Data Mining 2025

Link Analysis

Dept. of Computer Science and Information Engineering

National Cheng Kung University

Kun-Ta Chuang

ktchuang@mail.ncku.edu.tw

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]
- Issues
  - The number of citation base
  - Normalization?

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	
☐	12	NAT STRUCT MOL BIOL	1545-9993	22401	12.712	12.114	

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 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, $F \in (0, 1)$
  - **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

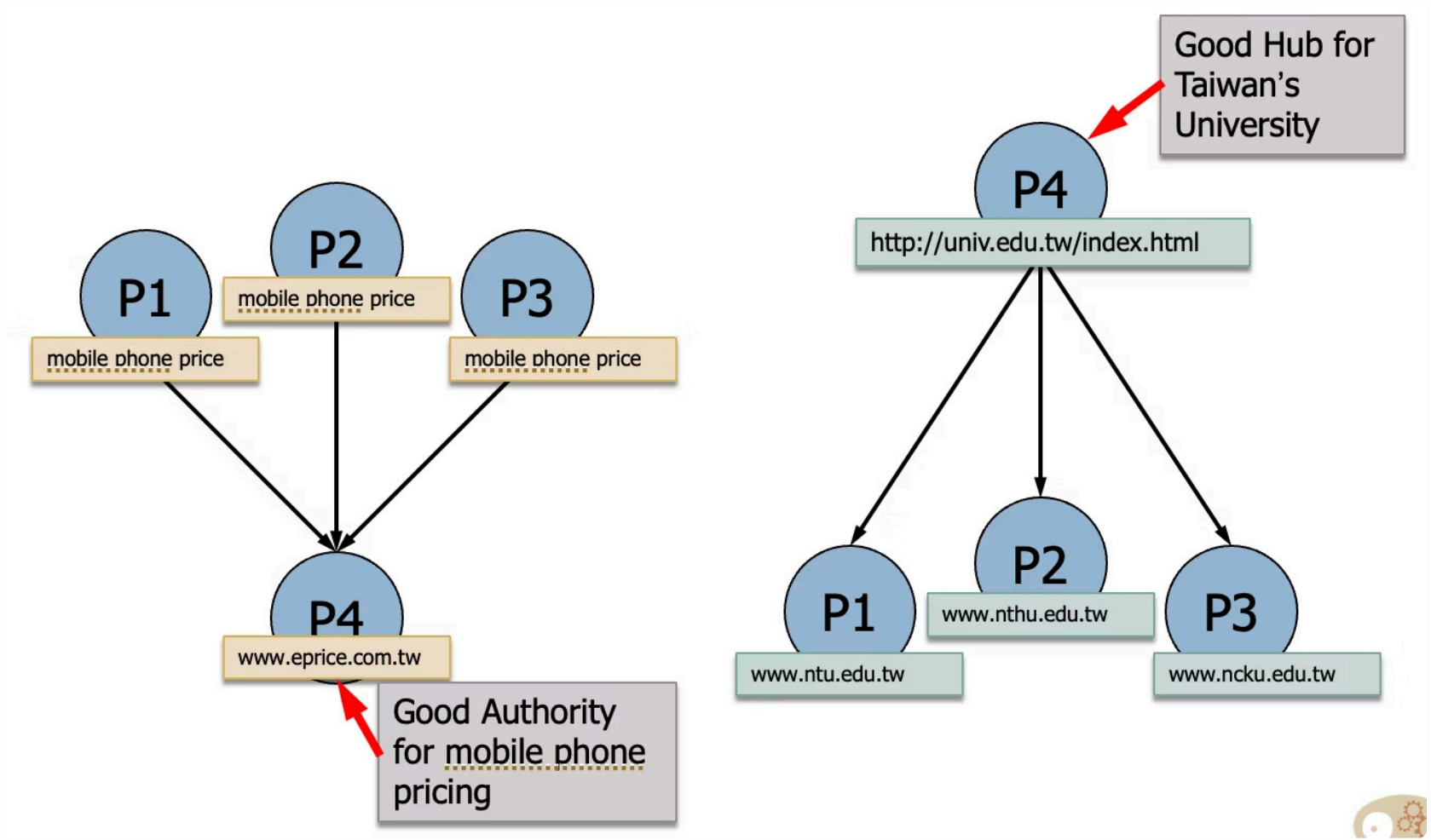
雞生蛋，蛋生雞?

