Welcome to

DS504/CS586: Big Data Analytics

Graph Mining II

Prof. Yanhua Li

Time: 6:00pm –8:50pm Mon. and Wed.
Location: SL105
Spring 2016

Reading assignments

We will increase the bar a little bit
Please add more of your ideas, and share them with us in the class.

Project 1

Team 1:

Utilizing the data from Allstate to predict the possible Bodily Injury Liability Insurance claim payments that the company may pay in line with the vehicle features.

Team 2:

3D Video Growing Trend on YouTube

Project 1

Team 3:

Designing a sampling method that can estimate the available capacity of hotels in a large area and a certain time range as accurate as possible.

Team 4:

A huge number of professional as well as amateur programmers use Stackoverflow.com to find solutions to their questions regularly. Thus, our team seeks to give such users a general idea of which skills to learn to create their projects and how to manage their skills by offering a project-related heat map of programming languages and knowledge points.

Project 1

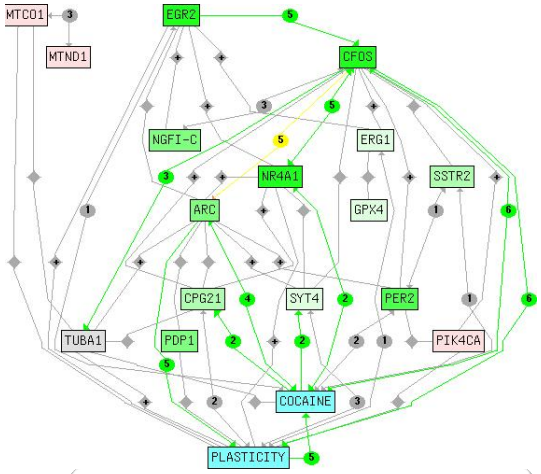
Team 5:

MEASURING RESTAURANT DIVERSITY INDEX FOR DIFFERENT CITIES in Yelp

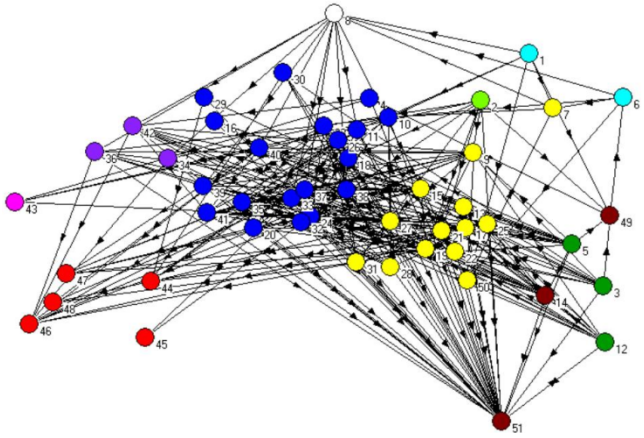
Team 6:

We will estimate the number of valid and invalid users among the population. Lastly, we will analyze other related web site statistics such as passive and active users.

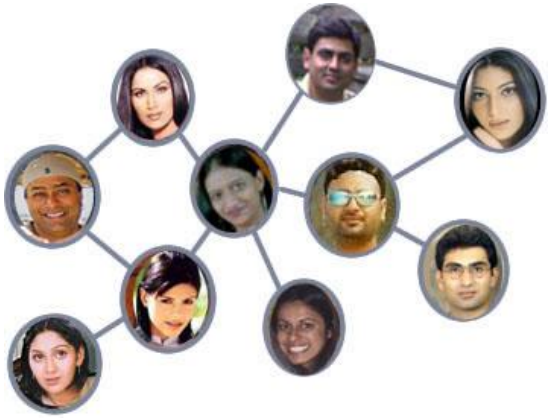
- Graphs are everywhere.



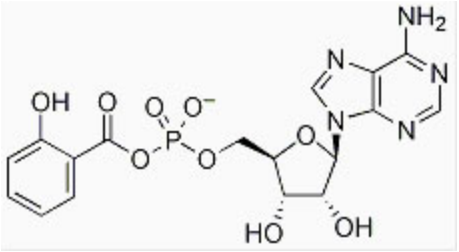
Biological Network



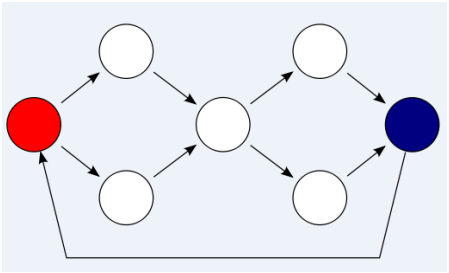
Ecological Network



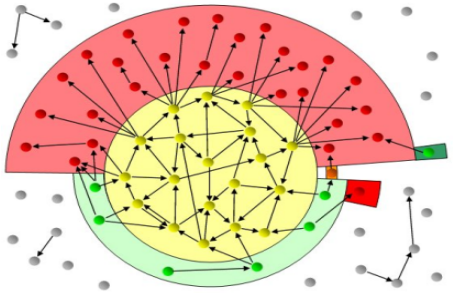
Social Network



Chemical Network

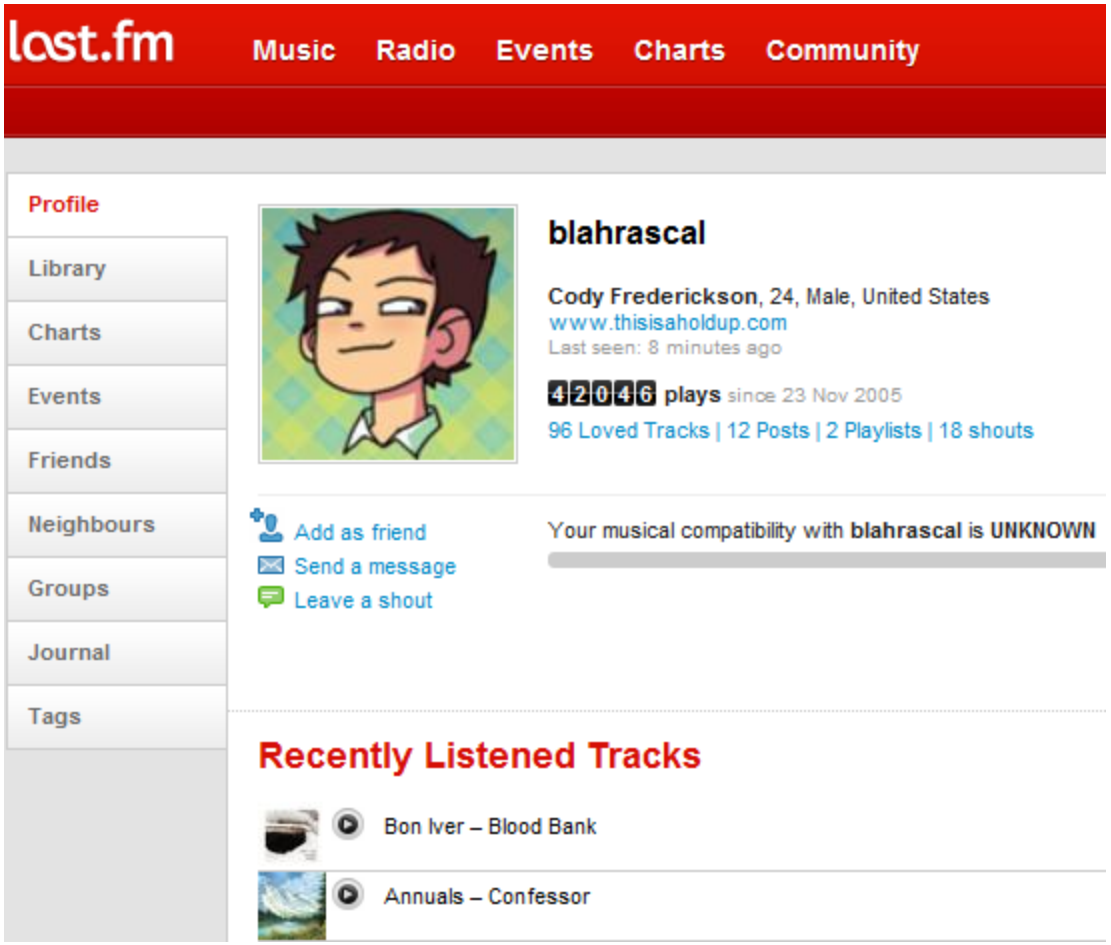


Program Flow



Web Graph

- Real-life graph contains complex contents – labels associated with nodes, edges and graphs.



