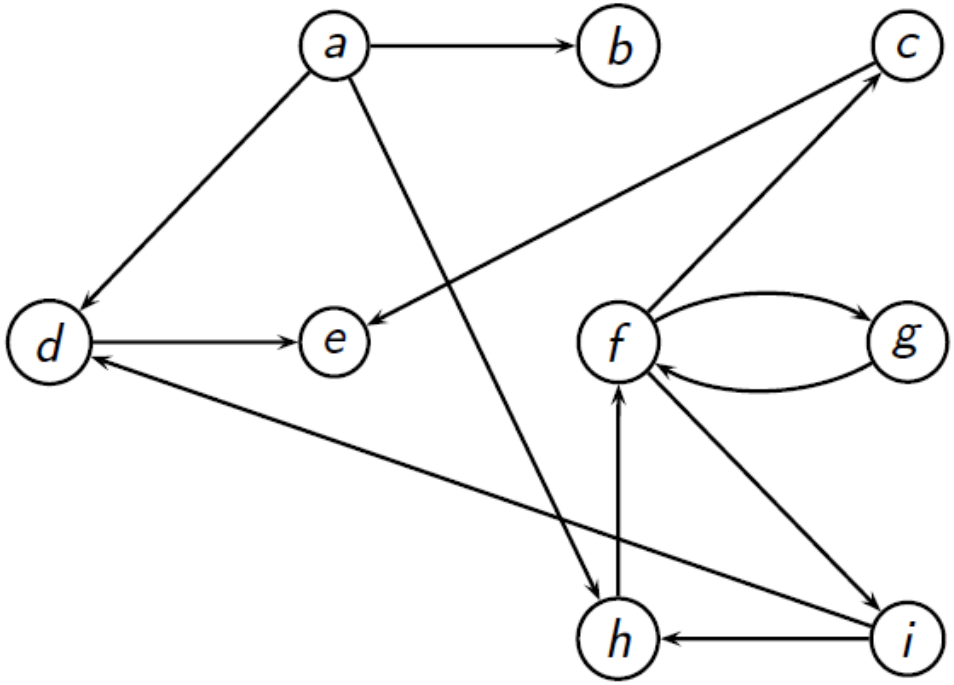


is a subgraph of : (here, no labels are considered)



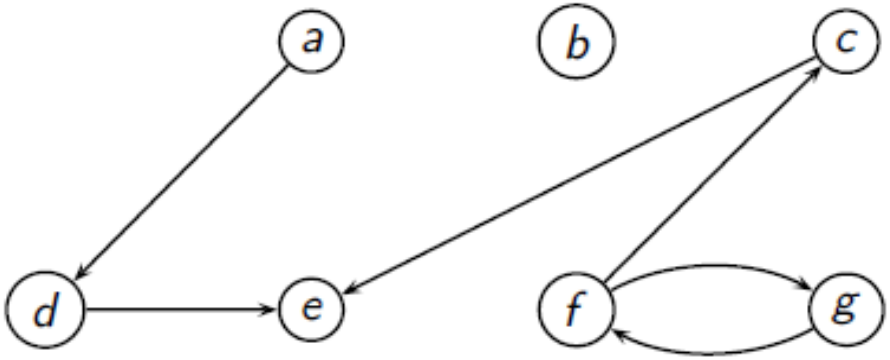
A portion of an existing graph (subset of vertex set and edges)

WHAT IS A SUBGRAPH?