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

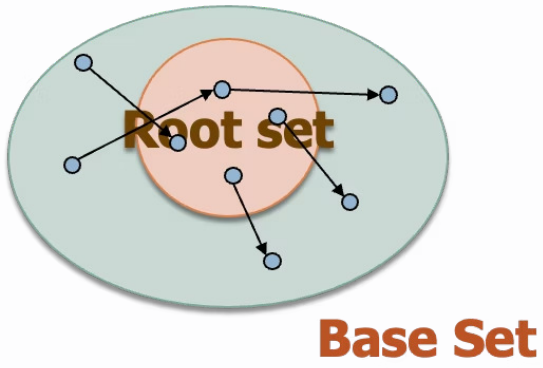


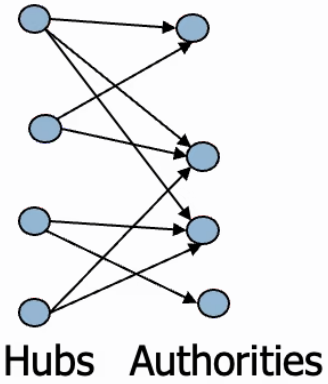
Introduction

- 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:

$a(v) \approx \Sigma w \in pa[v]\ h(w)$

$h(v) \approx \Sigma w \in ch[v]\ a(w)$

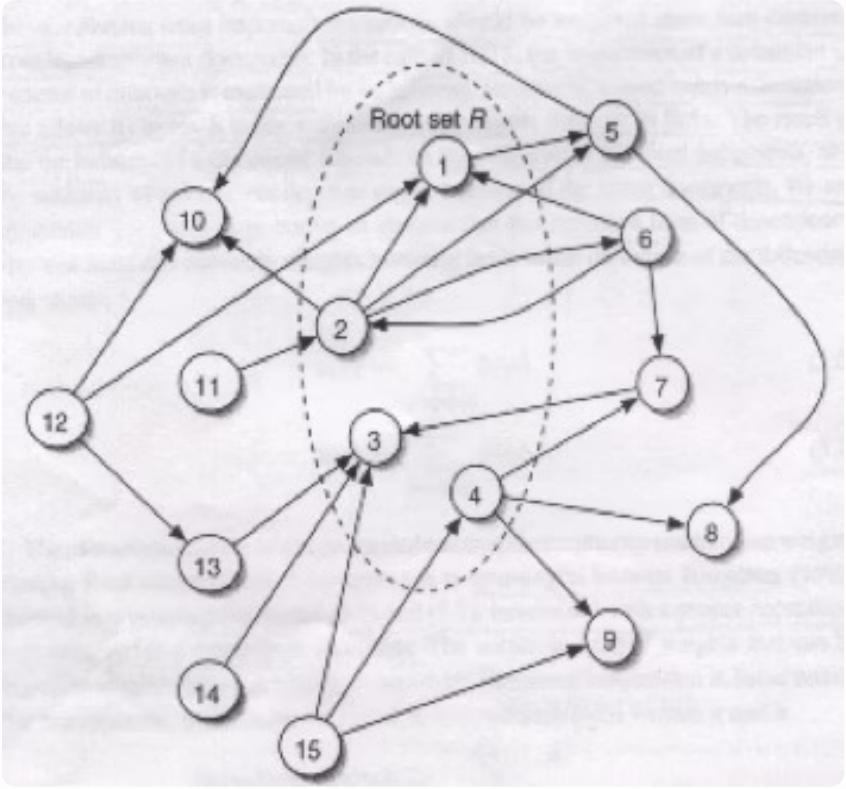
- Using Linear Algebra, we can prove:

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*



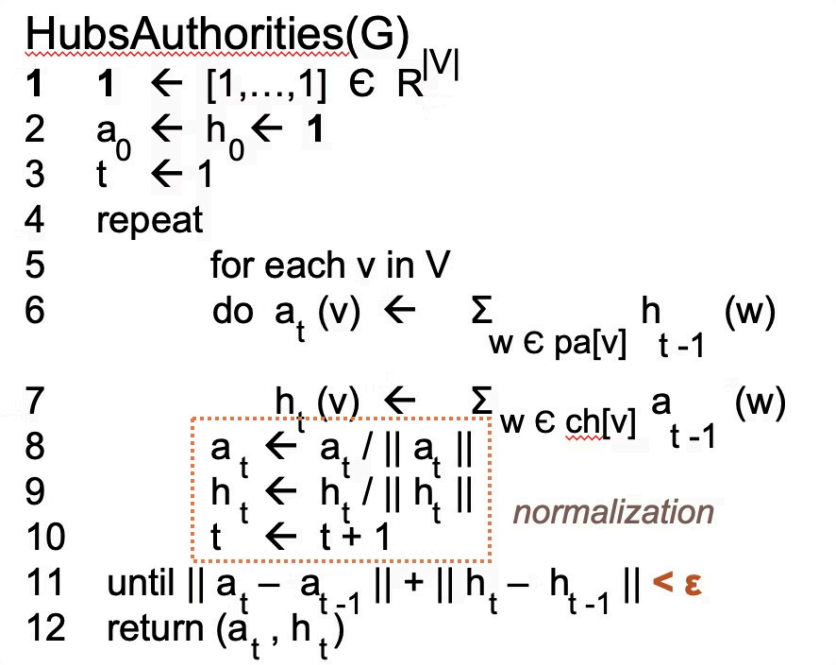
A new set S (base subgraph)

