Data Mining
資料探勘

Link Analysis

Hung-Yu Kao, Fall 2019

Objectives

- To review common approaches to link analysis
- To calculate the popularity of a site based on link analysis
- To model human judgments indirectly

Outline

1. Motivation
2. Early Approaches to Link Analysis
3. Hubs and Authorities: HITS
4. Page Rank
5. Other issues and Limitation of Link Analysis
6. Links in a social network

Motivation

- Human knowledge is real, convincing and trustable information
  - *E.g., classification by human in yahoo*
- Hyperlinks contain information about the *human judgment*
- Social sciences
  - Nodes: persons, organizations
  - Edges: social interaction
- Easy job ? *Counting in-links for popularity*

An example: scientific literature

- **Impact factor**
  (*http://scientific.thomson.com/free/essays/journalcitationreports/impactfactor/*)
  - for journal evaluation
  - *Garfield (Science 1955, 1972)*
  - The average number of citations per recently published item
  - **C / N**
    - **C**: the total number of citations in a given time interval [t, t + t1] to articles published by a given journal during [t – t2, t]
    - **N**: the total number of articles published by that journal in [t – t2, t]

Mark	Rank	Abbreviated Journal Title *(linked to journal information)*	ISSN	JCR Data Total Cites	Impact Factor	5-Year Impact Factor	Im
	1	NAT REV MOL CELL BIO	1471-0072	29222	39.123	42.508	
	2	CELL	0092-8674	171297	32.403	34.774	
	3	CANCER CELL	1535-6108	19726	26.566	28.174	
	4	CELL STEM CELL	1934-5909	10145	25.421	27.494	
	5	NAT MED	1078-8956	54228	22.462	26.418	
	6	NAT CELL BIOL	1465-7392	29959	19.488	20.116	
	7	ANNU REV CELL DEV BI	1081-0706	8399	15.836	19.733	
	8	MOL CELL	1097-2765	44493	14.178	14.202	
	9	DEV CELL	1534-5807	18481	14.030	14.202	
	10	CELL METAB	1550-4131	9907	13.668	17.770	
	11	CURR OPIN CELL BIOL	0955-0674	13795	12.897	12.594	
	[illegible]	NAT STRUCT MOL BIOL	1545-9993	22401	12.712	12.114	

- Issues
  - The number of citation base
  - Normalization?

ISI impact factor: http://isiknowledge.com/

Early Approaches

Basic Assumptions

- Hyperlinks contain information about the human judgment of a site
- The more incoming links to a site, the more it is judged ***important***

Bray 1996 *(Measuring the Web, WWW)*

- The **visibility** of a site is measured by the number of other sites pointing to it (indegree)
- The **luminosity** of a site is measured by the number of other sites to which it points (outdegree)

→ ***Limitation***: failure to capture the relative importance of different parents (children) sites

→ *But works in some recent reports !*

Early Approaches

Mark *(Commun ACM, 1988)*

- To calculate the score S of a document at vertex v

$$S(v) = s(v) + \frac{1}{|ch[v]|} \sum_{w \in |ch(v)|} S(w)$$

v: a vertex in the hypertext graph G = (V, E)
S(v): the global score
s(v): the score if the document is isolated
ch(v): children of the document at vertex v

- Limitation:
  - Require G to be a directed acyclic graph (DAG)
  - If v has a single link to w, S(v) > S(w)
  - If v has a long path to w and s(v) < s(w), then S(v) > S(w)

  → ***Unreasonable***, *users need go through the long path from the irrelevant document (v) to reach the important document (w)*

→ *But show the message passing schemes*

Early Approaches

Marchiori *(WWW, 1997)*

- ***Hyper information*** should complement textual information to obtain the overall information

  $S(v) = s(v) + h(v)$

  - S(v): overall information
  - s(v): textual information
  - h(v): hyper information

  Can't handle real world cases → a cyclic graph

- $h(v) = \sum_{w \in |ch[v]|} F^{r(v,w)} S(w)$
  - **F**: a fading constant, F Є (0, 1)
  - **r(v, w)**: the rank of w after sorting the children of v by S(w)

→ a remedy of the previous approach (Mark 1988)

HITS - Kleinberg's Algorithm

- HITS – **H**ypertext **I**nduced **T**opic **S**election
- For each vertex v Є V in a subgraph of interest:

  a(v) - the authority of v
  h(v) - the hubness of v

- A site is very authoritative if it receives many citations. *Citation from important sites weight more than citations from less-important sites*
- Hubness shows the importance of a site. A good hub is a site that links to many authoritative sites

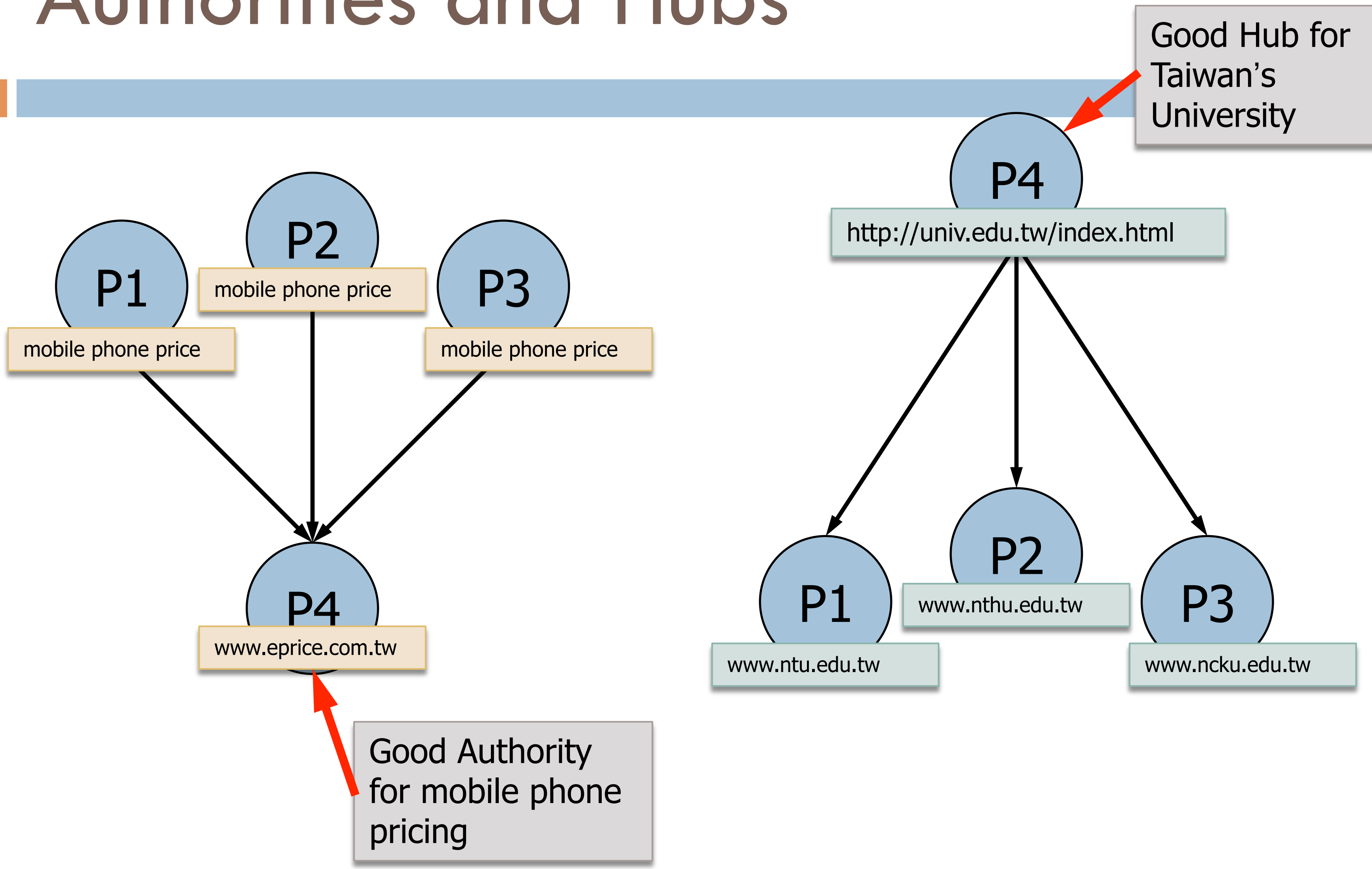
雞生蛋， 蛋生雞?

Twin relation v.s. triple relation or more

Motivation

- For a given query, which pages are the answer set?
  - Results of search engines
    - Rank manually
    - Rank by similarity
    - Rank by hit rate *(need usage log)*
    - Rank by link analysis (HITS, PageRank,…)
  - Relevant v.s. Authoritative
    - Intra-page v.s. inter-page
  - *Users need authoritative pages among relevant pages.*

Authorities and Hubs

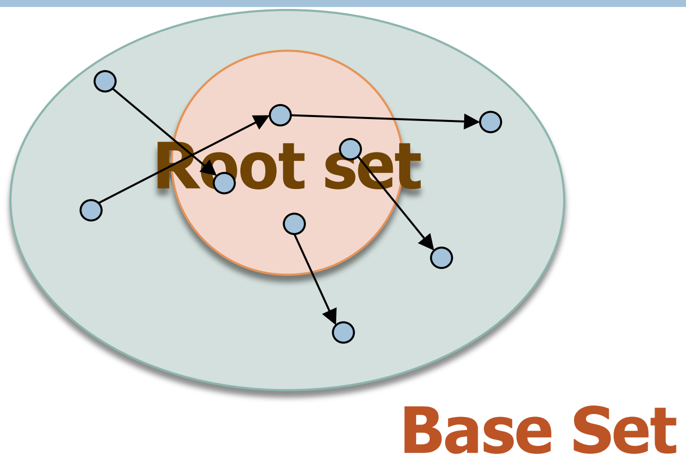


Introduction

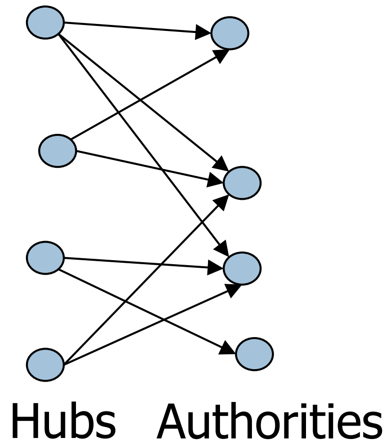
- How to find authoritative pages for queries
  - Step I: rank pages according to their **in-degree** in the **sub-graph** induced by the **root set S**
    - root set: top k pages indexed by search engines
    - Problems
      - very few edges, a large fraction of the nodes will be isolated
      - real authoritative pages are not included in the root set

