1.4.4 扩展

- 如何高效地计算出互联网上各网页的PageRank得分？针对该问题大致可以分为两大类:
  - 高效计算图中所有节点的PageRank得分
  - 互联网上少量特定网页的PageRank得分，该问题被称为single-node PageRank (单点PageRank计算)
- 针对单点PageRank计算, 有兴趣的同学自行阅读课外书籍和最新研究, 如Revisiting local computation of pagerank: Simple and optimal等

Some Problems with PageRank

- **Measures generic popularity of a page**
  - Biased against topic-specific authorities
  - **Solution:** Topic-Specific PageRank (**next**)
- **Susceptible to Link spam**
  - Artificial link topographies created in order to boost page rank
  - **Solution:** TrustRank (**next**)
- **Uses a single measure of importance**
  - Other models of importance
  - **Solution:** Hubs-and-Authorities (**next**)

Section 1.5: Topic-Specific PageRank

Content

1.5.1 Topic-Specific PageRank

- **Allows search queries to be answered also based on interests of the user**
  - **Example:** Query “Trojan” wants different pages depending on whether you are interested in sports, history and computer security

1.5.1 Topic-Specific PageRank

- **Instead of generic popularity, can we measure popularity within a topic?**

- **Goal:** Evaluate web pages not just according to their **popularity**, but by **how close they are to a particular topic**

1.5.1 Topic-Specific PageRank

- Random walker has a small probability of **teleporting** at any step, teleport can go to:
  - **Standard PageRank: Any page with equal probability**
    - To avoid dead-end and spider-trap problems
  - **Topic-Specific PageRank (面向主题的PageRank), also known as Personalized PageRank (个性化PageRank): A topic-specific set *S* of "relevant" pages (teleport set *S*, 随机跳转集合)**
- **Idea: Bias the random walk (有偏的随机游走模型)**
  - When walker teleports, she picks a page from a set ***S***
  - ***S*** contains only pages that are relevant to the topic
    - E.g., Open Directory (DMOZ) pages for a given topic/query
  - For each teleport set ***S***, we get a different vector r_S

1.5.1 Matrix Formulation

- **To make this work all we need is to update the teleportation part of the PageRank formulation:**

$$A_{ij} = \begin{cases} \beta M_{ij} + (1-\beta)/|S| & \text{if } i \in S \\ \beta M_{ij} + 0 & \text{otherwise} \end{cases}$$

$$A = \beta M + (1-\beta)\left[\frac{1}{N}\right]_{N\times N}$$

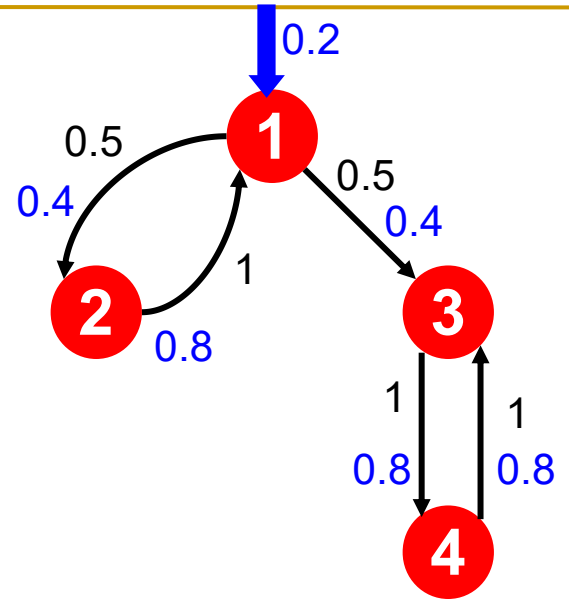
for basic PageRank

  - ***A*** is stochastic!

- We weighted all pages in the teleport set ***S*** equally
  - Could also assign different weights to pages!