HITS Example

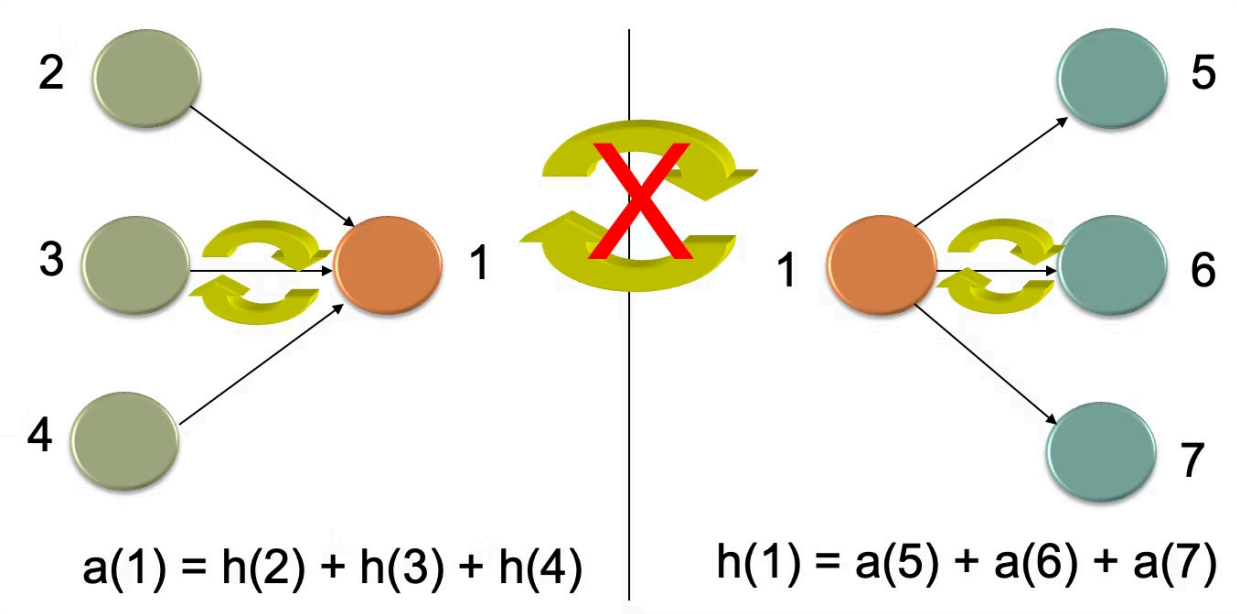
```
BaseSubgraph( R, d)
1.  S <- r
2.  for each v in R
3.      do S <- S U ch[v]
4.         P <- pa[v]
5.         if |P| > d
6.            then P <- arbitrary subset of P having size d
7.                 S <- S U P
8.  return S
```