Node Labels:

Location, Gender, Charts, Library, Events, Groups, Journal, Tags, Age, Tracks.

- Large Scale Graphs.

	# of Users	# of Links
Facebook	***400 Million***	52K Million
Twitter	105 Million	10K Million
LinkedIn	60 Million	0.9K Million
Last.FM	40 Million	2K Million
LiveJournal	25 Million	2K Million
del.icio.us	5.3 Million	0.7K Million
DBLP	0.7 Million	8 Million

Mining in Big Graphs

- Network Statistic Analysis (last lecture)
  - Network Size
  - Degree distribution.

- Node Ranking (this lecture)
  - Identifying most important/influential nodes
  - Viral Marketing, resource allocation

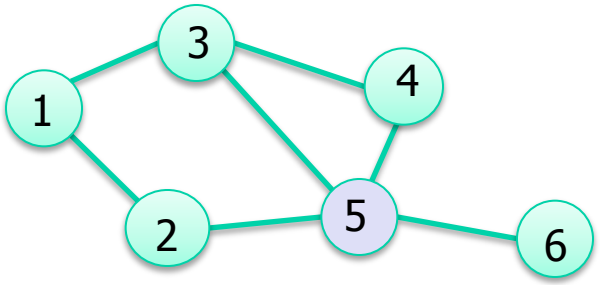
Characterize Node Importance

- Rank the webpages in search engine.
- Viral Marketing, resource allocation
- Open a new restaurant, find the optimal location
- …

Ranking nodes on an undirected graph

Node Degree

- **Local Importance**



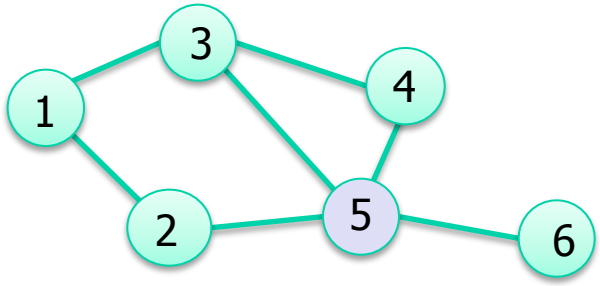
- **|V|=6**
- **|E|=7**

- **d(5)=4**
- **d(3)=3**
- **d(4)=2**
- **d(2)=2**
- **d(1)=2**
- **d(6)=1**

Stationary distribution

Connected Graphs

- **Global Importance**



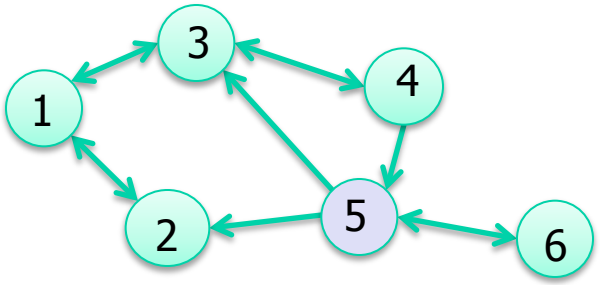
- **π(5)=4/14**
- **π(3)=3/14**
- **π(4)=2/14**
- **π(2)=2/14**
- **π(1)=2/14**
- **π(6)=1/14**

They are equivalent.

Ranking nodes on a directed graph

Node in & out Degree

- **Local Importance**



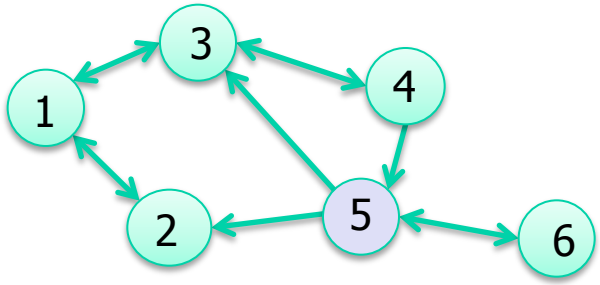
- **$d_{in}(3)=3$; $d_{out}(5)=3$;**
- **$d_{in}(5)=2$; $d_{out}(3)=2$;**
- **$d_{in}(1)=2$; $d_{out}(1)=2$;**
- **$d_{in}(2)=2$; $d_{out}(4)=2$;**
- **$d_{in}(4)=1$; $d_{out}(2)=1$;**
- **$d_{in}(6)=1$; $d_{out}(6)=1$;**

They are equivalent?

Stationary distribution

Strongly Connected Graphs & Aperiodic

- **Global Importance**



- **$\pi(5)$=?**
- **$\pi(4)$=?**
- **$\pi(3)$=?**
- **$\pi(2)$=?**
- **$\pi(1)$=?**
- **$\pi(6)$=?**

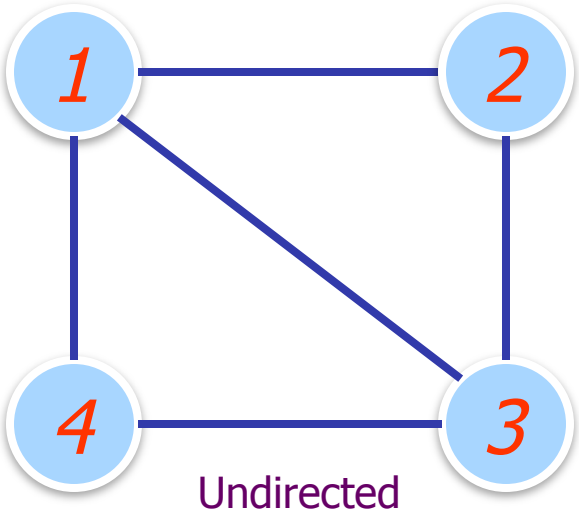
Random Walk (Undirected Graph)

Connected Graphs

- Adjacency matrix

$$A=\begin{pmatrix}0&1&1&1\\1&0&1&0\\1&1&0&1\\1&0&1&0\end{pmatrix}\quad D=\begin{pmatrix}3&0&0&0\\0&2&0&0\\0&0&3&0\\0&0&0&2\end{pmatrix}$$

Symmetric



Undirected

- Transition Probability Matrix

$$P_{ij}=\frac{1}{k_i}$$

$$x_{t,i}=\sum_j x_{t-1,j}p_{ji}$$

$$P=A\bullet D^{-1}=\begin{pmatrix}0&1/3&1/3&1/3\\1/2&0&1/2&0\\1/3&1/3&0&1/3\\1/2&0&1/2&0\end{pmatrix}$$

- |E|: *number of links*
- *Stationary Distribution*

$$\pi_i=\frac{d_i}{2|E|}$$

- **π(1)=3/10**
- **π(3)=3/10**
- **π(2)=2/10**
- **π(4)=2/10**

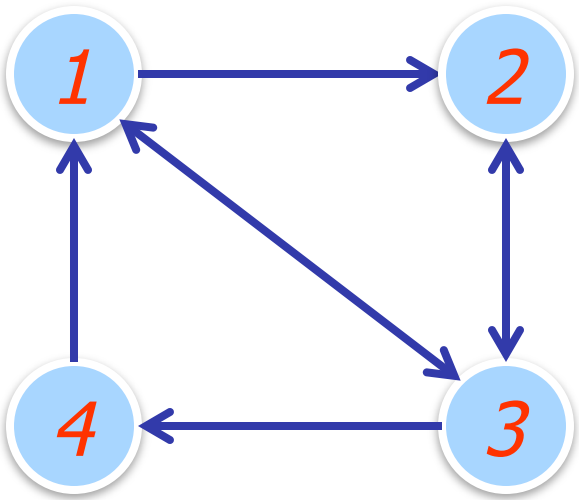
Random Walk (directed graph)