- **Compute as for regular PageRank:**
  - Multiply by ***M***, then add a vector
  - Maintains sparseness

1.5.1 Example



Suppose $S = \{1\}$, $\beta = 0.8$

Node	Iteration				
	0	1	2	…	stable
1	0.25	0.4	0.28		0.294
2	0.25	0.1	0.16		0.118
3	0.25	0.3	0.32		0.327
4	0.25	0.2	0.24		0.261

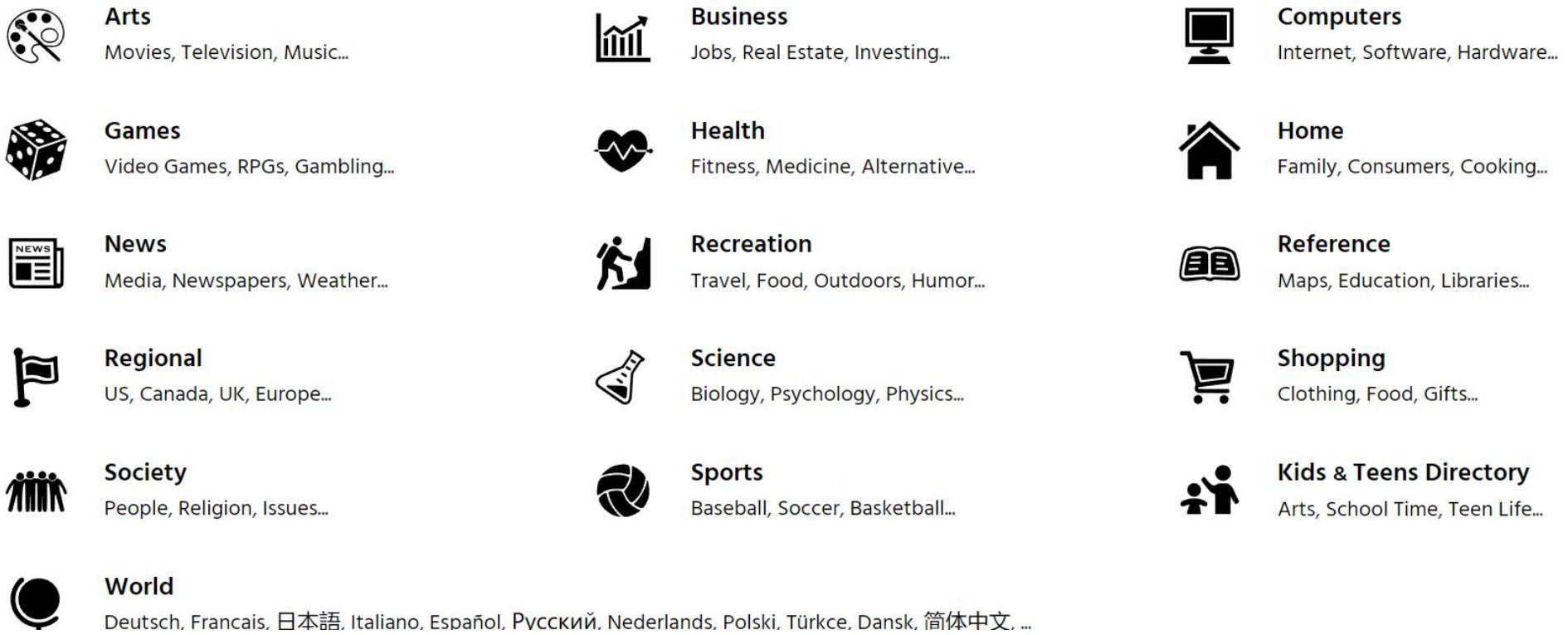
S={1}, β=0.90:
r=[0.17, 0.07, 0.40, 0.36]
S={1} , β=0.8:
r=[0.29, 0.11, 0.32, 0.26]
S={1}, β=0.70:
r=[0.39, 0.14, 0.27, 0.19]

S={1,2,3,4}, β=0.8:
r=[0.13, 0.10, 0.39, 0.36]
S={1,2,3} , β=0.8:
r=[0.17, 0.13, 0.38, 0.30]
S={1,2} , β=0.8:
r=[0.26, 0.20, 0.29, 0.23]
S={1} , β=0.8:
r=[0.29, 0.11, 0.32, 0.26]