Introduction

- Step II: **extend** the root set to **base set**
  - Problems
    - Unrelated page of large in-degree



- New approach (kleinberg '97)
  - There should also be considerable overlap in the sets of pages that point to authoritative pages.
    - Hub pages
    - *mutually reinforcing relationship*



Authority and Hubness Convergence

- Recursive dependency:

$$a(v) \leftarrow \Sigma_{w \in pa[v]} h(w)$$

$$h(v) \leftarrow \Sigma_{w \in ch[v]} a(w)$$

- Using Linear Algebra, we can prove:

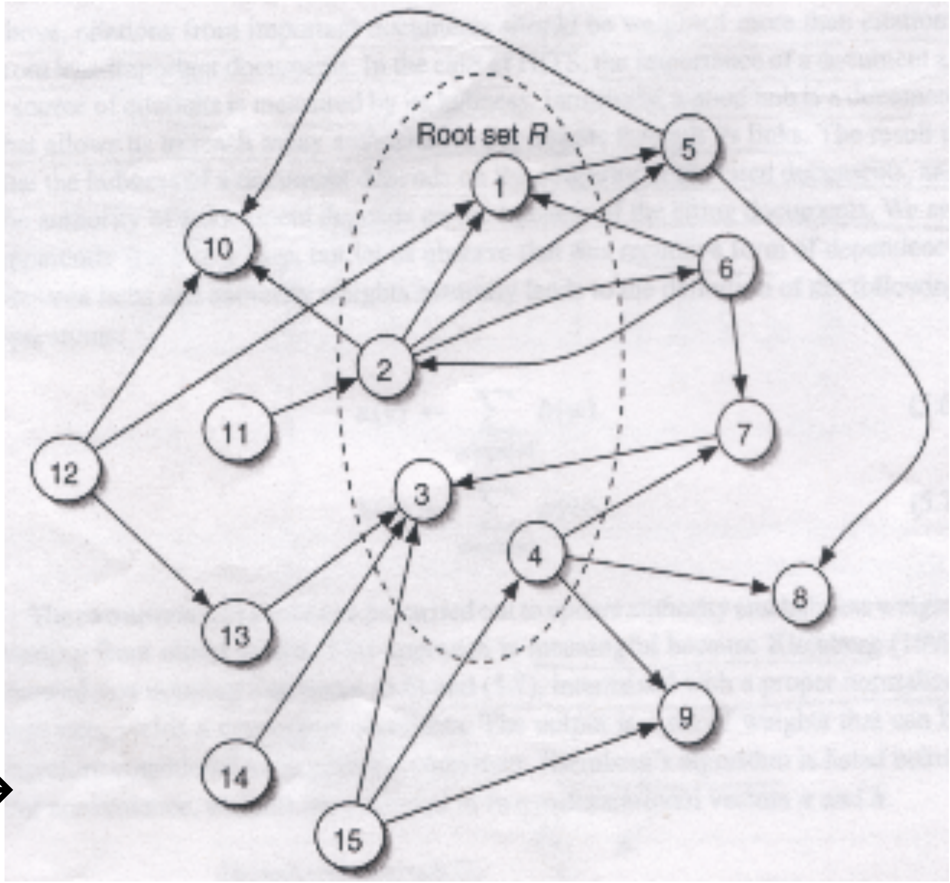
a(v) and h(v) converge

HITS Example

Find a base subgraph:

- Start with a root set R {1, 2, 3, 4}
- {1, 2, 3, 4} - nodes relevant to the topic
- Expand the root set R to include all the children and a fixed number of parents of nodes in R
  - *Indegree v.s. outdegree*

→ A new set S (base subgraph) →



HITS Example


```
BaseSubgraph( R, d)
      S ← r
      for each v in R
      do S ← S U ch[v]
            P ← pa[v]
            if |P| > d
            then P ← arbitrary subset of P having size d
                  S ← S U P
      return S
```

HITS Example

Hubs and authorities: two n-dimensional $\vec{\mathbf{a}}$ and $\vec{\mathbf{h}}$

HubsAuthorities(G)

$\mathbf{1} \leftarrow [1,\ldots,1] \in R^{|V|}$

$a_0 \leftarrow h_0 \leftarrow \mathbf{1}$

$t \leftarrow 1$

repeat

  for each v in V

    do $a_t(v) \leftarrow \sum_{w \in pa[v]} h_{t-1}(w)$

    $h_t(v) \leftarrow \sum_{w \in ch[v]} a_{t-1}(w)$

  $a_t \leftarrow a_t / \|a_t\|$

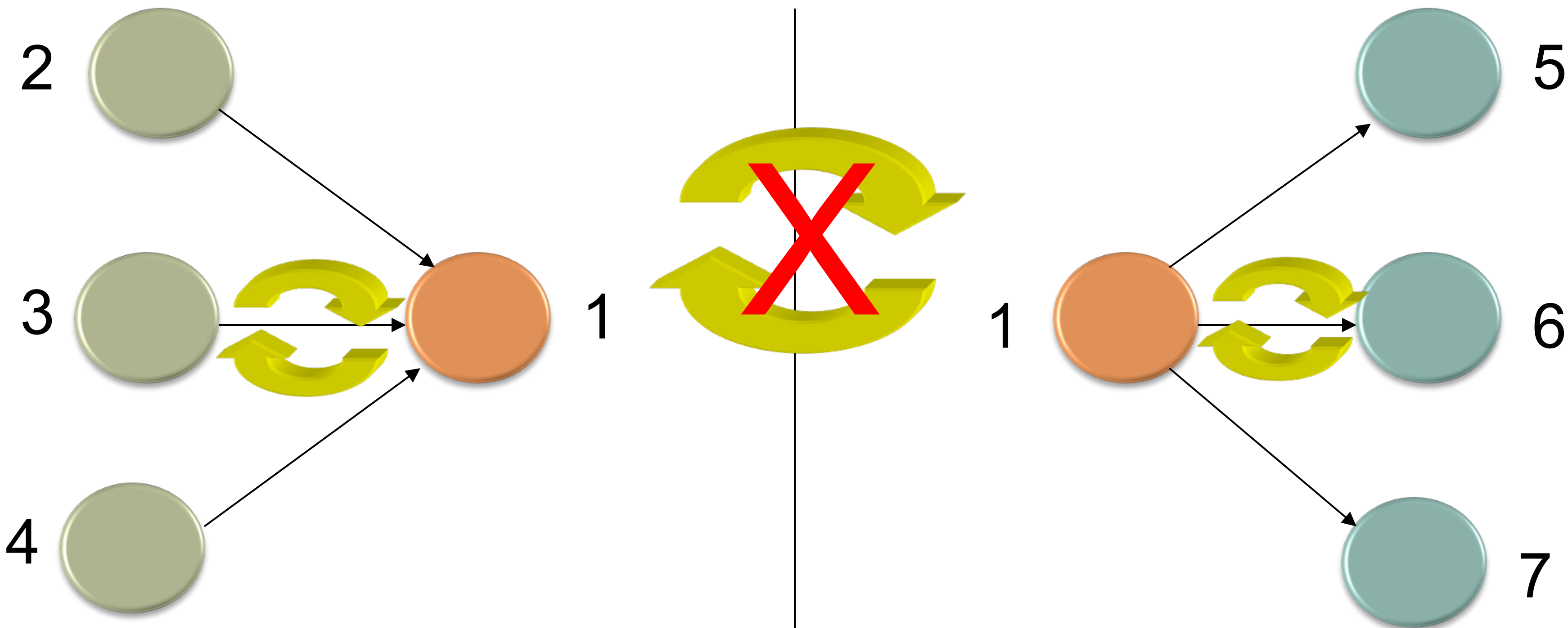
  $h_t \leftarrow h_t / \|h_t\|$ *normalization*

  $t \leftarrow t + 1$

until $\|a_t - a_{t-1}\| + \|h_t - h_{t-1}\| < \varepsilon$

return (a_t, h_t)

Authority and Hubness



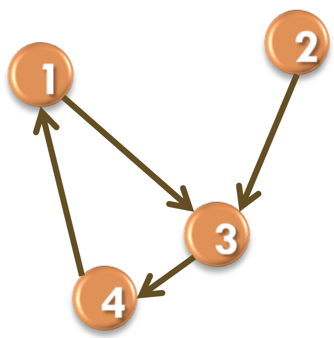
$a(1) = h(2) + h(3) + h(4)$

$h(1) = a(5) + a(6) + a(7)$

Basic Link Analysis

- Let A denote the **adjacency matrix** of the graph, $\boldsymbol{a}_t \leftarrow A^t \boldsymbol{h}_{t-1}$, $\boldsymbol{h}_t \leftarrow A\, \boldsymbol{a}_{t-1}$
  - $\boldsymbol{a}_n$ is the unit vector in the direction of $(A^tA)^{n-1}A^tz$
  - $\boldsymbol{h}_n$ is the unit vector in the direction of $(AA^t)^nz$
- $\boldsymbol{a}^*$ is the principal eigenvector of A^tA, and $\boldsymbol{h}^*$ is the principal eigenvector of AA^t

Adjacency matrix



$$A = \begin{bmatrix} 0010 \\ 0010 \\ 0001 \\ 1000 \end{bmatrix}$$

$$A^t = \begin{bmatrix} 0001 \\ 0000 \\ 1100 \\ 0010 \end{bmatrix}$$

In-Out

$$A^t A = \begin{bmatrix} 1000 \\ 0000 \\ 0020 \\ 0001 \end{bmatrix}$$

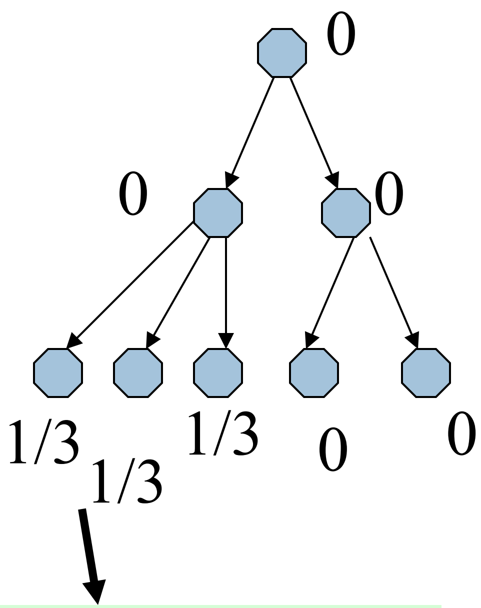
Out-In

$$AA^t = \begin{bmatrix} 1100 \\ 1100 \\ 0010 \\ 0001 \end{bmatrix}$$

$$AA = \begin{bmatrix} 0001 \\ 0001 \\ 1000 \\ 0010 \end{bmatrix}$$

Example (1-norm normalization)