$(a, h), (h, f), (f, c), (c, e)$ is a path
$(f, i), (i, h), (h, f)$ is a cycle

PATH AND CYCLES



	a	b	c	d	e	f	g
a				×			
b							
c					×		
d					×		
e							
f			×				×
g						×	

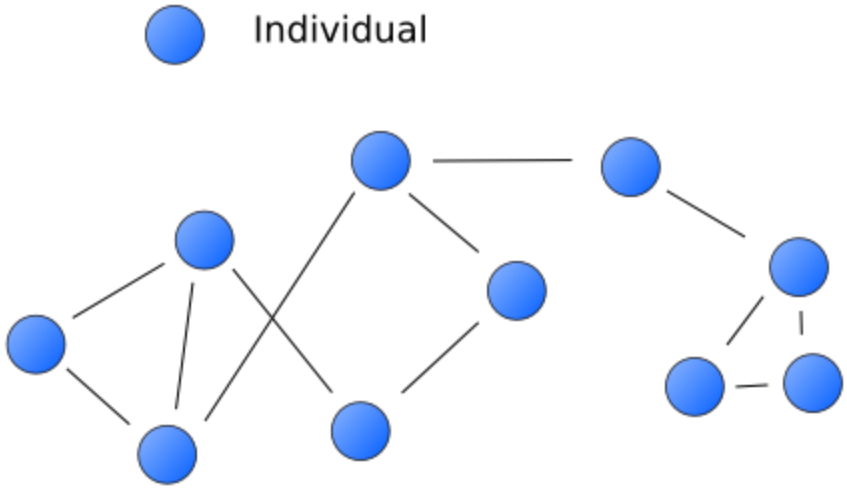
ADJACENCE MATRIX

Social network analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers or other information/knowledge processing entities.

The nodes in the network are the people and groups while the links show relationships or flows between the nodes.

Mining Subgraphs
- Topological orderings
- Strongly Connected Components

Graph Analysis
- Page Rank
- HITS



CLASSICAL GRAPH PROBLEMS IN SNA

SNA helps in analyzing the following facts:

- Short distances transmit information accurately and in a timely way, while long distances transmit slowly and can distort the information.
- Isolation - People that are not integrated well into a group and therefore, represent both untapped skills and a high likelihood of turnover.
- Highly expert people - Not being utilized appropriately.
- Organizational subgroups or cliques - Can develop their own subcultures and negative attitudes toward other groups.

CLASSICAL GRAPH PROBLEMS IN SNA

APPLICATION OF SNA:

- Realizing 9/11 Al- Qaeda Network.
- Build a grass roots political campaign.
- Determine influential journalists and analysts in the IT industry.
- Map executive's personal network based on email flows.
- Discover the network of Innovators in a regional economy.
- Analyze book selling patterns to position a new book and many more……

CLASSICAL GRAPH PROBLEMS IN SNA

CLASSICAL GRAPH PROBLEMS IN SNA

1. Degree Centrality:

   The number of direct connections a node has. What really matters is where those connections lead to and how they connect the otherwise unconnected.

2. Betweenness Centrality:

   A node with high betweenness has great influence over what flows in the network indicating important links and single point of failure.

3. Closeness Centrality:

   The measure of closeness of a node which are close to everyone else. The pattern of the direct and indirect ties allows the nodes any other node in the network more quickly than anyone else. They have the shortest paths to all others.

Frequent pattern: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set

Motivation: Finding inherent regularities in data

- What products were often purchased together?
- What are the subsequent purchases after buying a PC?
- What kinds of DNA are sensitive to this new drug?
- Can we classify web documents using frequent patterns?

MINING FREQUENT PATTERNS: WHAT IS IT GOOD FOR?

- Detection of cycles, search of strongly connected subgraphs.
- Route problems for labeled graphs : shortest path, minimum spanning tree, Chinese postman problem,
- Topological sort (nodes are ordered accordingly to the successor relation induced by edges and paths)
- Flow computation in networks
- Remark : Algorithm complexity is usually computed with respect to the number of nodes and/or the number of edges

CLASSICAL GRAPH PROBLEMS IN SNA

DEFINITION OF GRAPH

Empty Graph: For simplicity and expediency we ignore the possibility of a graph G being empty

Graph: is a data structure G = { V, E } consisting of a set E of edges and a set of V vertices, AKA nodes. Any node $v_i \in V$ may be connected to any other node v_j. Such a connection is called an edge. Edges may be directed, or even bi-directed. Different from a tree, a node in G may have any number of predecessors –or incident edges