1.5.2 Discovering the Topic Vector S

Create different PageRanks for different topics

- The 16 DMOZ top-level categories: arts, business, sports,…



【备注】https://dmoztools.net/

1.5.2 Discovering the Topic Vector S

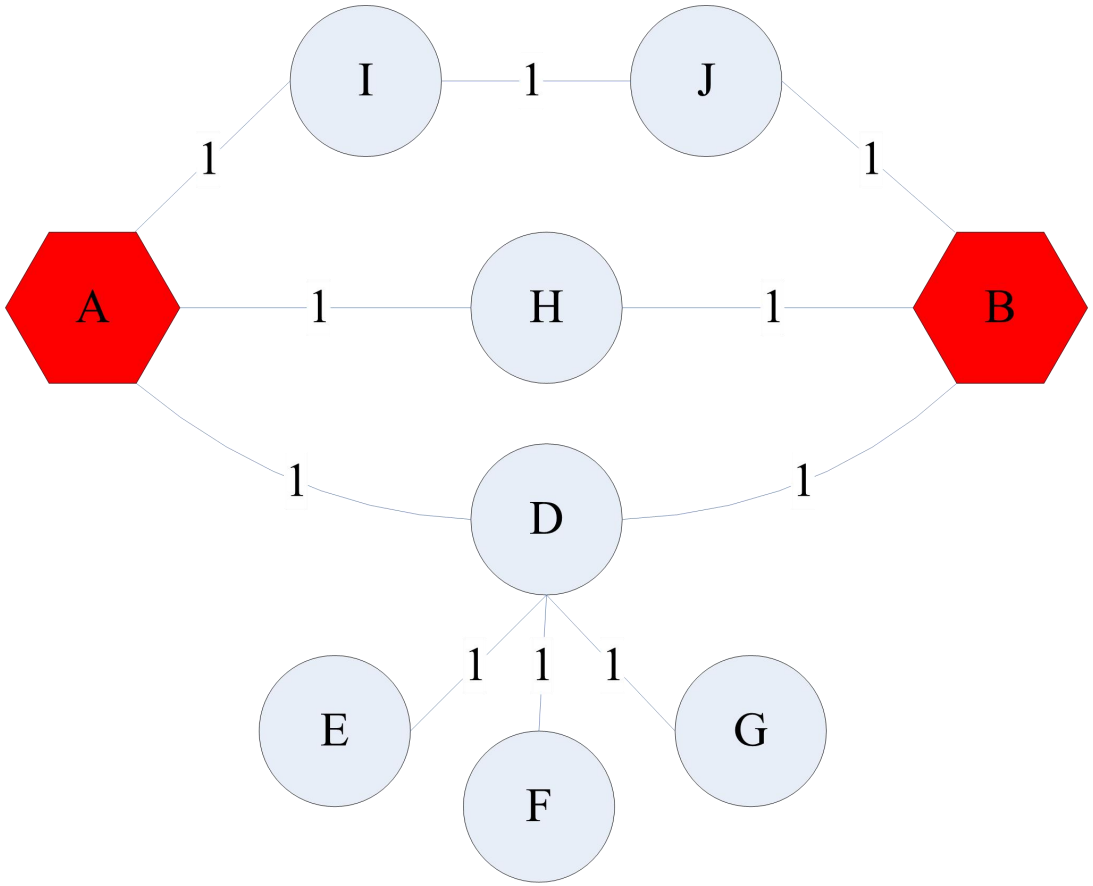
- **Which topic ranking to use?**
  - User can pick from a menu
  - Classify query into a topic
  - Can use the **context** of the query
    - E.g., query is launched from a web page talking about a known topic
    - History of queries e.g., "basketball" followed by "Jordan"
  - User context, e.g., user's bookmarks, …