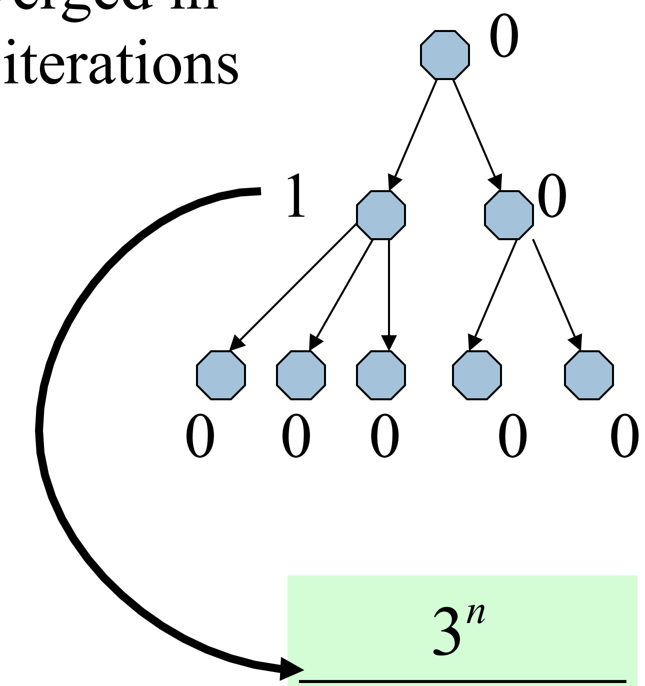
Authority



$$\frac{3^n}{3*3^n+2*2^n+2*2^n}$$

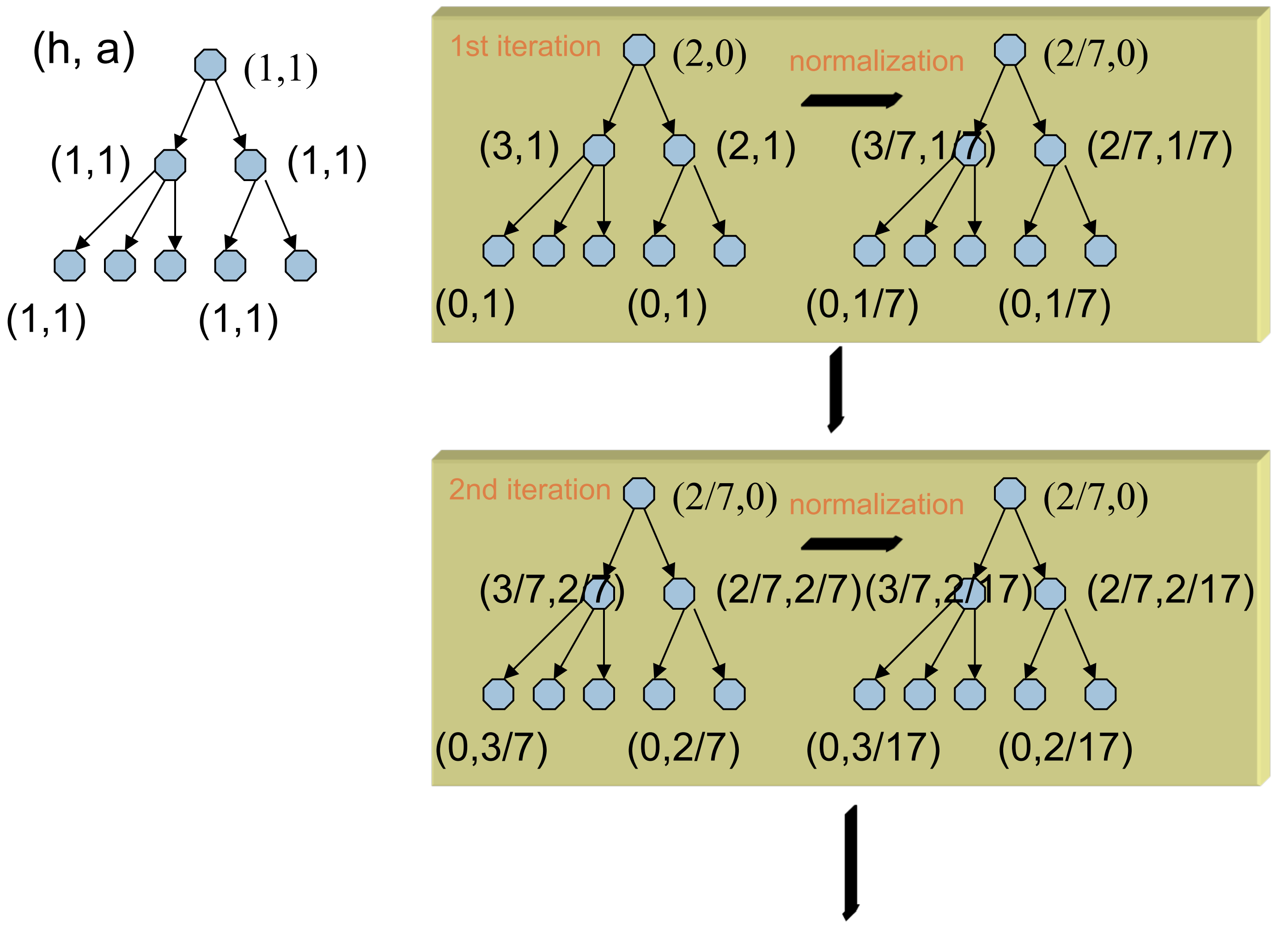
Converged in 72‘th iterations

Hub

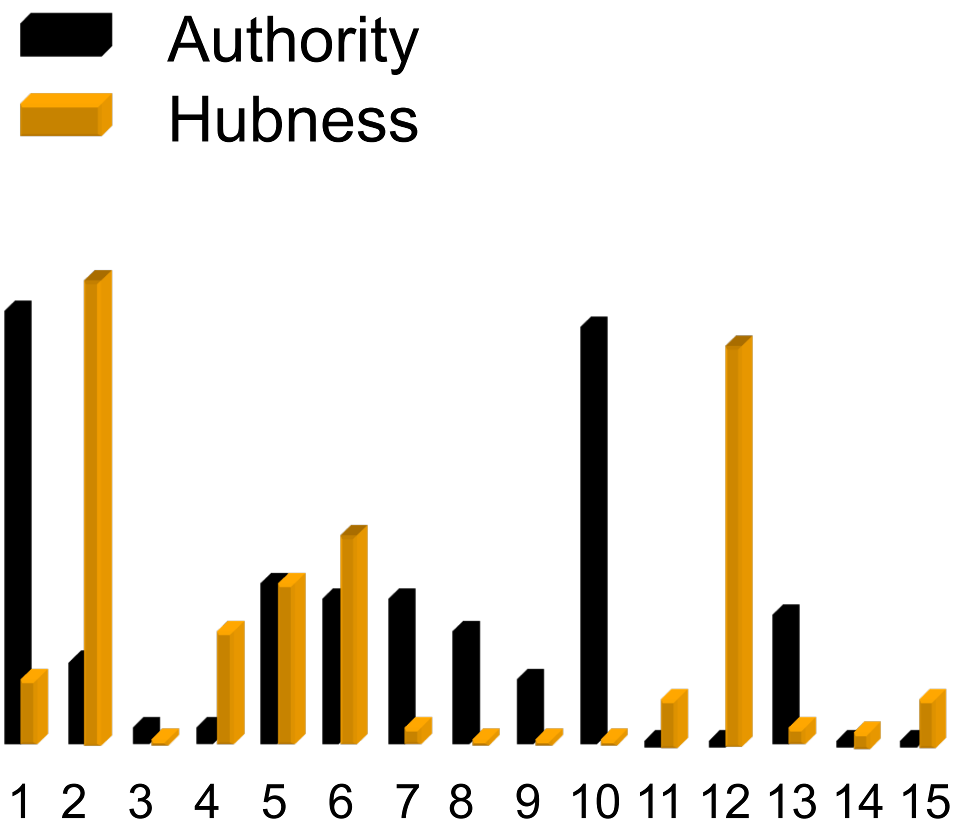


$$\frac{3^n}{3^n+2^n+2^n}$$

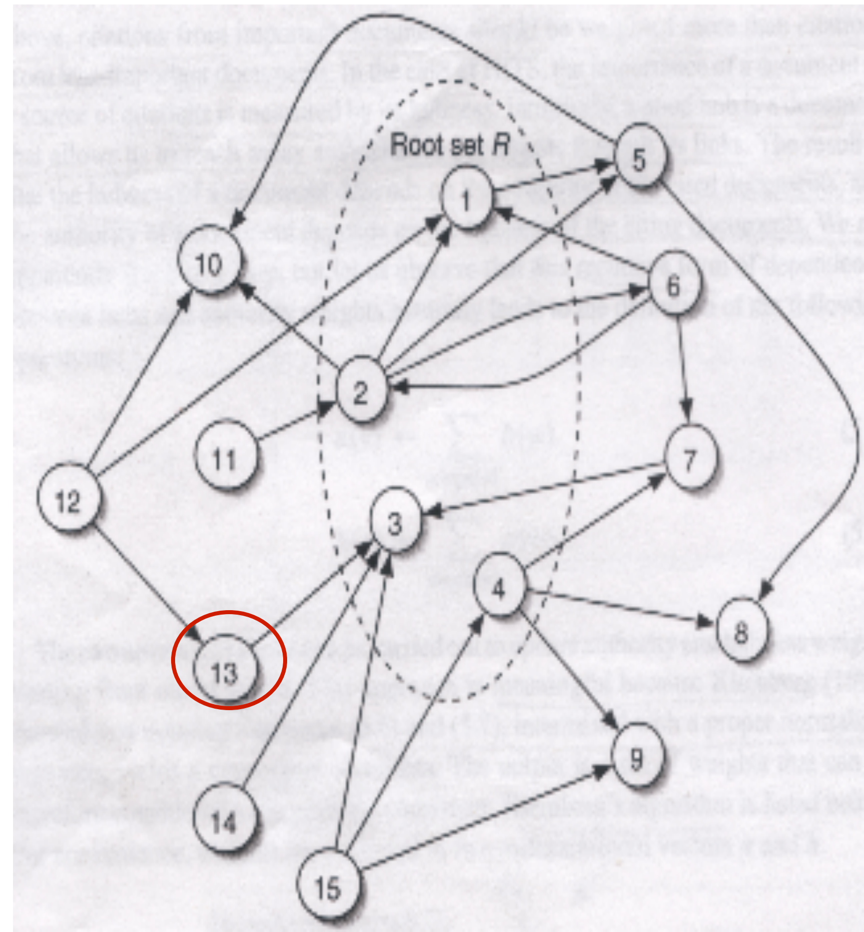
Example (1-norm normalization)