HITS Example

Hubs and authorities: two n-dimensional **a** and **h**



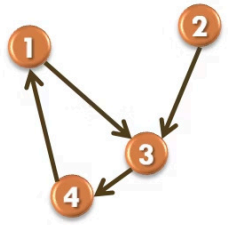
Authority and Hubness



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



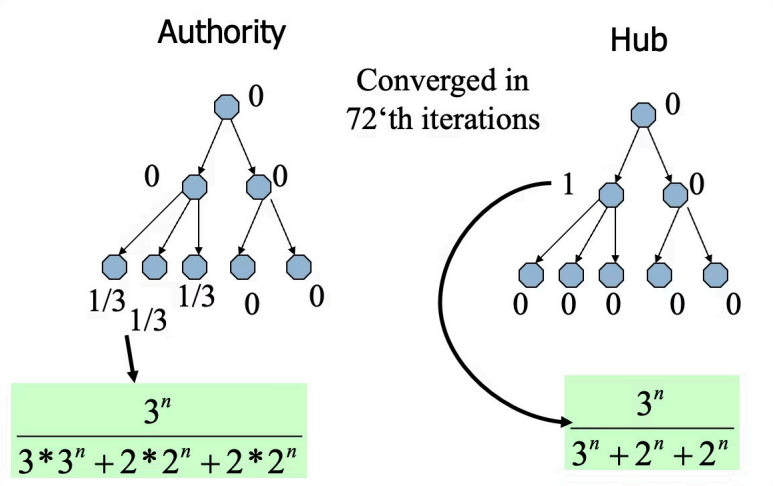
$$A = \begin{bmatrix} 0010 \\ 0010 \\ 0001 \\ 1000 \end{bmatrix} \qquad A^t A = \begin{bmatrix} 1000 \\ 0000 \\ 0020 \\ 0001 \end{bmatrix} \qquad AA = \begin{bmatrix} 0001 \\ 0001 \\ 1000 \\ 0010 \end{bmatrix}$$

In-Out

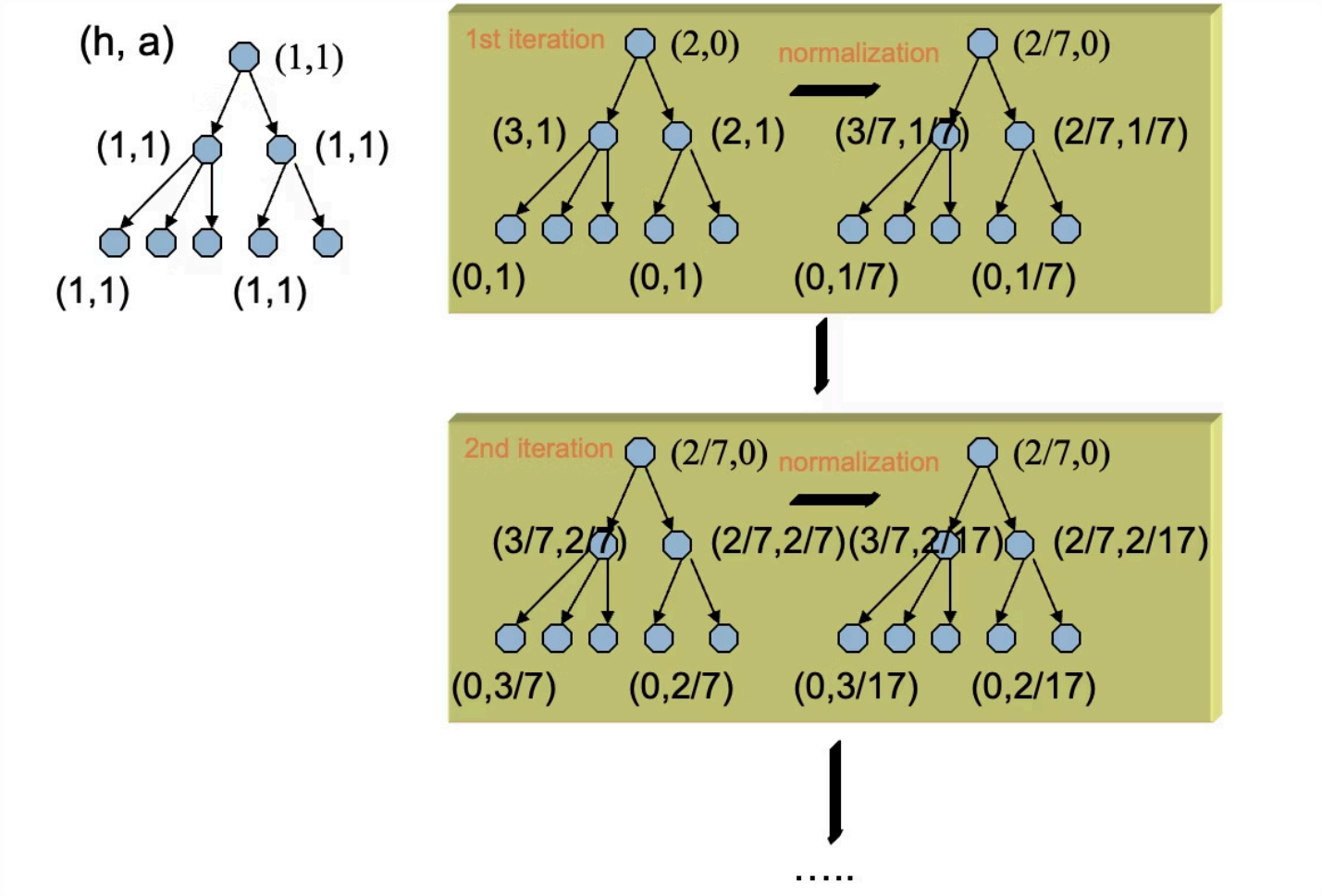
$$A^t = \begin{bmatrix} 0001 \\ 0000 \\ 1100 \\ 0010 \end{bmatrix} \qquad AA^t = \begin{bmatrix} 1100 \\ 1100 \\ 0010 \\ 0001 \end{bmatrix}$$