Strongly Connected Graphs & Aperiodic

- Adjacency matrix

$$A=\begin{pmatrix}0&1&1&0\\0&0&1&0\\1&1&0&1\\1&0&0&0\end{pmatrix}\quad D=\begin{pmatrix}2&0&0&0\\0&1&0&0\\0&0&3&0\\0&0&0&1\end{pmatrix}$$

Asymmetric



- Transition Probability Matrix

$$P_{ij}=\frac{1}{k_{out,i}}$$

$$x_{t,i}=\sum_j x_{t-1,j}p_{ji}$$

$$P=A\bullet D^{-1}=\begin{pmatrix}0&1/2&1/2&0\\1/2&0&1/2&0\\1/3&1/3&0&1/3\\1&0&0&0\end{pmatrix}$$

- |E|: *number of directed links*
- *Stationary Distribution*

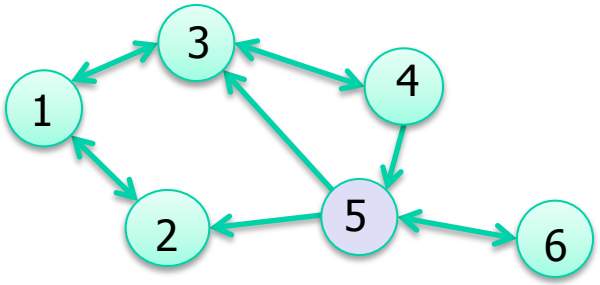
$$\pi_i \neq \frac{d_i}{2|E|}$$

- **π(1)=6/18=1/3**
- **π(2)=4/18=2/9**
- **π(3)=3/18=1/6**
- **π(4)=5/18**

Ranking nodes in a directed graph

Node in & out Degree

- **Local Importance**

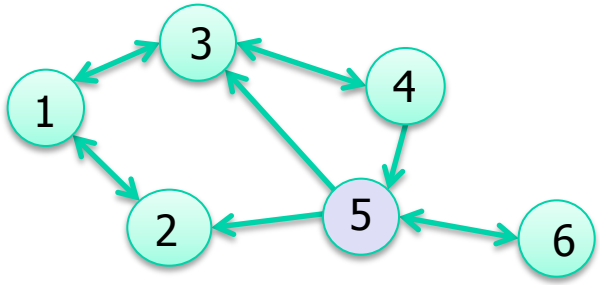


- $d_{in}(3)=3$; $d_{out}(5)=3$;
- $d_{in}(5)=2$; $d_{out}(3)=2$;
- $d_{in}(1)=2$; $d_{out}(1)=2$;
- $d_{in}(2)=2$; $d_{out}(4)=2$;
- $d_{in}(4)=1$; $d_{out}(2)=1$;
- $d_{in}(6)=1$; $d_{out}(6)=1$;

Stationary distribution

Strongly Connected Graphs & Aperiodic

- **Global Importance**



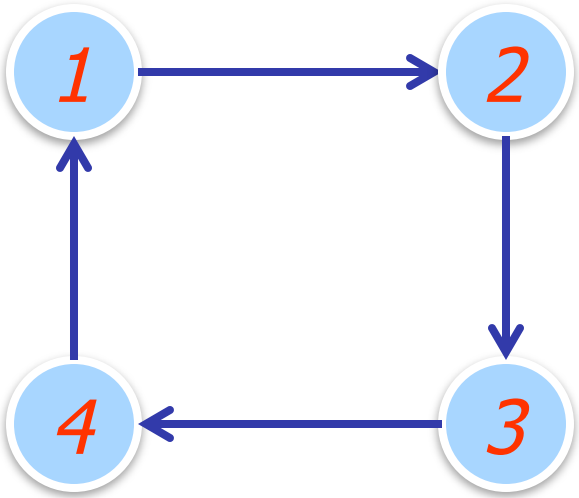
- $\pi(1)=5/16$
- $\pi(3)=1/4$
- $\pi(2)=3/16$
- $\pi(4)=1/8$
- $\pi(5)=3/32$
- $\pi(6)=1/32$

They are no longer equivalent.

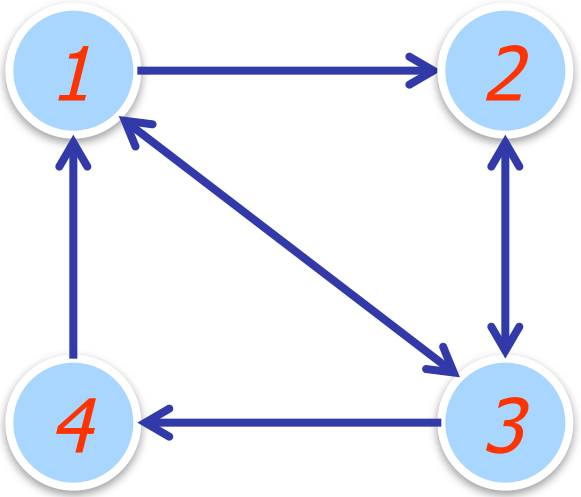
directed graphs

Strongly Connected Graphs & Aperiodic

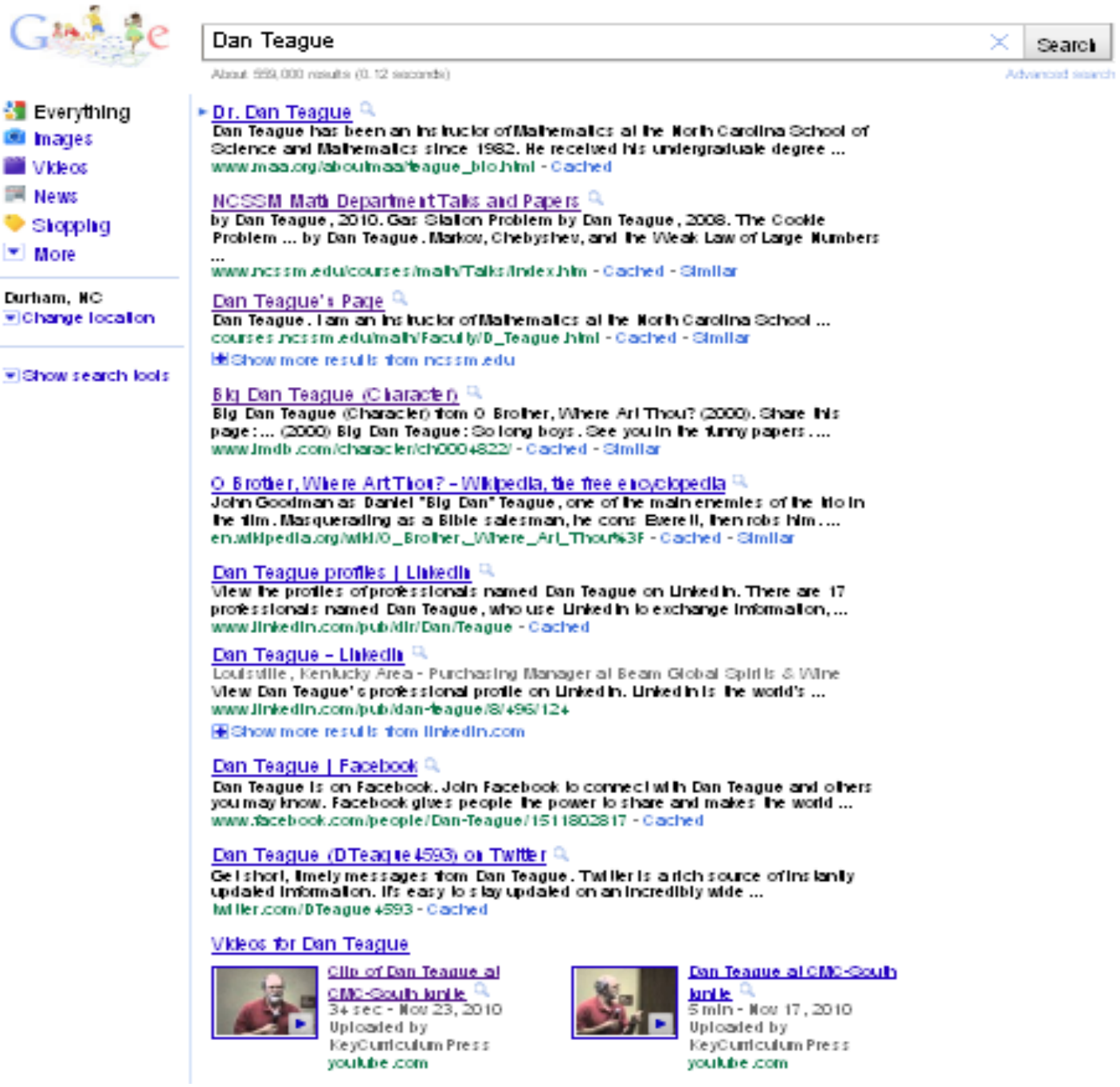
- Periodic
- vs
- Aperiodic Graphs
  - **The greatest common divisor of the lengths of its cycles is one or not**