1.5.3 Application to Measuring Proximity

- **Special Case** in Topic-Specific PageRank: **Random Walk with Restarts (重启随机游走算法, PWR)**: *S* is **a single element**

e.g., S = {A}

1.5.3 Proximity on Graphs

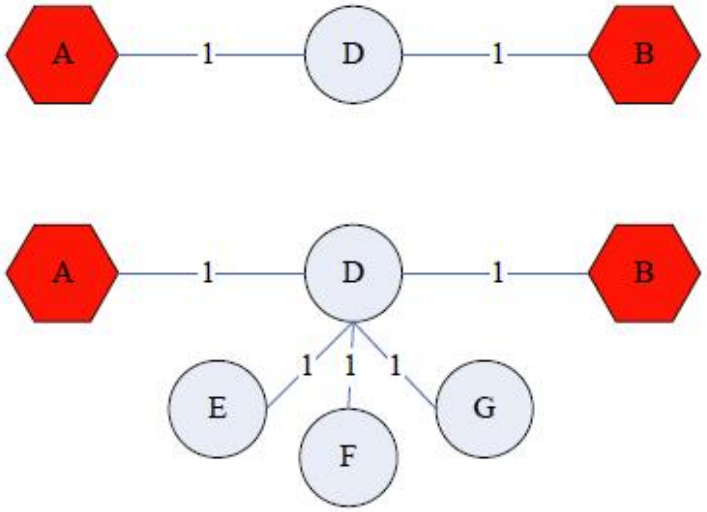


a.k.a.: Relevance, Closeness, 'Similarity'…

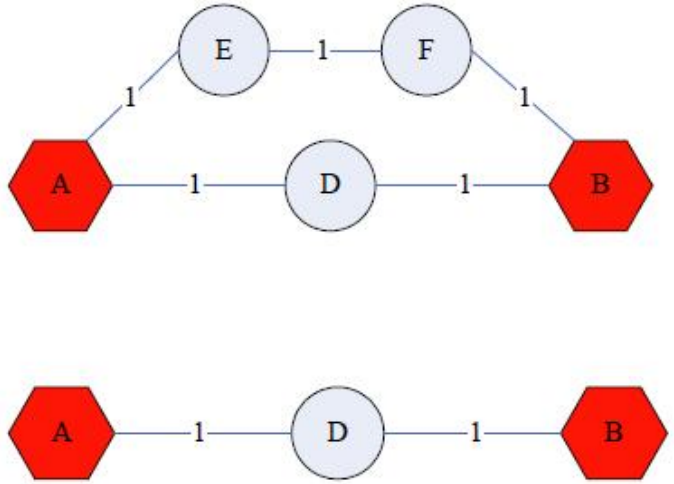
[Tong-Faloutsos, '06]

1.5.3 Good proximity measure?

- **Shortest path is not good:**



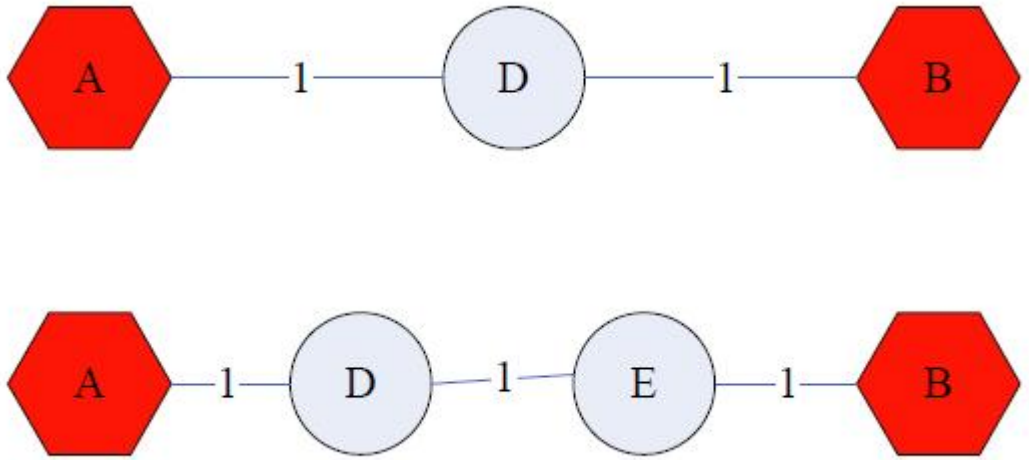
No effect of degree-1 nodes (E, F, G)!