Connected Graph: If all $n > 0$ nodes v_n in G are connected somehow, the graph G is called connected, regardless of edge directions

Strongly Connected Component: A subset $SG \subseteq G$ is strongly connected, if every $i > 0$ nodes v_i in SG can reach all v_i nodes in SG somehow

Directed Acyclic Graph (DAG): A DAG is a graph with directed edges that form no cycle. A node may still have multiple predecessors

A graph G(v, e) consists of nodes v and edges e

G is identified and thus accessible via one select node, called *entry node*, or simply entry, AKA *head*

G is not necessarily connected

- If parts of G are unconnected, how can they be retrieved in case of a necessary, complete graph traversal?

Several methods of forcing complete access:

- Either create a super-node, not specified by the user of G, in a way that each unconnected region is pointed at
- Or have a linked-list (LL) meandering through each node of G, without this link field being part of G proper

GRAPH DATA STRUCTURE

TOPOLOGICAL SORT

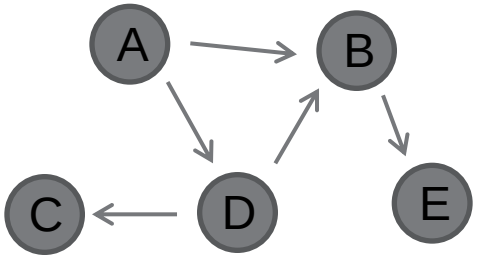
Topological sort of a DAG (directed acyclic graph):

- Linear ordering of all vertices in graph G such that vertex *u* comes before vertex *v* if edge $(u, v) \in G$

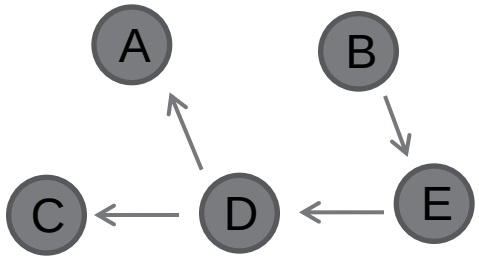
Real-world application:

✓Scheduling

✓Spread of information



politics



sports

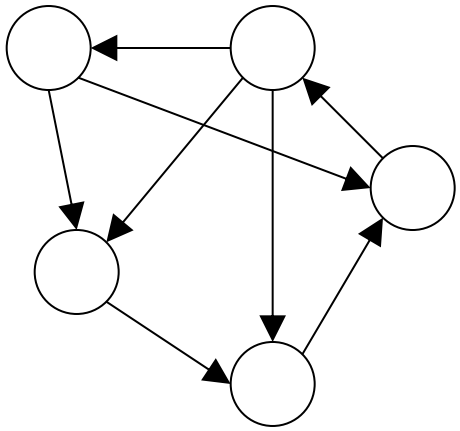
Example of topological sort: to trace the **spread** of the information (also termed as *contamination* in literature) in the network along with the cause of its spread.

EXAMPLE: INFORMATION FLOW

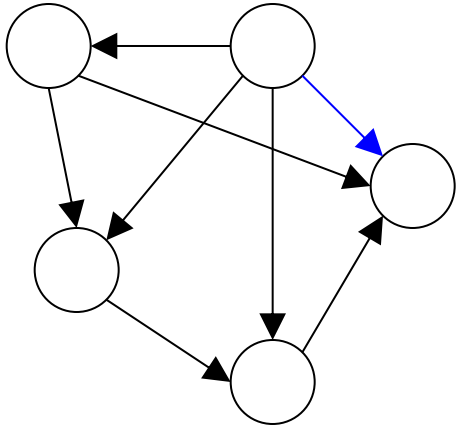
Every pair of vertices are reachable from each other

Graph *G* is *strongly connected* if, for every *u* and *v* in *V*, there is some path from *u* to *v* and some path from *v* to *u*.