Out-In

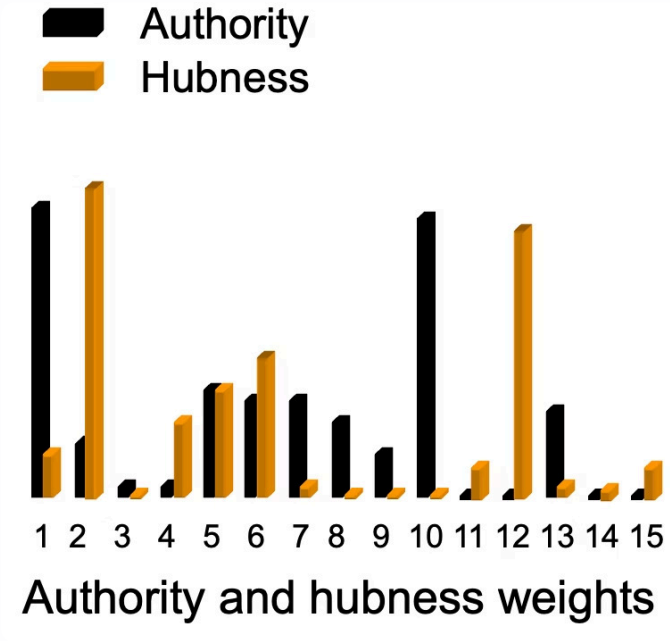
Example (1-norm normalization)

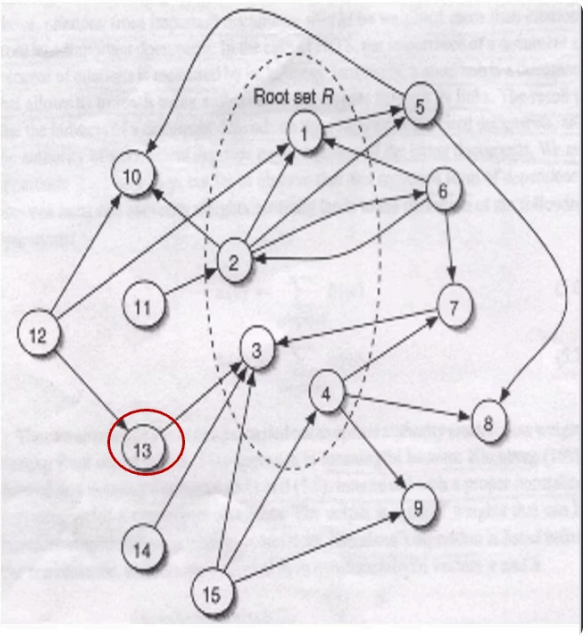


Example (1-norm normalization)



HITS Example Results



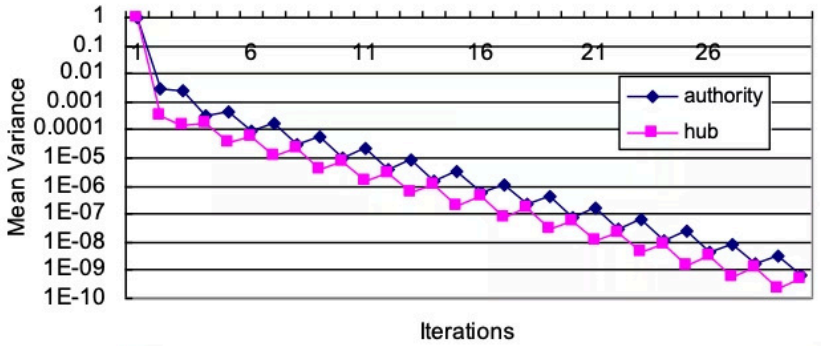


Issues for HITS

- 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



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

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