Multi-faceted relationships

1.5.3 Good proximity measure?

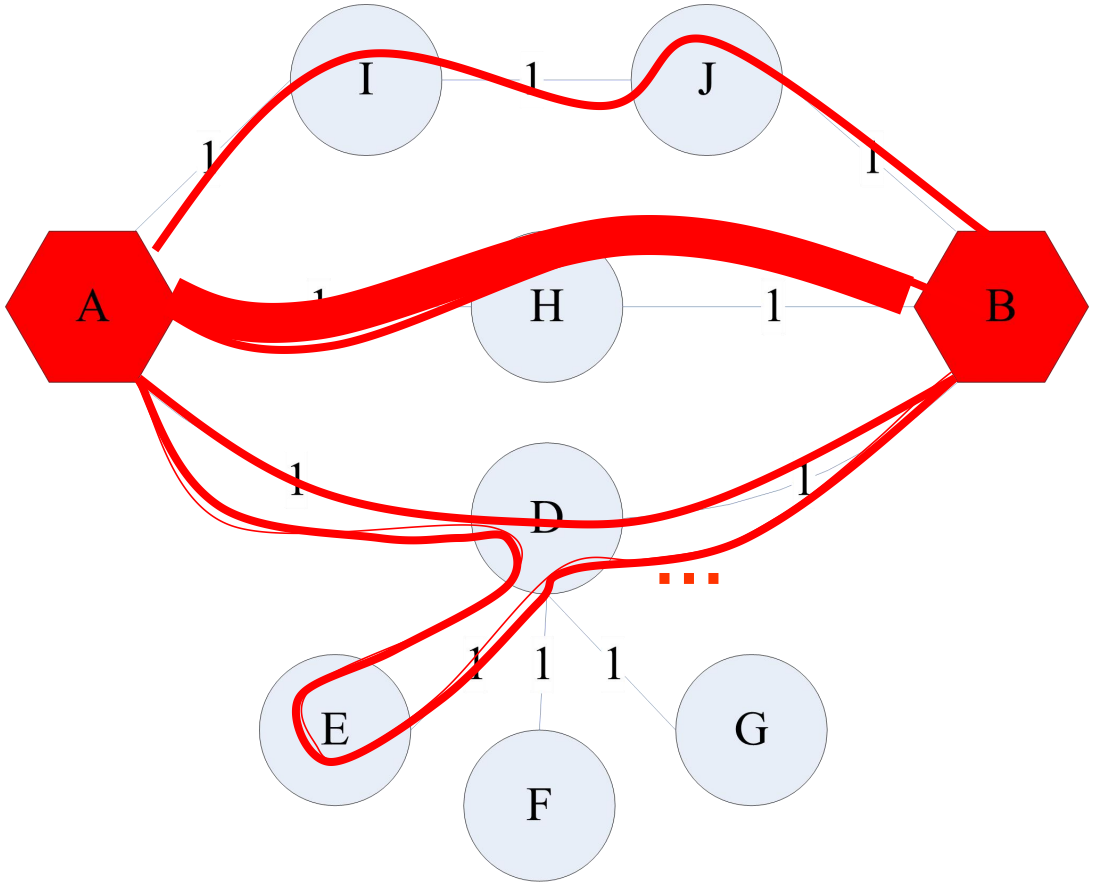
- Network flow is not good:



Does not punish long paths

1.5.3 What is good notion of proximity?

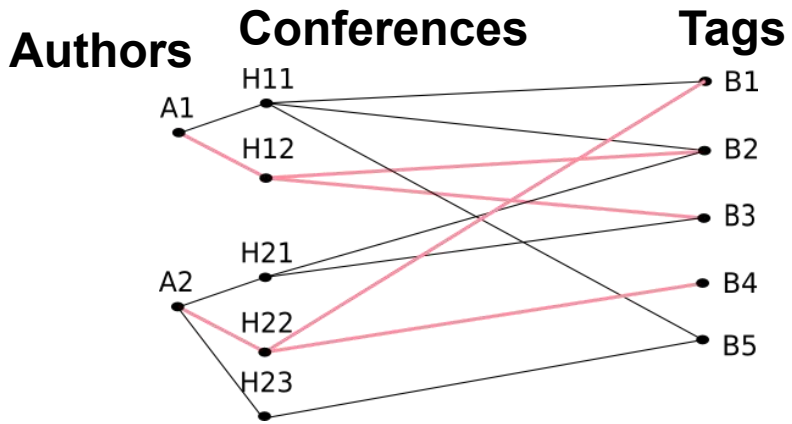
[Tong-Faloutsos '06]