- Disconnected graph
- vs
- Connected graph
  - Strongly Connected
  - vs
  - Weakly Connected



- Ergodic: Strongly Connected and Aperiodic

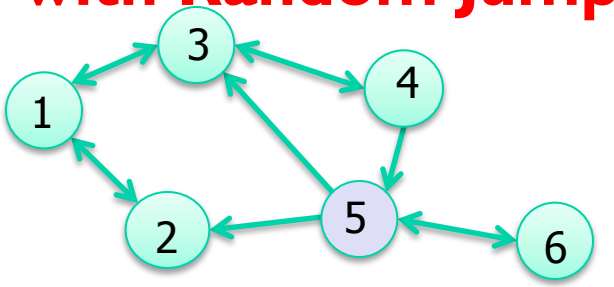


Why This Order?

Ranking nodes in a directed graph (II)

PageRank

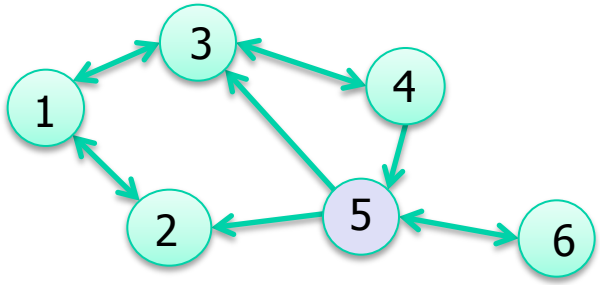
- **Random Walk**
- **with Random Jumps**



- **R(3)=?;**
- **R(5)=?;**
- **R(1)=?;**
- **R(2)=?;**
- **R(4)=?;**
- **R(6)=?;**

HITS

- Hub & Authority



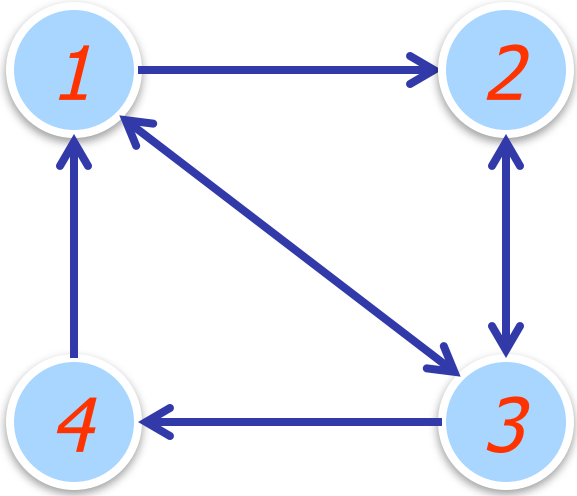
- **$R_a(3)$=?; $R_h(5)$=?;**
- **$R_a(5)$=?; $R_h(3)$=?;**
- **$R_a(1)$=?; $R_h(1)$=?;**
- **$R_a(2)$=?; $R_h(4)$=?;**
- **$R_a(4)$=?; $R_h(2)$=?;**
- **$R_a(6)$=?; $R_h(6)$=?;**

They are no longer equivalent.

Naïve PageRank

- Adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad D = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



- Transition Probability Matrix

$$P_{ij} = \frac{1}{k_{out,i}}$$

$$R_i = \sum_j R_j p_{ji}$$

$$P = A \bullet D^{-1} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

- *Stationary Distribution*

$$R_i = \pi_i$$

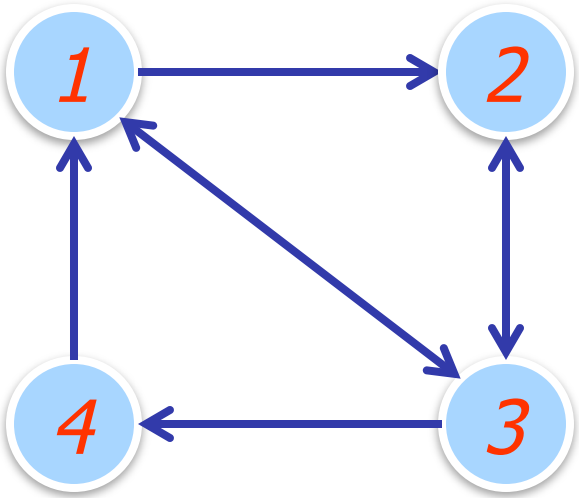
- **π(1)=6/18=1/3**
- **π(2)=4/18=2/9**
- **π(3)=3/18=1/6**
- **π(4)=5/18**

- ***Disconnected Graph & Random surfing behaviors***

Standard PageRank

- Adjacency matrix

$$A=\begin{pmatrix}0&1&1&0\\0&0&1&0\\1&1&0&1\\1&0&0&0\end{pmatrix}\quad D=\begin{pmatrix}2&0&0&0\\0&1&0&0\\0&0&3&0\\0&0&0&1\end{pmatrix}$$



- Transition Probability Matrix (d=0.85)

$$P_{ij}=\frac{1}{k_{out,i}}$$

$$R_i=d\sum_j R_j p_{ji}+(1-d)\frac{1}{n}$$

$$P=A\bullet D^{-1}=\begin{pmatrix}0&1/2&1/2&0\\1/2&0&1/2&0\\1/3&1/3&0&1/3\\1&0&0&0\end{pmatrix}$$

- *Stationary Distribution (J is all-1 matrix).*

$$R_i=\pi_{pr,i}$$

$$P_{pr}=d\bullet P+(1-d)\frac{1}{n}J=\begin{pmatrix}0.0375&0.4625&0.4625&0.0375\\0.0375&0.0375&0.0375&0.8875\\0.3208&0.3208&0.0375&0.3208\\0.8875&0.0375&0.0375&0.0375\end{pmatrix}$$

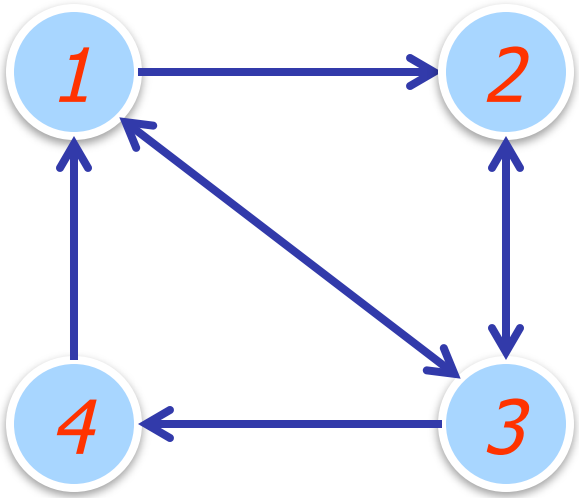
- *Convergence*
  - *Leading eigenvector of P_{pr}*

- How to quantify the importance as a hub and authority separately?