HITS Example Results



Authority and hubness weights

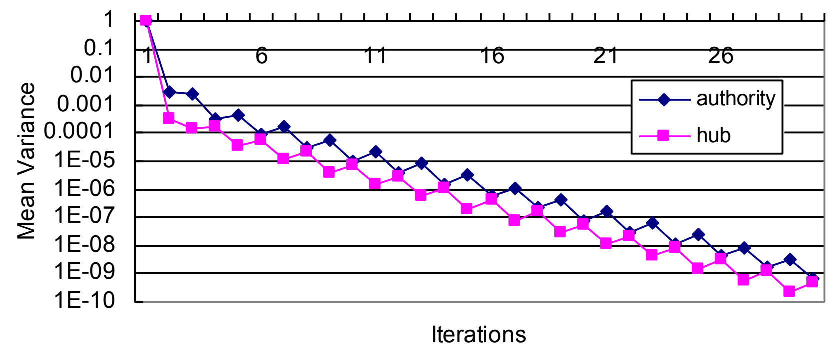


Issues for HITS

- Mutually reinforcing relationships between hosts
  - Nepotistic links cancellation
    - Nepotistic links: links between pages that are present for reasons other than merit
      - Menu links
      - Link-based spam
  - Link normalization

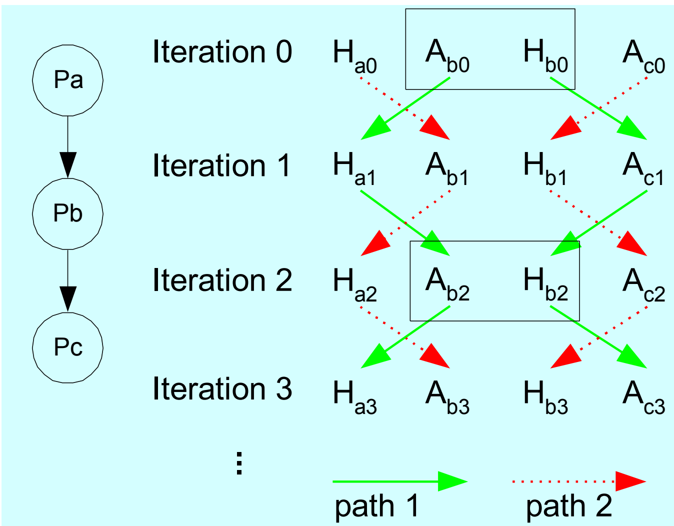
One important observation

- The process of link analysis
  - Convergence of values of hubs and authorities
  - Two (hub, authority) pairs



$$\{(A_{a3},H_{a3}),(A_{b2},H_{b2}),(A_{c3},H_{c3})\}$$

$$\{(A_{a2},H_{a2}),(A_{b3},H_{b3}),(A_{c2},H_{c2})\}$$



HITS Improvements

Bharat and Henzinger *(1998, SIGIR, 1068 citation counts)*

-- Improved algorithms for topic distillation in a hyperlinked environment

- HITS problems
  1) The document can contain many ***identical links*** to the same document in another host (投票部隊)
  2) Links are generated automatically (e.g. messages posted on newsgroups)
     - *Containing human's opinion ?*
  3) Non-relevant Nodes
     - *Topic drift*

Solutions – *Combining Connectivity and Content Analysis*

- Assign weight to *identical* multiple edges, which are inversely proportional to their multiplicity
- Prune irrelevant nodes or regulating the influence of a node with a ***relevance weight***

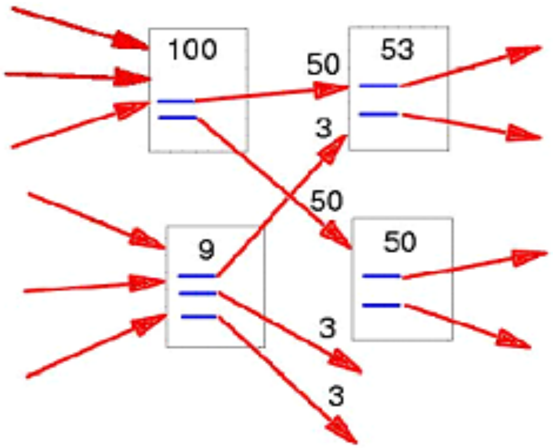
$$similarity(Q, D_j) = \frac{\sum_{i=1}^{t}(w_{iq} \times w_{ij})}{\sqrt{\sum_{i=1}^{t}(w_{iq})^2 \times \sum_{i=1}^{t}(w_{ij})^2}}$$

where
$w_{iq} = freq_{iq} \times IDF_i$,
$w_{ij} = freq_{ij} \times IDF_i$,
$freq_{iq}$ = the frequency of the term i in query Q,
$freq_{ij}$ = the frequency of the term i in document D_j,
IDF_i = an estimate of the inverse document frequency of term i on the World Wide Web.

PageRank

- Introduced by Page et al (*1998, WWW*)
  - The weight is assigned by the rank of parents

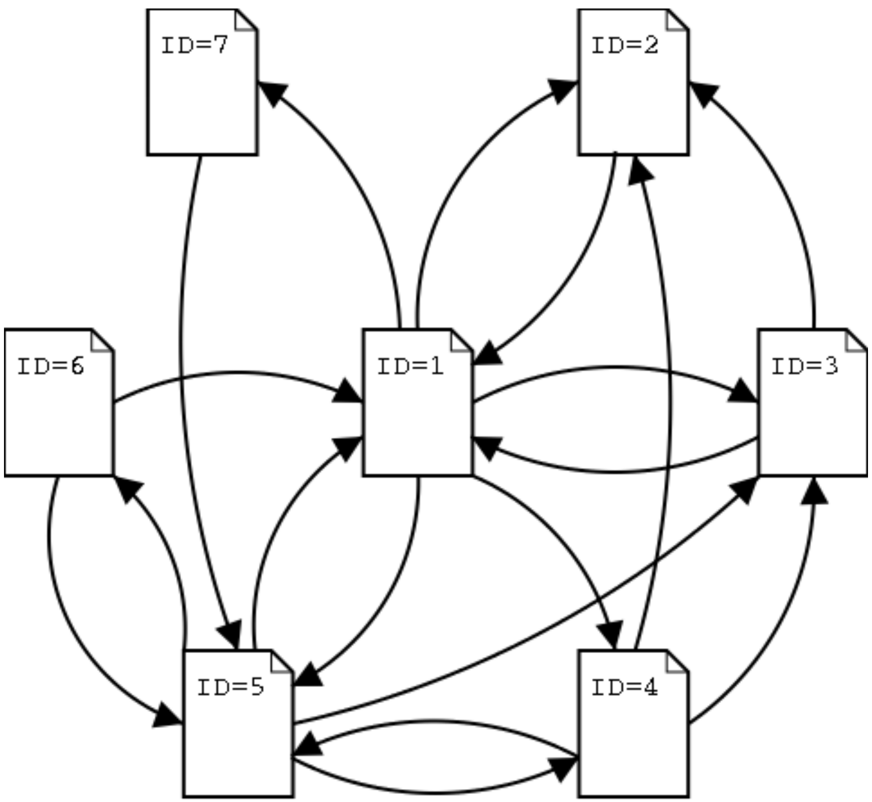
$$r(v) = \alpha \sum_{w \in pa[v]} \frac{r(w)}{|ch[w]|},$$



- Difference with HITS
  - HITS takes Hubness & Authority weights
  - The page rank is proportional to its parents' rank, but inversely proportional to its parents' outdegree
  - Query independent

[Google's Pagerank]

Matrix Notation



Page ID	OutLinks
1	2,3,4,5,7
2	1
3	1,2
4	2,3,5
5	1,3,4,6
6	1,5
7	5

Adjacent Matrix

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

Calculator: http://www.webworkshop.net/pagerank_calculator.php3

Matrix Notation

- Matrix Notation

  $\mathbf{r} = \alpha\,\mathbf{B}\,\mathbf{r} = \mathbf{M}\,\mathbf{r}$

  α : eigenvalue

  r : eigenvector of B

$$b_{uv} = \begin{cases} \dfrac{a_{uv}}{\sum_w a_{uw}} & \text{if } ch[u] \neq 0, \\ a_{uv} = 0 & \text{otherwise} \end{cases}$$

$$A\,x = \lambda\,x$$

$$|\,A - \lambda I\,|\;x = 0$$

$$B = \begin{pmatrix} 0 & 1/5 & 1/5 & 1/5 & 1/5 & 0 & 1/5 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 & 0 & 0 \\ 1/4 & 0 & 1/4 & 1/4 & 0 & 1/4 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Finding Pagerank
→ to find eigenvector of B with an associated eigenvalue α

Matrix Notation

PageRank: eigenvector of **P** relative to max eigenvalue

$$\mathbf{B} = \mathbf{P}\ \mathbf{D}\ \mathbf{P}^{-1}$$

$$\begin{pmatrix} \lambda_1 & & O \\ & \lambda_2 & \\ & & \ddots & \\ O & & & \lambda_n \end{pmatrix}$$

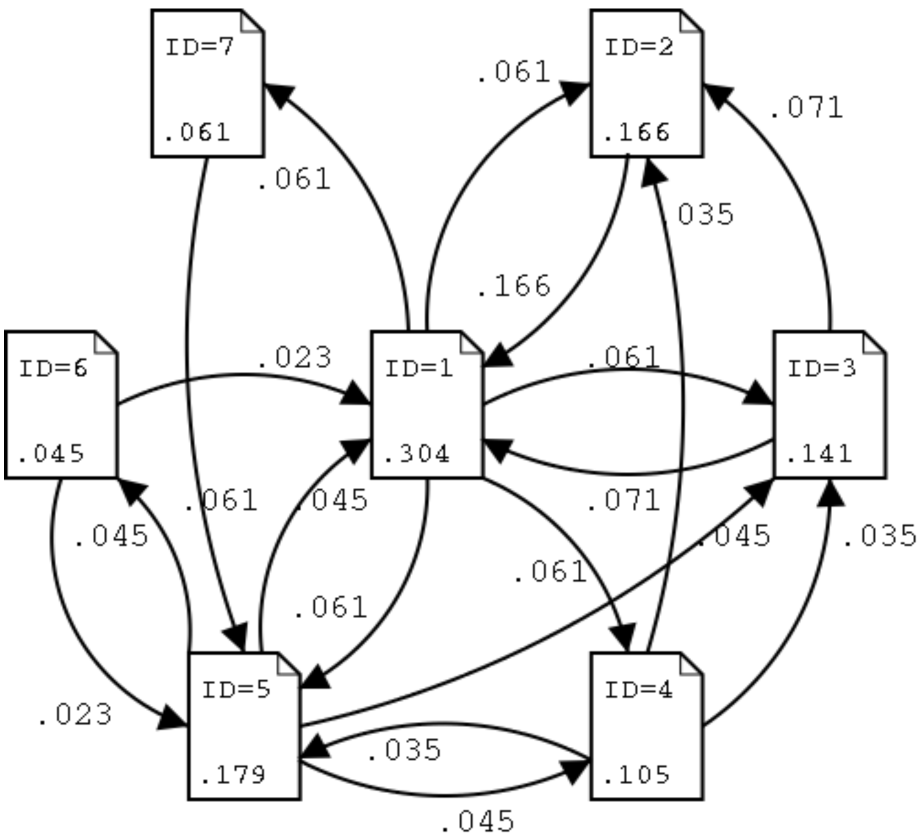
D: diagonal matrix of eigenvalues $\{\lambda_1, \ldots \lambda_n\}$