Strongly Connected



Not Strongly Connected



STRONGLY CONNECTED COMPONENTS

Every node v_i in a strongly connected component SCC of graph G can reach every node v_i, (not necessarily in one single step).

An SCC is a subgraph SG of graph G, $SG \subseteq G$

By definition then, a singleton node graph is strongly connected; (not very interesting…)

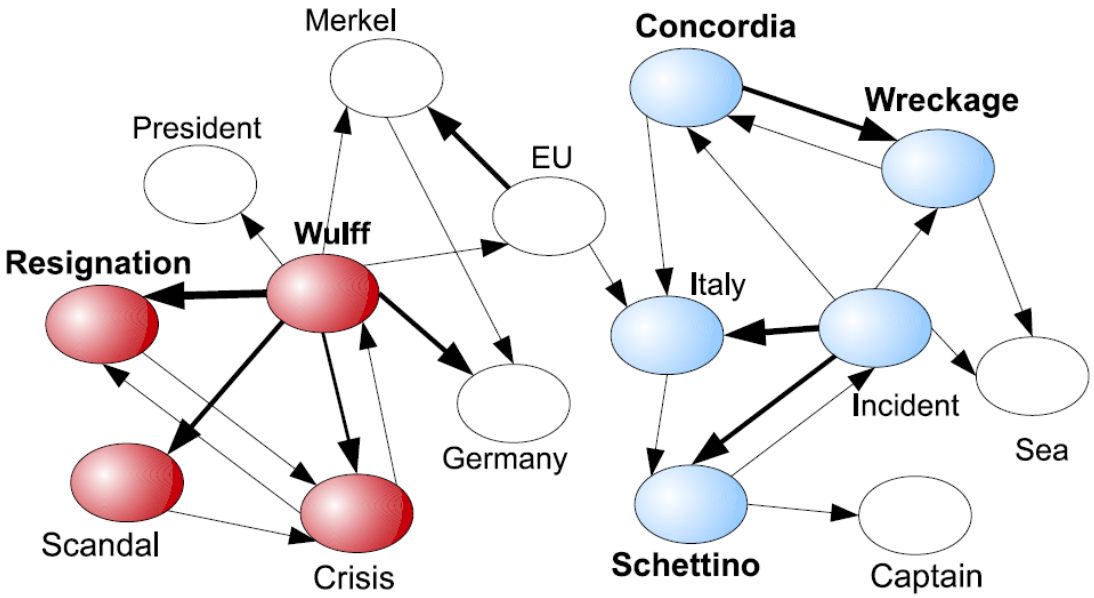
Graph needs defined entry point: named entry or head

STRONGLY CONNECTED COMPONENT

STRONGLY CONNECTED COMPONENTS: APPLICATION

Findig SCC could be useful to detect topics

Topics can be defined as a coherent set of semantically related terms that express a single argument.



In order to retrieve the most emerging topics, we consider the set of emerging keywords, computed as before, and we search for the strongly connected components (SCC) rooted on them in the topic graph (formed by keywords).

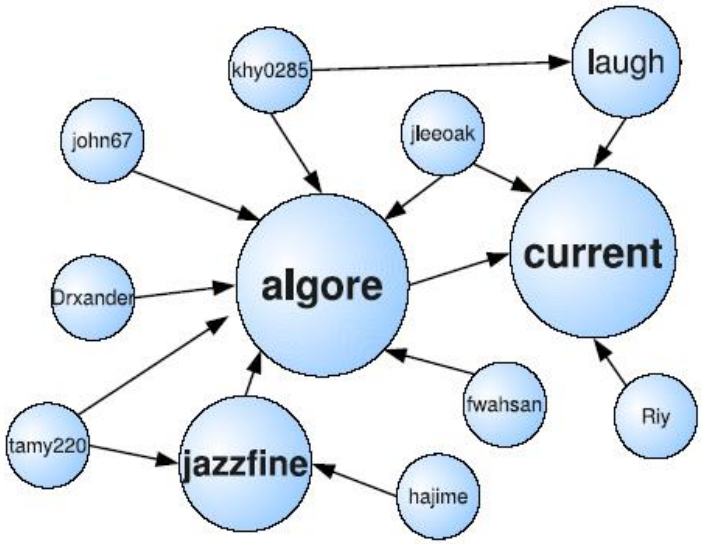
LINK ANALYSIS: PAGE RANK

In a graph, we could also take into account the existing links among the nodes to perform alternative analysis.