Hub & Authority (HITS)

- Adjacency matrix

$$A=\begin{pmatrix}0&1&1&0\\0&0&1&0\\1&1&0&1\\1&0&0&0\end{pmatrix}\quad D=\begin{pmatrix}2&0&0&0\\0&1&0&0\\0&0&3&0\\0&0&0&1\end{pmatrix}$$



- Hub and authority
  - Initial Step: $hub(p)=1; auth(p)=1;$
  - Each step with normalization:

$$hub(p)=\sum_{i=1}^{n} auth(i); \qquad hub(p)=\frac{hub(p)}{\sqrt{\sum_{i=1}^{n} hub(i)^2}};$$

$$auth(p)=\sum_{i=1}^{n} hub(i); \qquad auth(p)=\frac{auth(p)}{\sqrt{\sum_{i=1}^{n} auth(i)^2}};$$

- *Convergence*
  - *hub and authority are the left and right singular vector of the adjacency matrix A.*

A Note on Maximizing the Spread of Influence in Social Networks

E. Even-Dar and A. Shapira

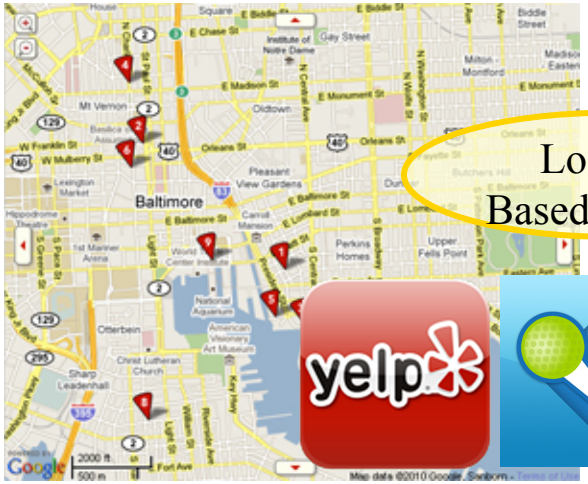
Social Influence

Social Influence

Instant Messaging

Collaboration networks

Sharing sites



Location Based Services

Social networks

Microblogs

Social Influence

LinkedIn Account Type: Basic | Upgrade

The Power of Social Recommendation

Kai Fu Lee · December 03, 2012

A few days ago, I had a chance to try a particular brand of frozen dessert maker, which was only available in the United States. I sent out a "tweet" in China on Sina Weibo, sharing my experience.

To my surprise, this tweet was retweeted over 100,000 times, as Chinese young people and parents showed great interest in having such a machine. Even more surprising, within 3 hours, Taobao (China's eBay) had hundreds of sellers, offering to buy such a machine in the US and shipping it to China for the buyer. Even though the shipping cost was more than the machine, thousands were sold within a day (one store reported 51 sales, and there were over

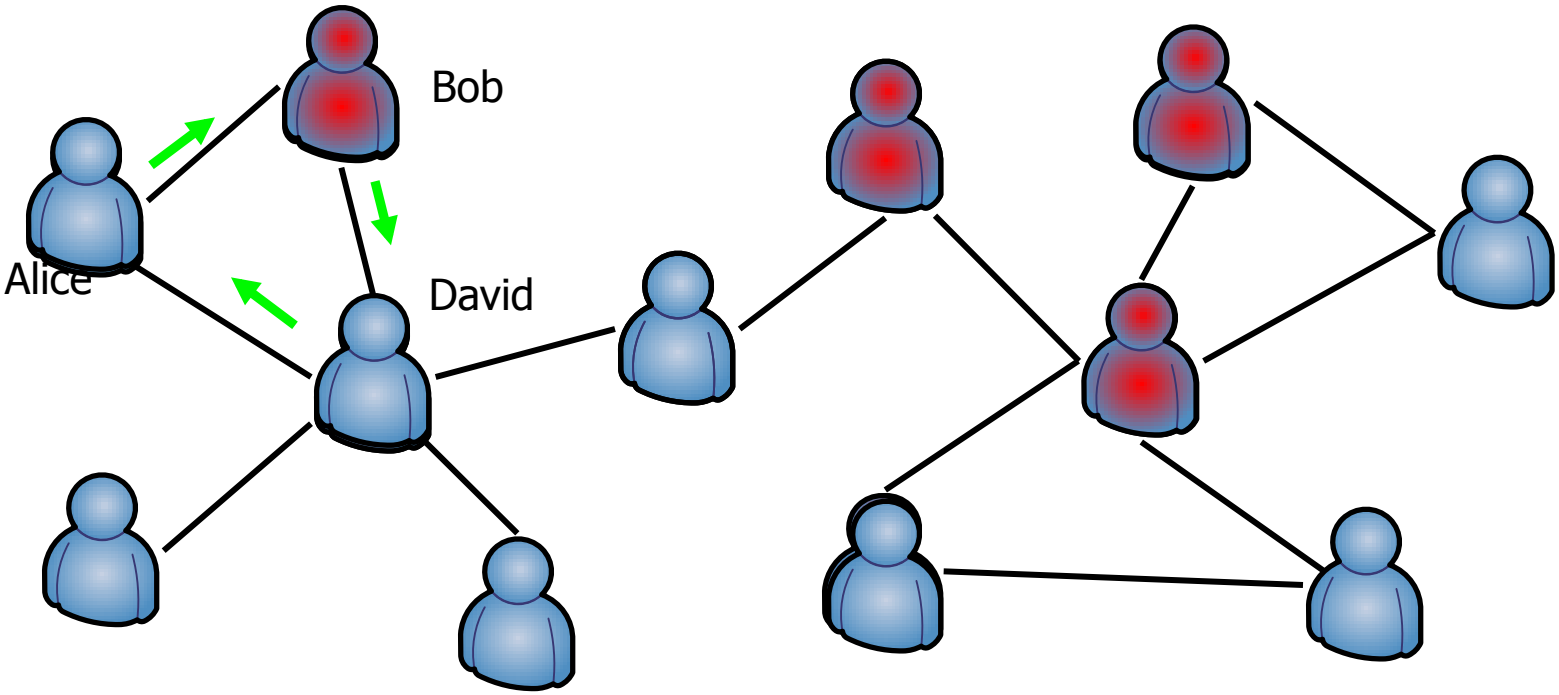
Voter Influence Model

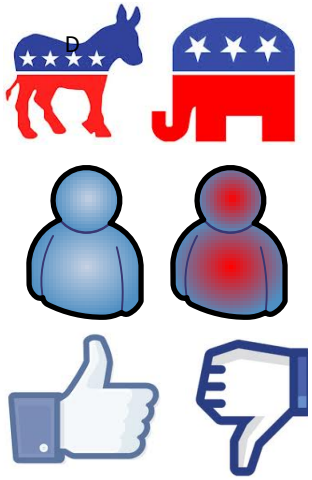
Opinion diffusions

Switch opinions back and forth

Word of mouth effect!