P: regular matrix that consists of eigenvectors

$$\begin{pmatrix} \mathbf{r}_1 & \mathbf{r}_2 & \cdots & \mathbf{r}_n \end{pmatrix}$$

$$\text{PageRank } \mathbf{r}_1 = \begin{pmatrix} 0.69946 \\ 0.38286 \\ 0.32396 \\ 0.24297 \\ 0.41231 \\ 0.10308 \\ 0.13989 \end{pmatrix} \xrightarrow{\text{normalized}} \begin{pmatrix} 0.303514 \\ 0.166134 \\ 0.140575 \\ 0.105431 \\ 0.178914 \\ 0.044728 \\ 0.060703 \end{pmatrix}$$

Matrix Notation



PR	ID	OutLink	InLink
0.304	1	2,3,4,5,7	2,3,5,6
0.179	5	1,3,4,6	1,4,6,7
0.166	2	1	1,3,4
0.141	3	1,2	1,4,5
0.105	4	2,3,5	1,5
0.061	7	5	1
0.045	6	1,5	5

- Confirm the result

  # of inlinks from high ranked page

  hard to explain about 5&2, 6&7

- Interesting Topic

  * *How do you create your homepage highly ranked / lowly ranked?*

  * *How to detect it ?*

Markov Chain Notation

- Random surfer model
  - Description of a **random walk** through the Web graph
  - Interpreted as a transition matrix with asymptotic probability that a surfer is currently browsing that page

$$r_t(v) = P(S_t = v) = \sum_w P(S_t = v \mid S_{t-1} = w) P(S_{t-1} = w)$$
$$= \sum_w m_{wv} r_{t-1}(w).$$

$$\boldsymbol{r}_t = \mathbf{M}\,\boldsymbol{r}_{t-1}$$

M: transition matrix for a first-order Markov chain (stochastic)

Does it converge to some sensible solution (as $t \to \infty$) **regardless of the initial ranks** *(equal or non-equal)* ?

Problem

- "Rank Sink" Problem
  - never pass the rank to others
  - In general, many Web pages have no inlinks / outlinks
  - It results in dangling edges in the graph

E.g.

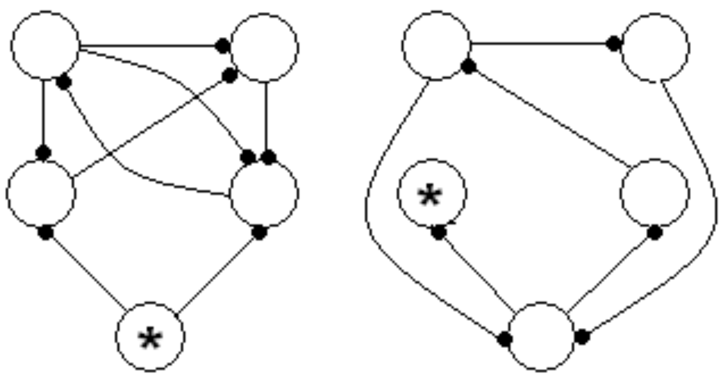
no parent → rank 0

M^T converges to a matrix

whose last column is all zero

no children → no solution

M^T converges to zero matrix



Modification

- Surfer will restart browsing by *picking a new Web page at random*

  $$\mathbf{M} = (\mathbf{B} + \mathbf{E})$$

  E : escape matrix

  M : stochastic matrix

  $$e_{vw} = \begin{cases} 0 & \text{if } |ch[v]| > 0 \\ \frac{1}{n} & \text{otherwise} \end{cases}$$

- Problem still exists?
  - It is not guaranteed that **M** is primitive
  - If **M** is stochastic and primitive, PageRank converges to corresponding stationary distribution of **M**

PageRank Algorithm

PageRank$(\boldsymbol{M}, n, \epsilon)$

$\mathbf{1} \leftarrow [1, \ldots, 1] \in \mathbb{R}^n$

$z \leftarrow \frac{1}{n}\mathbf{1}$

$\boldsymbol{x}_0 \leftarrow z$

$t \leftarrow 0$

repeat

  $t \leftarrow t + 1$

  $\boldsymbol{x}_t \leftarrow \boldsymbol{M}^{\mathrm{T}} \boldsymbol{x}_{t-1}$

  $d_t \leftarrow \|\boldsymbol{x}_{t-1}\|_1 - \|\boldsymbol{x}_t\|_1$

  $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_1 + d_t z$

  $\delta \leftarrow \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_t\|_1$

 until $\delta < \epsilon$

return $\boldsymbol{x}_t$

dt is the total rank being lost in sinks

Normalization

* Page et al, 1998

Quick reference

$$PR(P_i) = \frac{(d)}{n} + (1-d) \times \sum_{l_{j,i} \in E} PR(P_j) / \mathrm{Outdegree}(P_j)$$

D(damping factor)=0.1~0.15
n=|page set|

Stability

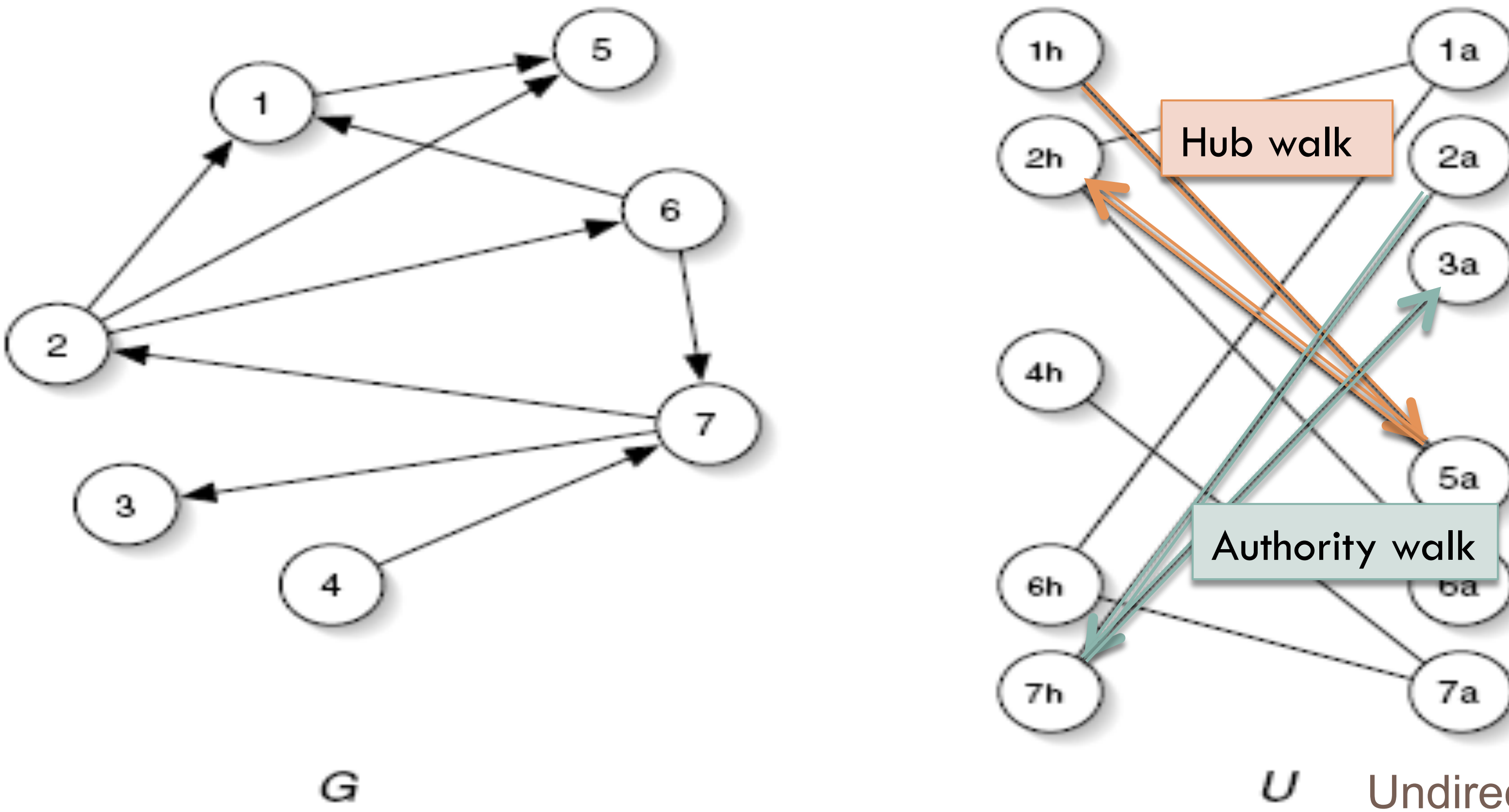
- Whether the link analysis algorithms based on eigenvectors are stable in the sense that results don't change significantly?
- The **connectivity** of a portion of the graph is **changed** arbitrary
  - How will it affect the results of algorithms?

Ng et al (***2001, SIGIR***) – "stable algorithms for link analysis"

SALSA

- SALSA *(Lempel, Moran 2001, ACM TOIS)*
  - Probabilistic extension of the HITS algorithm
  - Random walk is carried out by following hyperlinks both in the forward and in the backward direction
- Two **separate** random walks
  - Hub walk
  - Authority walk

Forming a Bipartite Graph in SALSA



Undirected / bidirected

Figure 5.6 Forming a bipartite graph in SALSA.