The page rank algorithm could help estimate the popularity of a node

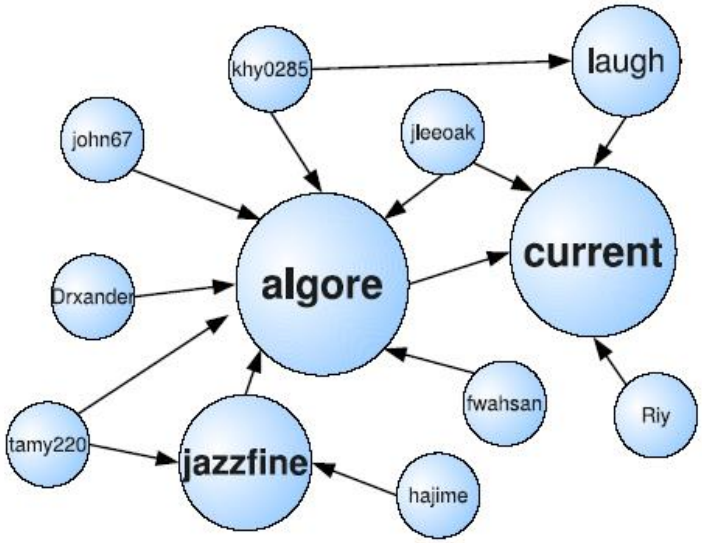
Links can bring a lot of information

Link between nodes= perticene relation



LINK ANALYSIS: PAGE RANK

We define an author-based graph ***G(U, F)*** where ***U*** is the set of users and ***F*** is the set of directed edges; thus, given two users *u1* and *u2*, the edge <*u1, u2*> exists only if u1 is a follower of u2 .



..thus, we measure the degree of importance of each user by analyzing the connectivity in G;

Page Rank on G

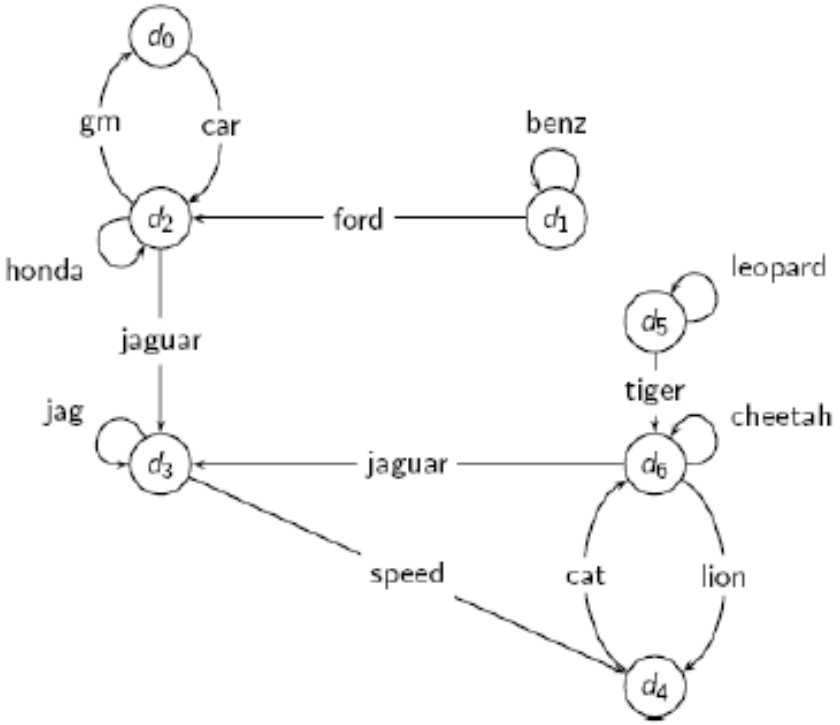
LINK ANALYSIS: PAGE RANK

***Idea*:** a surfer doing a random walk within a graph

- At every step, the surfer exits from the current node following an existing link (same probability for each link).