Randomly selecting one neighbor to adopt its opinion





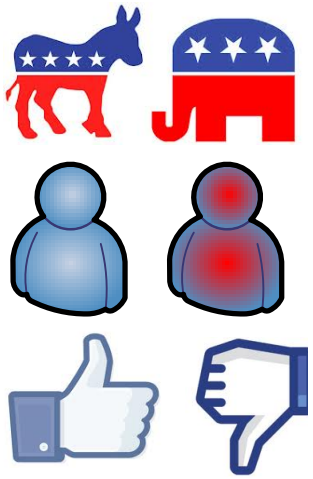
[1] P. Clifford and A. Sudbury. A model for spatial conflict. *Biometrika*, 60(3):581, 1973.

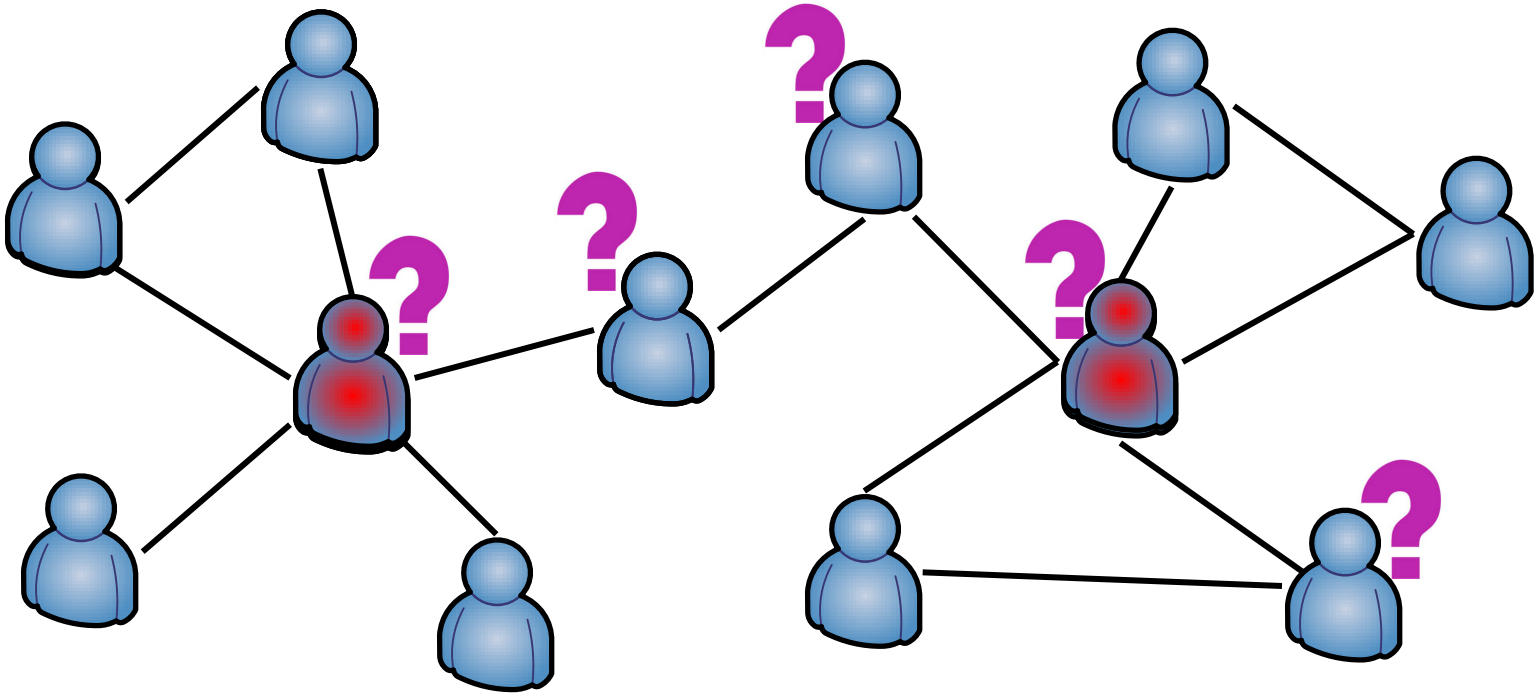
Influence Maximization

Budget: Selecting k individuals as initial red seeds

Assumption: Uniform cost of selecting each initial seed

Goal: Maximize the number of future red nodes

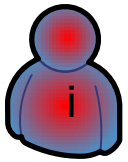


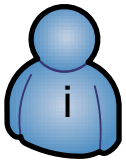


[15] E. Even-Dar and A. Shapira. A note on maximizing the spread of influence in social networks. In *WINE*, 2007.

Formulation

Probability of node *i* being red at step t: $x_t(i)$

At step t>0, $x_t(i)$ 

$1 - x_t(i)$ 

At step t+1, $$x_{t+1}(i) = \sum_{j:a_{ij}>0} x_t(j) p_{ji}$$

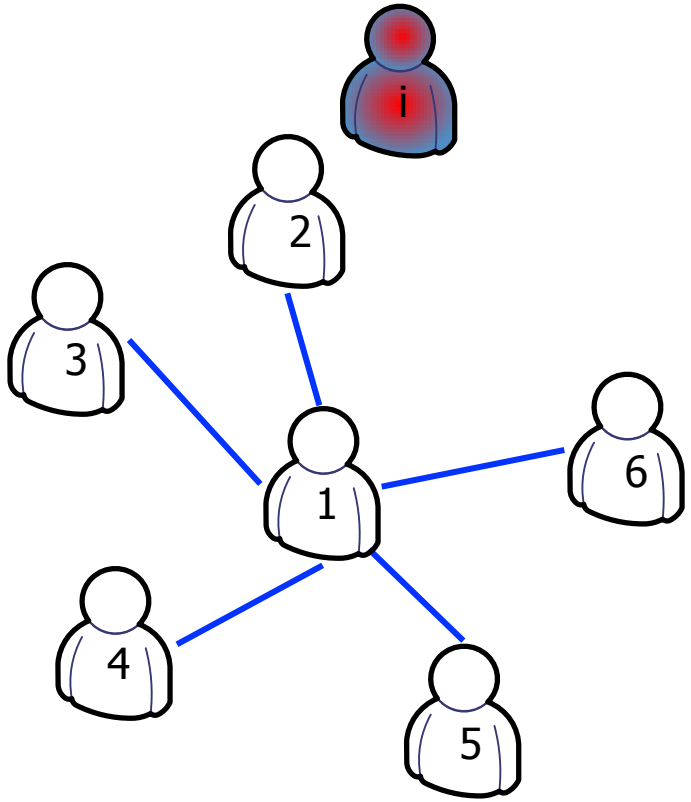
$$p_{ij} = a_{ij} / \sum_{j \in V} a_{ij}$$

Influence at step t: $$f_t(x_0) = \sum_{i \in V} x_t(i)$$

Influence contribution:

Short term $$\max_{x_0} : f_t(x_0) - f_0(x_0)$$

Long term $$\max_{x_0} : \lim_{t \to \infty} f_t(x_0) - f_0(x_0)$$



Formulation (Random Walk)

Influence at step t: $\mathbf{1}x_t^T$

Influence contribution:

Short term $\max_{x_0} : \mathbf{1}x_t^T$

Long term $\max_{x_0} : \lim_{t\to\infty} \mathbf{1}x_t^T$

x_t^T is a column vector, which is the transpose of row vector x_t

---

Matrix form: $x_t = x_0 P^t$

$$\lim_{t\to\infty} x_t = \lim_{t\to\infty} x_0 P^t = \pi$$

Influence contribution:

Short term $\max_{x_0} : f_t(x_0) - f_0(x_0)$

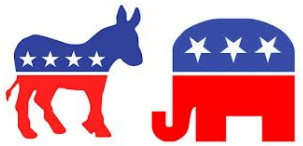
Long term $\max_{x_0} : \lim_{t\to\infty} f_t(x_0) - f_0(x_0) = x_0\pi^T - f_0(x_0)$