Random Walks

- Hub walk
  - Follow a Web link from a page u_h to a page w_a (a **forward** link) and then
  - Immediately traverse a **backlink** going from w_a to v_h, where $(u,w) \in E$ and $(v,w) \in E$
- Authority Walk
  - Follow a Web link from a page w_a to a page u_h (a **backward** link) and then
  - Immediately traverse a **forward** link going back from u_h to x_a where $(u,w) \in E$ and $(u,x) \in E$

Computing Weights

- Hub weight computed from the sum of the product of the inverse degree of the in-links and the out-links

$$\tilde{h}_{uv} = \sum_{\substack{w:(u,w)\in E,\\ (v,w)\in E}} \frac{1}{\deg(u_{\mathrm{h}})} \frac{1}{\deg(w_{\mathrm{a}})},$$

$$\tilde{t}_{uv} = \sum_{\substack{w:(w,u)\in E,\\ (w,v)\in E}} \frac{1}{\deg(v_{\mathrm{a}})} \frac{1}{\deg(w_{\mathrm{h}})}.$$

Why We Care

- Lempel and Moran (2001) showed theoretically that SALSA weights are more robust that HITS weights in the presence of the **Tightly Knit Community** (TKC) Effect.
  - This effect occurs when a small collection of pages (related to a given topic) is connected so that ***every hub links to every authority*** and includes as a special case the mutual reinforcement effect
  - ***highly ranked*** by HITS

- TKC could be exploited by **spammers** hoping to increase their page weight (e.g. link farms)

A Similar Approach

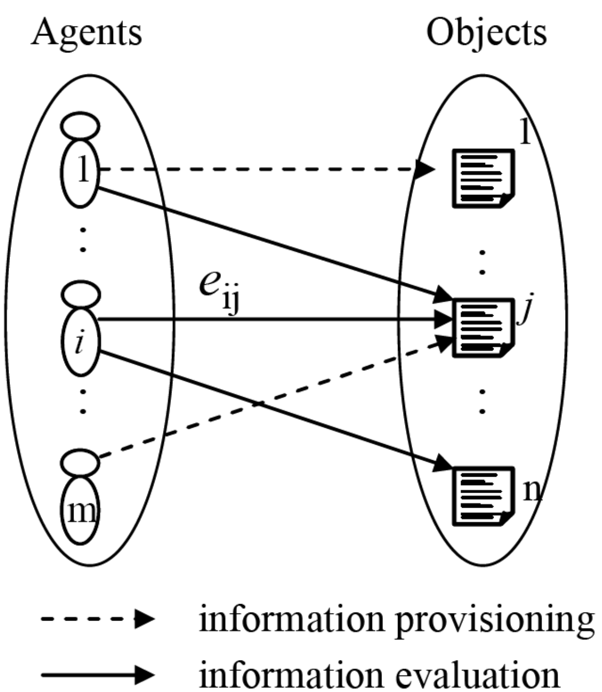
- Rafiei and Mendelzon (*2000, WWW*) and Ng *et al.* (2001) propose similar approaches using *reset* as in PageRank
  - Unlike PageRank, in this model the surfer will follow a forward link on odd steps but a backward link on even steps

- The stability properties of these ranking distributions are similar to those of PageRank (Ng *et al.* 2001)
- Borodin, *2001 WWW*

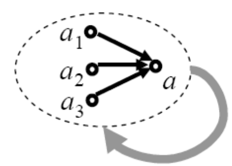
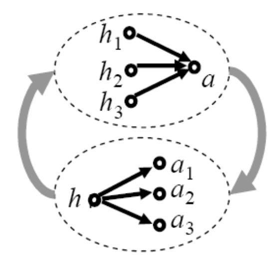
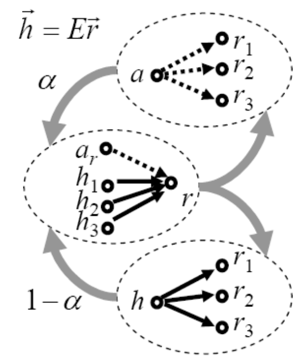
PHITS and More

- PHITS: Cohn and Chang (*2000, ICML*)
  - Only the principal eigenvector is extracted using HITS/SALSA, so the authority along the remaining eigenvectors is completely neglected
    - Account for more eigenvectors of the co-citation matrix
- See also Lempel, Moran (2003, 2004)

An Example of Three-tier HITS: EigenRumor (www2005)



EigenRumor community model

	PageRank	HITS	EigenRumor
Entities	Web page	Web page	Agent/Object
Link types	Evaluation (E)	Evaluation (E)	Evaluation (E) Provisioning (P)
Scores	Authority ($\vec{a}$)	Authority($\vec{a}$) Hub($\vec{h}$)	Authority($\vec{a}$) } Agent Hub($\vec{h}$) Reputation($\vec{r}$) Object
Algorithm	$\vec{a} = (\frac{d}{N}\mathbf{1}_N + (1-d)E^T)\vec{a}$ 	$\vec{h} = E\vec{a}$ $\vec{a} = E^T\vec{h}$ 	$\vec{r} = \alpha P^T\vec{a} + (1-\alpha)E^T\vec{h}$ $\vec{a} = P\vec{r}$ $\vec{h} = E\vec{r}$ 

Comparison with PageRank and HITS Algorithms

Limits of Link Analysis

- META tags/ invisible text
  - Search engines relying on meta tags in documents are often misled (intentionally) by web developers
- Pay-for-place
  - Search engine bias : organizations pay search engines and page rank
  - Advertisements: organizations pay high ranking pages for advertising space
    - With a primary effect of increased visibility to end users and a secondary effect of increased respectability due to relevance to high ranking page
    - Ad-sense
- *Inside Web Page Patron Graph*

Limits of Link Analysis

- Stability
  - Adding even a small number of nodes/edges to the graph has a significant impact
    - *reference Project #3*
- Topic drift – similar to TKC
  - A top authority may be a hub of pages on a different topic resulting in increased rank of the authority page
- Content evolution
  - Adding/removing links/content can affect the intuitive authority rank of a page requiring recalculation of page ranks
  - *Incremental link analysis*
- 子日: 眾好之, 必查之, 眾惡之, 必查之 (論語衛靈公篇)

Similarity measurement by links

- How similar two objects are within a network?
- How to measure the similarity between two objects based on links relationship?
  - E.g., similar friendship

- Measure the similarity between two objects
  - Based on linked-structure
    - Measure the *object-to-object relations*
  - Based on textual content
    - Measure the *keywords co-currency*

  - Linked-based structural similarity measures produce systematically better correlation with human judgements compared to the text-based one [Maguitman etc. WWW06]

Related Work

- Coupling
  - M. M. Kessler, American Documentation, 1963
- Co-Citation
  - H. G. Small, J. of American Society for Information Science, 1973
- ***SimRank***
  - Glen Jeh, Jennifer Widom, KDD'02
  - Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, Denis Turdakov, VLDB'08
- LinkClus
  - Xiaoxin Yin, Jiawei Han, Philip S. Yu VLDB'06
- ***P-Rank***
  - Peixiang Zhao, Jiawei Han, Yizhou Sun, CIKM'09
- RankClus
  - Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, Tianyi We, EDBT'09
- ***NetClus***
  - Yizhou Sun, Yintao Yu, Jiawei Han, KDD'09

SimRank

- Basic idea
  - Based on **Random Surfer model**
  - Two objects are *similar* if they *are linked with* the same or similar objects
  - Consider the ***inlink*** relationship
  - Defined by recursively and computed by iteratively

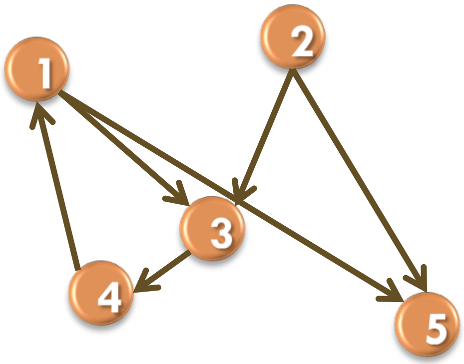
- Discussion in the *Homogeneous* Networks

SimRank

- SimRank formula

$$S(a,b)=\frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)|}\sum_{j=1}^{|I(b)|}S\left(I_i(a),I_j(b)\right)$$

  - *I*(*a*), *I*(*b*): all in-neighbors
  - C is decay factot, 0<C<1
  - *S*(*a*, *b*)∈[0, 1]
  - *S*(*a*, *a*)=1



1'st iteration
S(3, 5)=C/4 * 2
S(4, 5)=0

How about S(4,5) while e(1,2) is added?

P-Rank

- P-Rank formula

$$s(\mathrm{a},\mathrm{b}) = \lambda \times \frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(\mathrm{a})|}\sum_{j=1}^{|I(\mathrm{b})|} \mathrm{s}(I_i(\mathrm{a}), I_j(\mathrm{b})) + (1-\lambda)\times \frac{C}{|O(a)||O(b)|}\sum_{i=1}^{|O(\mathrm{a})|}\sum_{j=1}^{|O(\mathrm{b})|} \mathrm{s}(O_i(\mathrm{a}), O_j(\mathrm{b}))$$

  - *I(a)*, *I(b)*: all in-neighbors
  - O*(a)*, O*(b)*: all out-neighbors
  - C is damping factor , C∈[0, 1]
  - λ is a parameter to balance the relative ***weight of in-link and out-link directions***, λ∈[0, 1]
  - *s(a, b)*∈[0, 1]
  - *s(a, a)*=1

Link analysis in a social network

- Node → entity
- Edge → relationship
- We want to know in this social network
  - Which (group of) node / edge is influential
  - Which (group of) node / edge is important
  - Which node is an outlier
  - Information flow

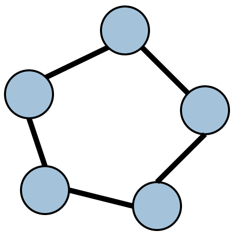
Centrality

- **Degree** centrality
  - In-degree, out-degree
  - Localization, isolation
- **Closeness** centrality
  - Geodesic distance between the entity and all other entities
- **Betweeness** centrality
  - Gendesic path
- **Eigenvector** centrality
  - Central entity receiving many communications from other well-connected entities (central entities)
- **Power** centrality

Network centralization

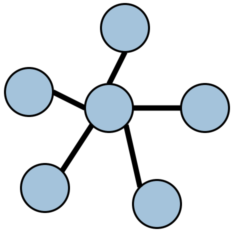
- Summary of centralization of a network
  - E.g.,

$$NET_{Degree} = \frac{\sum_{v \in V} Max_{v \in V} Degree(v) - Degree(v)}{(n-1)*(n-2)}$$