There is also the probability to remain on the same node (same node).

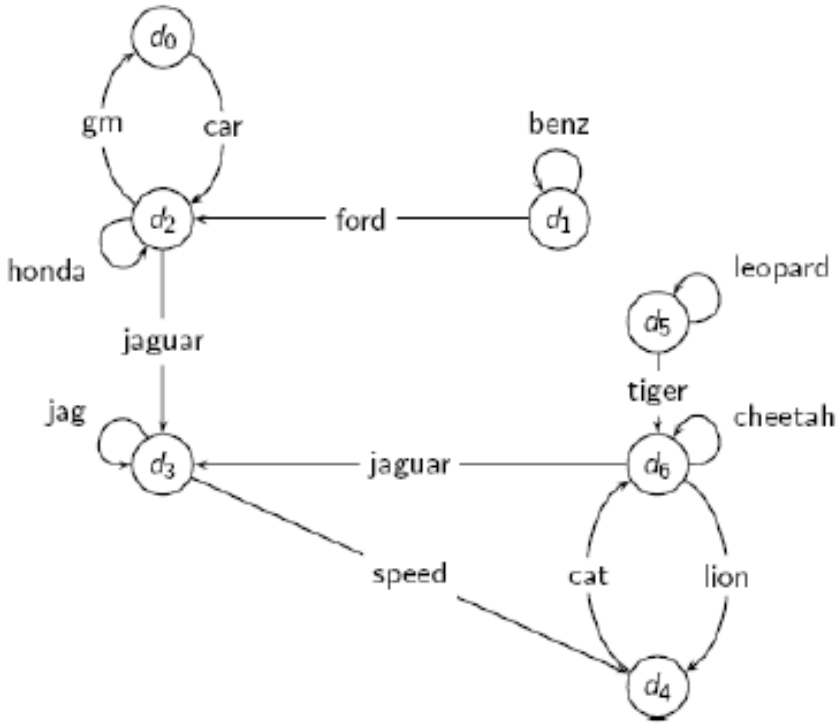
It is easy to imagine that the most visited nodes are the ones with highest degree of incoming edges *(ie, the most important/popular).*



Link Matrix

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0	0	1	0	0	0	0
d_1	0	1	1	0	0	0	0
d_2	1	0	1	1	0	0	0
d_3	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1
d_5	0	0	0	0	0	1	1
d_6	0	0	0	1	1	0	1

EXAMPLE



Transition Matrix

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

EXAMPLE

But, for each step, we still have the probability p to jump to another randomly selected node.

…and a probability 1-p to follow one of the outcoming edges..

LINK ANALYSIS: PAGE RANK

But, for each step, we still have the probability p to jump to another randomly selected node.

…and a probability 1-p to follow one of the outcoming edges..

P=0.14

0.02	0.02	0.88	0.02	0.02	0.02	0.02
0.02	0.45	0.45	0.02	0.02	0.02	0.02
0.31	0.02	0.31	0.31	0.02	0.02	0.02
0.02	0.02	0.02	0.45	0.45	0.02	0.02
0.02	0.02	0.02	0.02	0.02	0.02	0.88
0.02	0.02	0.02	0.02	0.02	0.45	0.45
0.02	0.02	0.02	0.31	0.31	0.02	0.31

LINK ANALYSIS: PAGE RANK

Page Rank:

- we start from a randomly picked node
- We start surfing by using the transition matrix.
- We then follow the links with the probability calculted in the matrix.