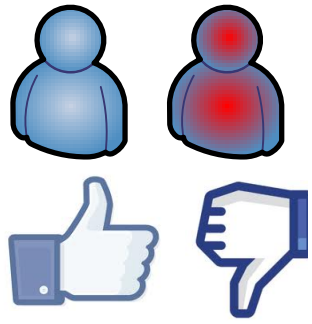
Influence Maximization

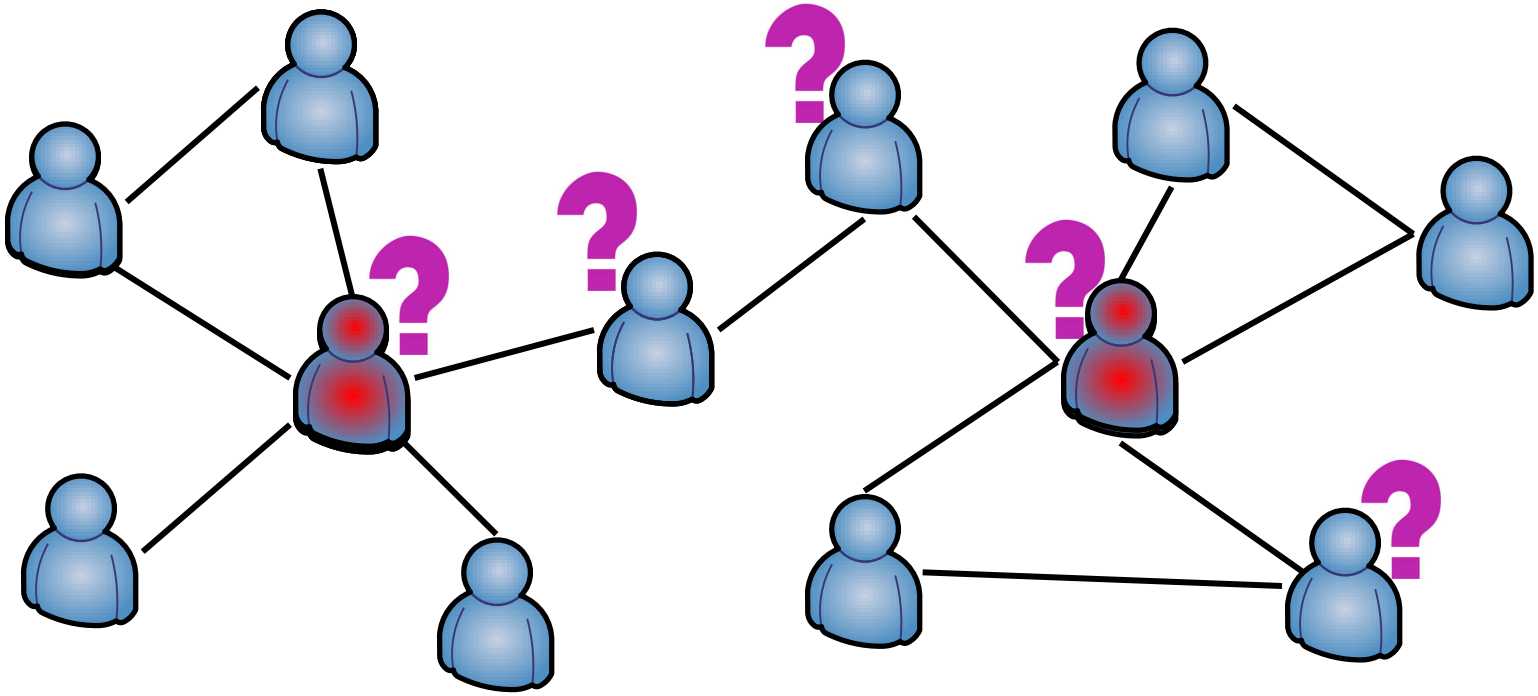
Budget: Selecting C for initial red seeds

Assumption: Heterogeneous costs of selecting different initial seeds (c_i)

Goal: Maximize the number of future red nodes







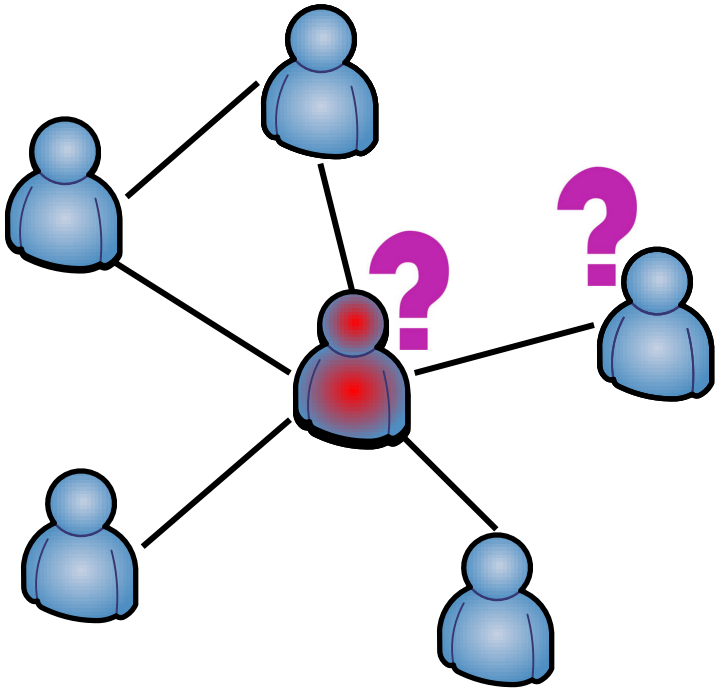
[15] E. Even-Dar and A. Shapira. A note on maximizing the spread of influence in social networks. In *WINE*, 2007.