Centralization = 0

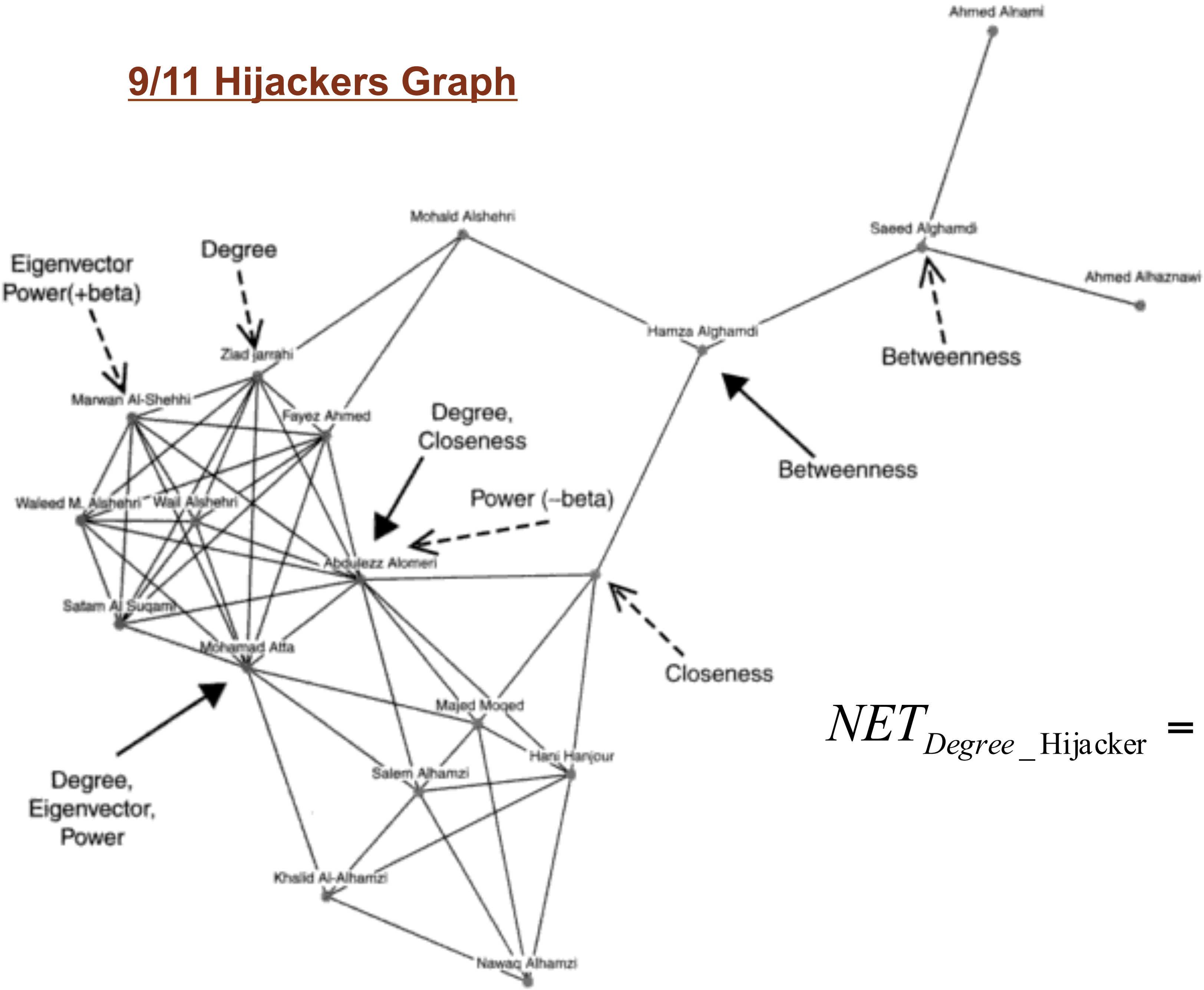
$$NET_{Degree} = \frac{\sum_{v \in V} 2-2}{(n-1)*(n-2)}$$



Centralization = 1

$$NET_{Degree} = \frac{\sum_{v \in V} (n-1)-1}{(n-1)*(n-2)} = \frac{(n-1)(n-2)}{(n-1)(n-2)} = 1$$

9/11 Hijackers Graph

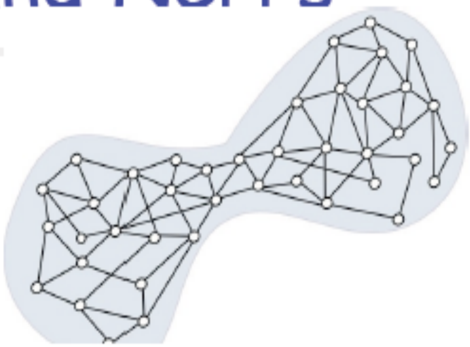


$$NET_{Degree\_\text{Hijacker}} = 0.31$$

Reference from "The Text Mining Handbook", Ronen Feldman, James Sanger, P257.

http://snap.stanford.edu/ncp/

Communities, Conductance, and NCPPs



Let A be the adjacency matrix of G=(V,E).

The conductance ϕ of a set S of nodes is:

$$\phi(S) = \frac{\sum_{i \in S, j \notin S} A_{ij}}{\min\{A(S), A(\bar{S})\}}$$

=s/(s+2e),
s: #edges with one endpoint in S and one endpoint in S complement
e: #edges with both endpoints in S

$$A(S) = \sum_{i \in S} \sum_{j \in V} A_{ij}$$

A: adjacency matrix of G

The Network Community Profile (NCP) Plot of the graph is:

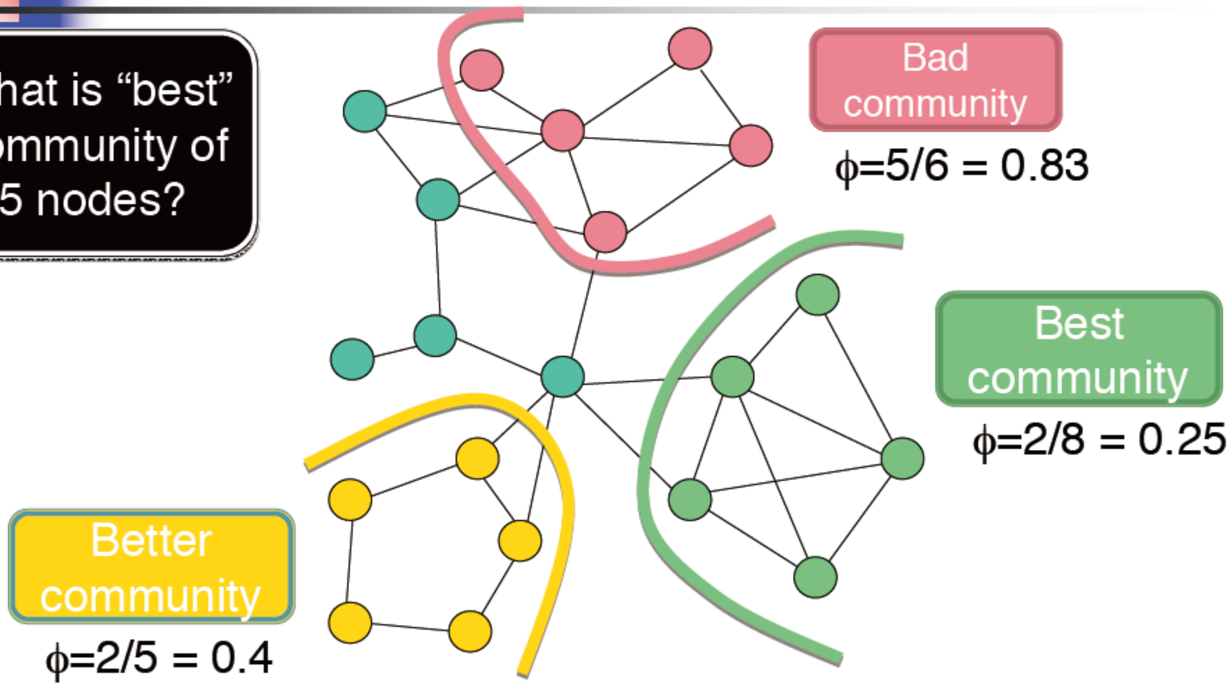
$$\Phi(k) = \min_{S \subset V, |S| = k} \phi(S)$$

Just as conductance captures the "gestalt" notion of cluster/community quality, the NCP plot measures cluster/community quality as a function of size.

NCP is intractable to compute --> use approximation algorithms!

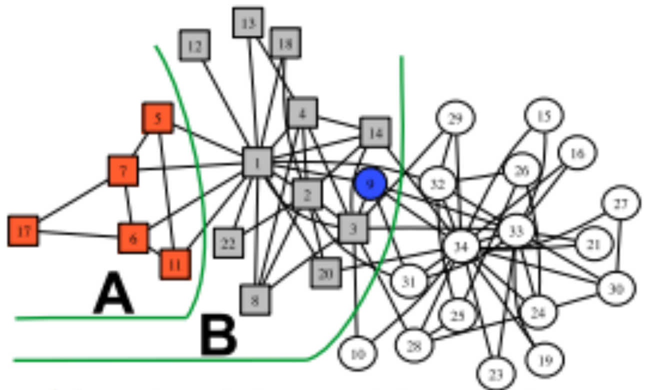
Conductance

Community Score: Conductance

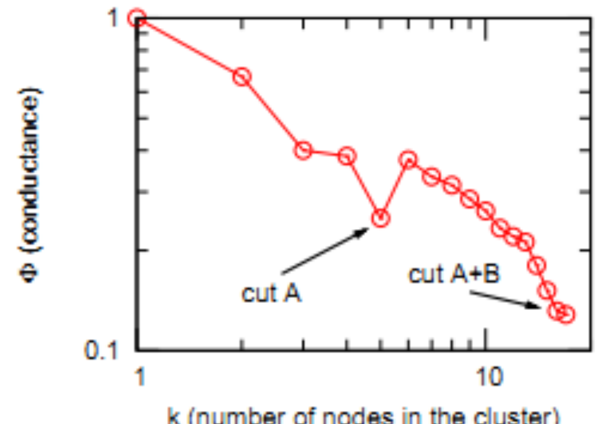


Score: $\phi(S)$ = # edges cut / # edges inside

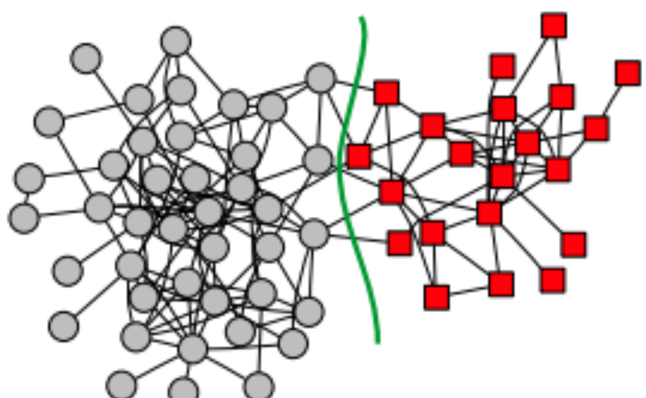
NCPP examples



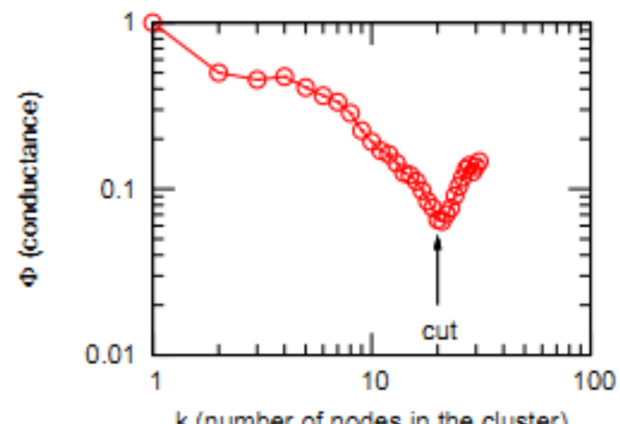
(a) Zachary's karate club network . . .