The probability to get a single node is the page rank value.

LINK ANALYSIS: PAGE RANK

But we could also distinguish between two types of nodes…

- Hubs : nodes that point out to important nodes.
- Authorities : important nodes (reference node with respect to a subject)

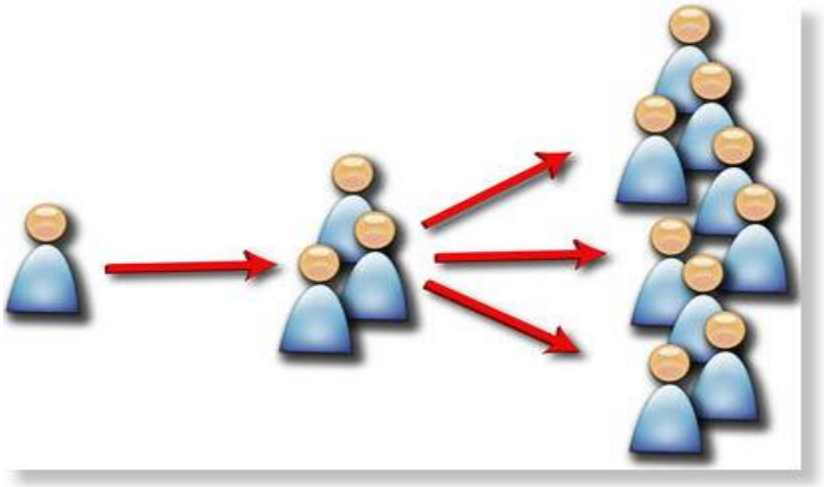
Idea:

- Every node will have two values: H and A

HITS

HITS algorithm could be easily applied to analyze the “contagion” and modeling the paths of its propagation in the network in order to understand the *cause* of a popularity.

Some node is a provider (authority), many more act as simple distributors (hubs).



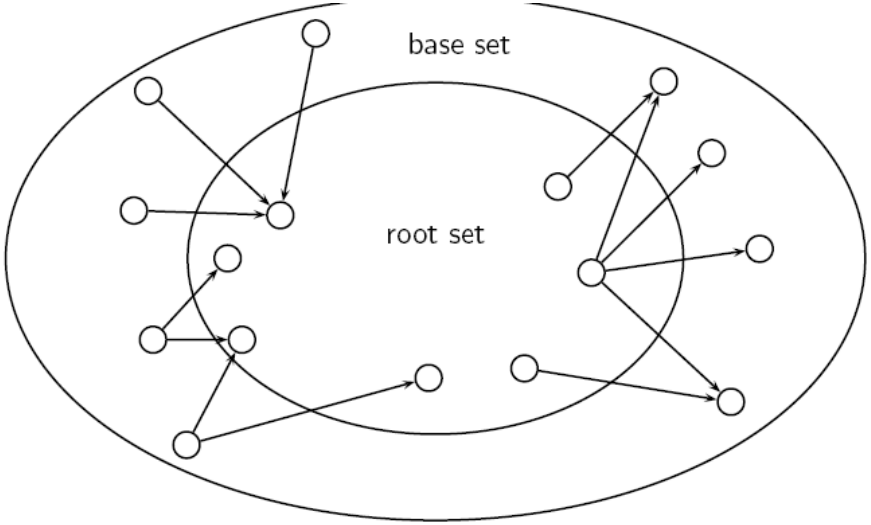
HITS: APPLICATION

HITS

We start searching the pages related to one argument: results: seed set

Then we search for the connected nodes -> base set

This is the base for the hits alorithim.



Iterative algorithm

These slides are available at:

http://www.ai.univ-paris8.fr/~cataldi/DSMS/slides.pdf

The project document description is moreover available at

http://www.ai.univ-paris8.fr/~cataldi/DSMS/project.pdf

m.cataldi@iut.univ-paris8.fr

AT WORK!