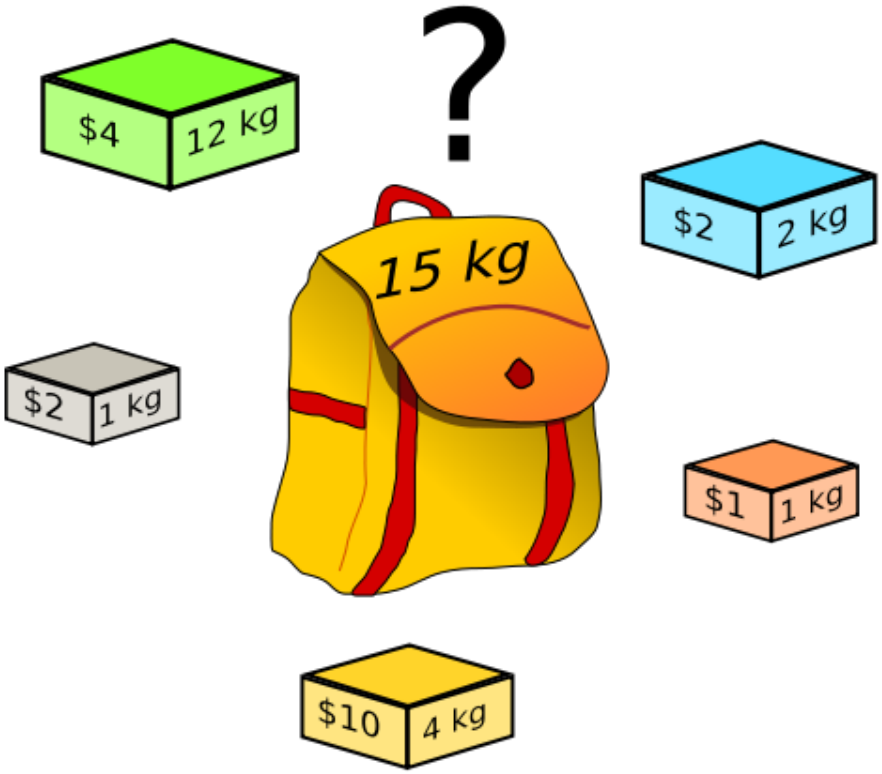
Knapsack problem

Knapsack problem

Weight = Influence Value/ Stationary distribution

Size = Cost c_i of choosing a node n_i

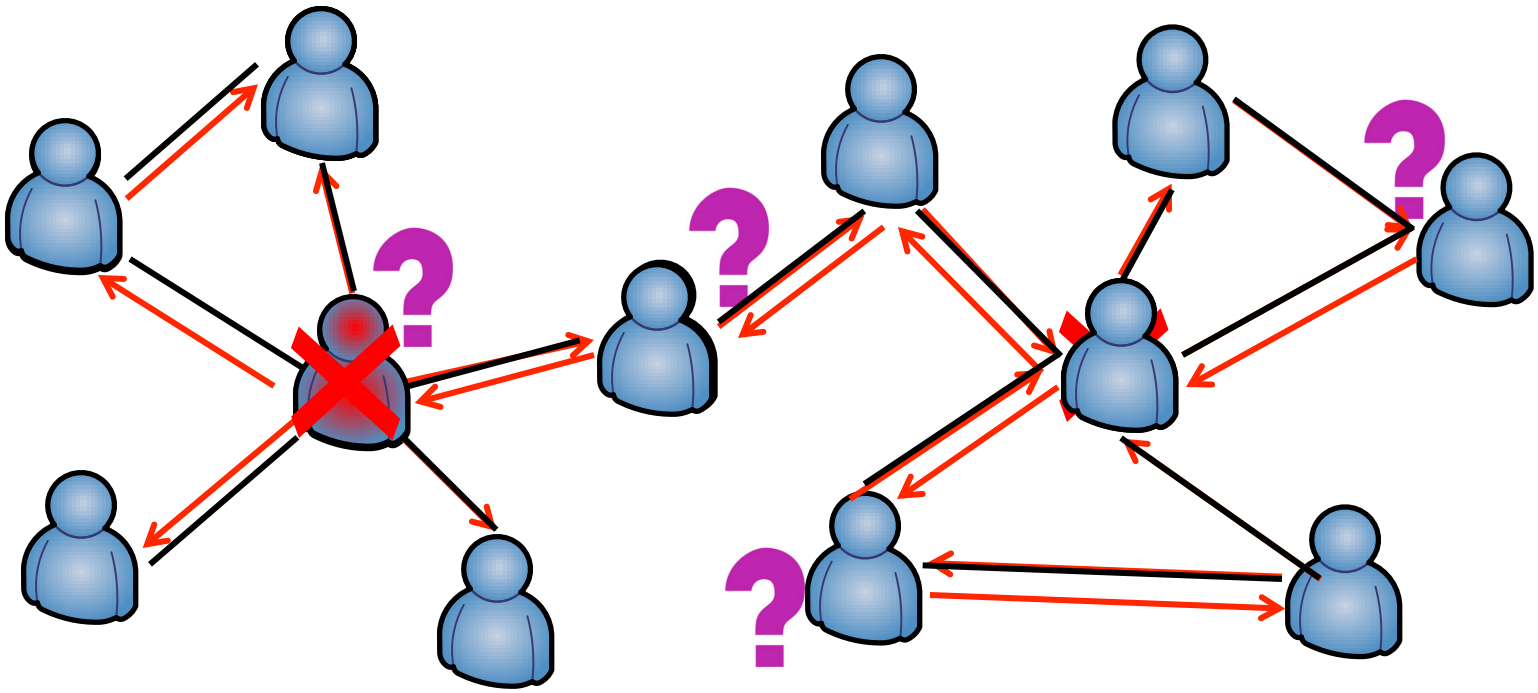




What if directed social graph

One-way connection

Randomly select an out-going neighbor

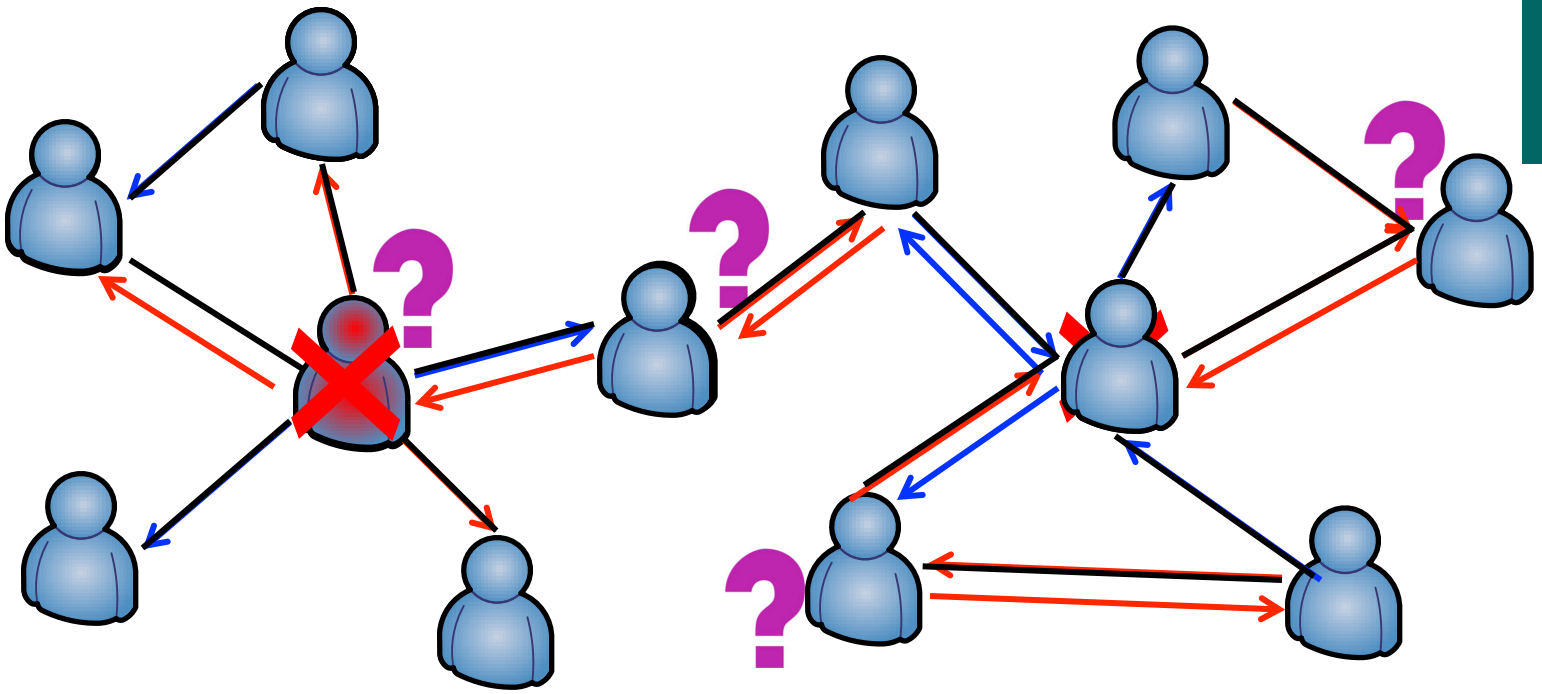


Adopt the opinion of one of the outgoing neighbors.

Directed and Signed Networks

One-way signed connection

Randomly select an out-going neighbor



Friend

Foe

Adopt the opposite opinion of foe, the same opinion of friend

[WSDM'13] Yanhua Li, Wei Chen, Yajun Wang, Zhi-Li Zhang, **Influence Diffusion Dynamics and Influence Maximization in Social Networks with Friend and Foe Relationships**. *The 6th ACM International Conference on Web Search and Data Mining*, February 4-8, 2013, Rome, Italy.

Any Comments & Critiques?

Next Class: Graph Mining (Presentation)

- Do assigned readings before class
- Submit reviews/critiques
- Attend in-class discussions