(b) . . . and it's community profile plot

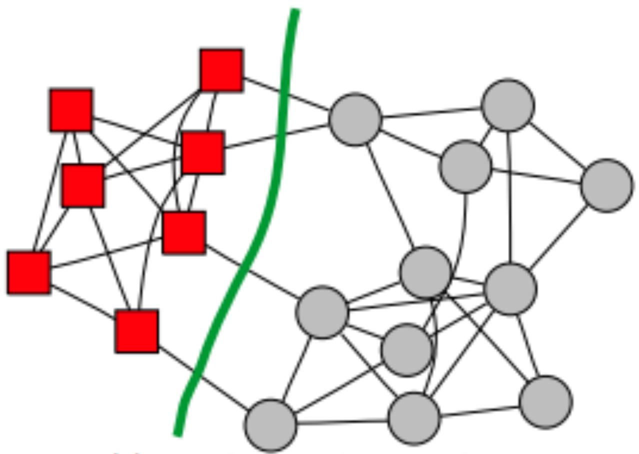


(c) Dolphins social network . . .

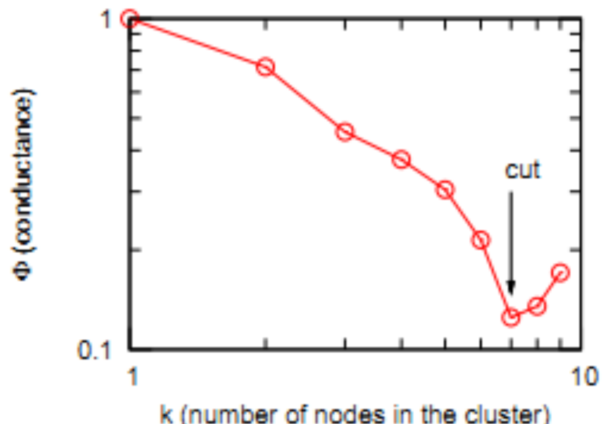


(d) . . . and it's community profile plot

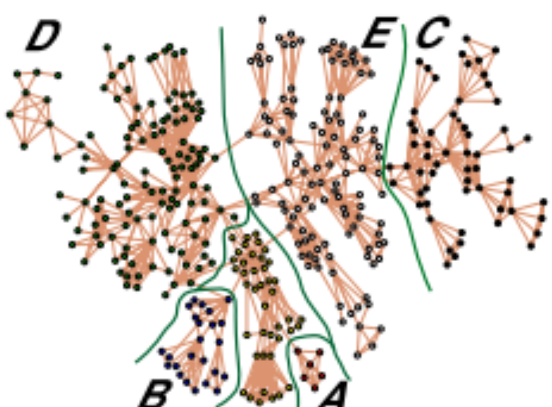
NCPP examples



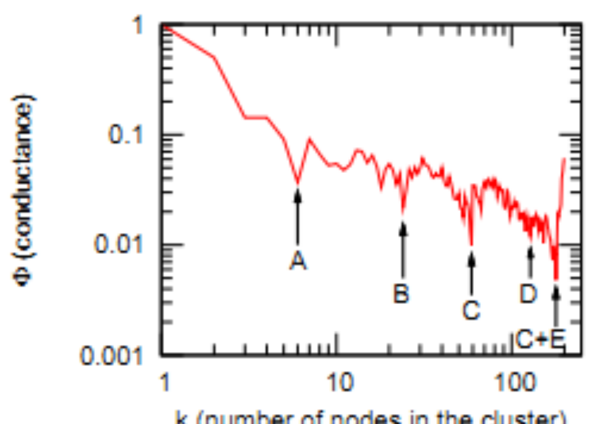
(e) Monks social network . . .



(f) . . . and it's community profile plot



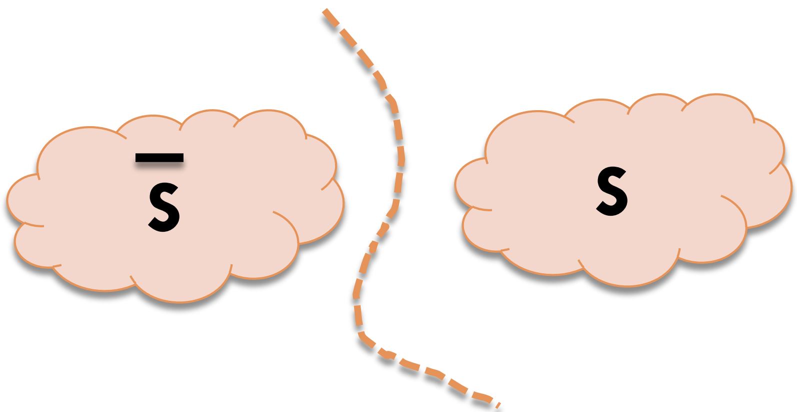
(g) Network science network . . .



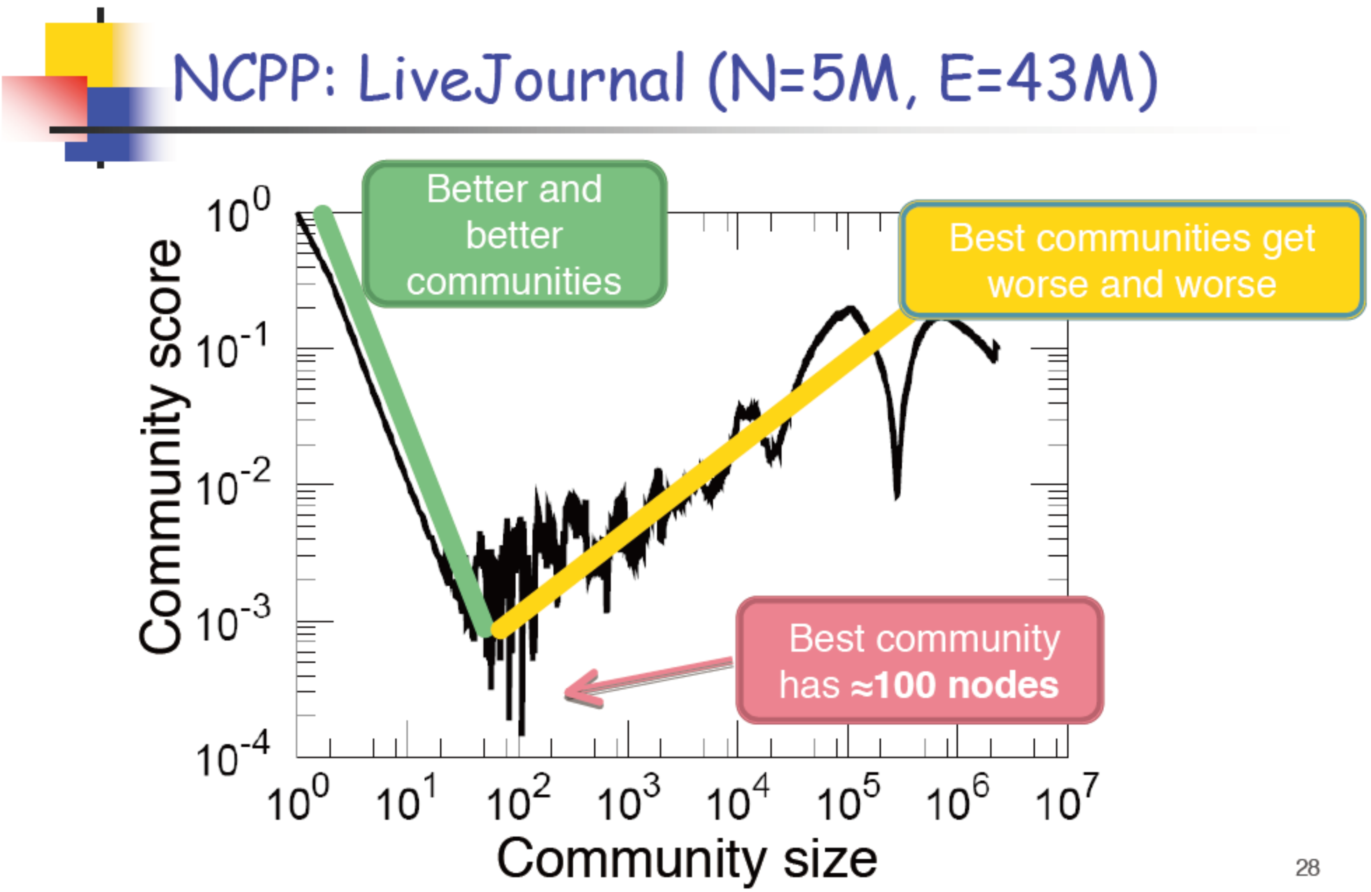
(h) . . . and it's community profile plot

Conductance

- conductance $\Phi(S)$ of a set S of nodes in a graph that equals to the number of edges between S and its complement divided by the sum of the degrees of the nodes inside S
- The lower the conductance the more expressed and more community-like a set of nodes is



Conductance



Reference paper

- **Statistical Properties of Community Structure in Large Social and Information Networks.** Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, Michael W. Mahoney. *WWW 2008*

Further Reading

- R. Lempel and S. Moran, Rank Stability and Rank Similarity of Link-Based Web Ranking Algorithms in Authority Connected Graphs, Inf. Retrieval. Vol 8(2): 245-264 (2005)
- M. Henzinger, Link Analysis in Web Information Retreival, Bulletin of the IEEE computer Society Technical Committee on Data Engineering, 2000.
- L. Getoor, N. Friedman, D. Koller, and A. Pfeffer. *Relational Data Mining*, S. Dzeroski and N. Lavrac, Eds., Springer-Verlag, 2001

Can you think of any circumstances where being "central" might make one less influential? less powerful?

Adversarial Information Retrieval on the Web

- search engine spam and optimization (SEO)
- link-bombing (a.k.a. Google-bombing)
- comment spam, referrer spam
- blog spam (splogs)
  - 部落格觀察 (http://look.urs.tw/) (close, 2006~2010)
- malicious tagging
- reverse engineering of ranking algorithms