- **Multiple connections**
- **Quality of connection**
  - **Direct & Indirect connections**
  - **Length, Degree, Weight…**

1.5.3 SimRank

- **SimRank:** Random walks from a **fixed node** on ***k***-partite graphs (***k***部图)
  - Setting: ***k***-partite graph with ***k*** types of nodes. e.g.: Authors, Conferences, Tags

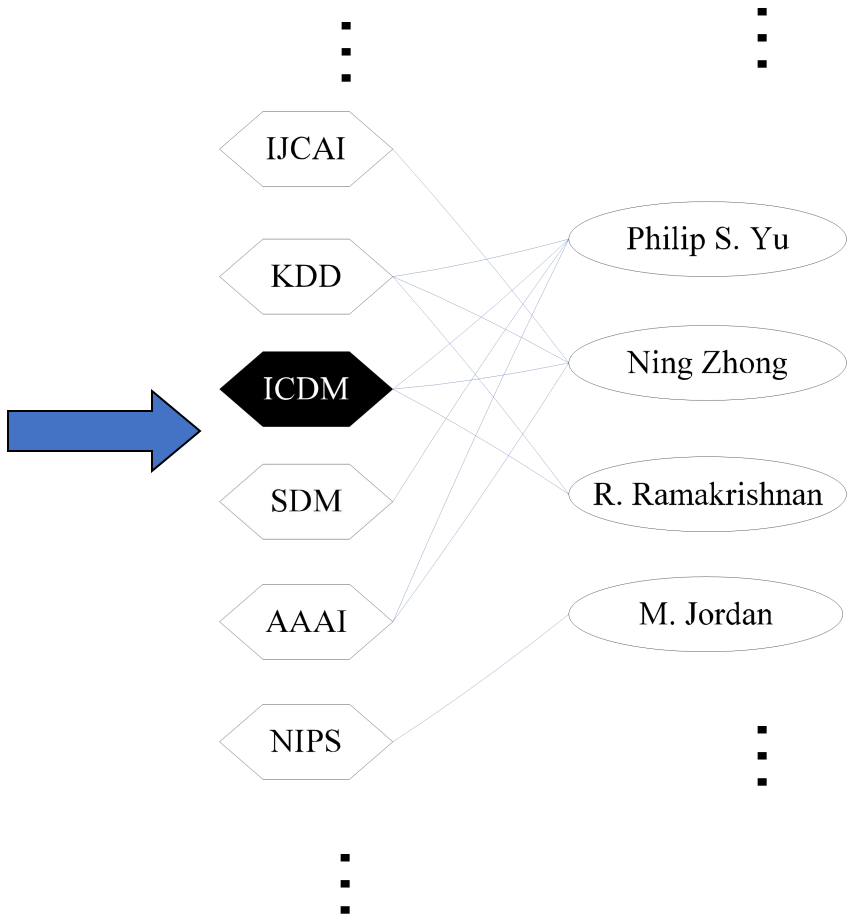


  - Use to measure proximity in graph, page rank, or recommend, …
  - 采用协同过滤思想: 如果a和b分别与c和d关联, 且已知c与d是相似, 则a与b也相似 (推荐系统章节会再详细讲)
  - 更多细节课外阅读相关文献. e.g., "Simrank: a measure of structural-context similarity"

1.5.3 Random Walk with Restarts

- **Random Walk with Restarts(重启随机游走算法, PWR):** Topic Specific PageRank from the node *u*, **teleport set *S* = {*u*}**
- Resulting scores measures similarity to node ***u***
- **Problem:**
  - Must be done once for each node ***u***
  - Suitable for sub-Web-scale applications

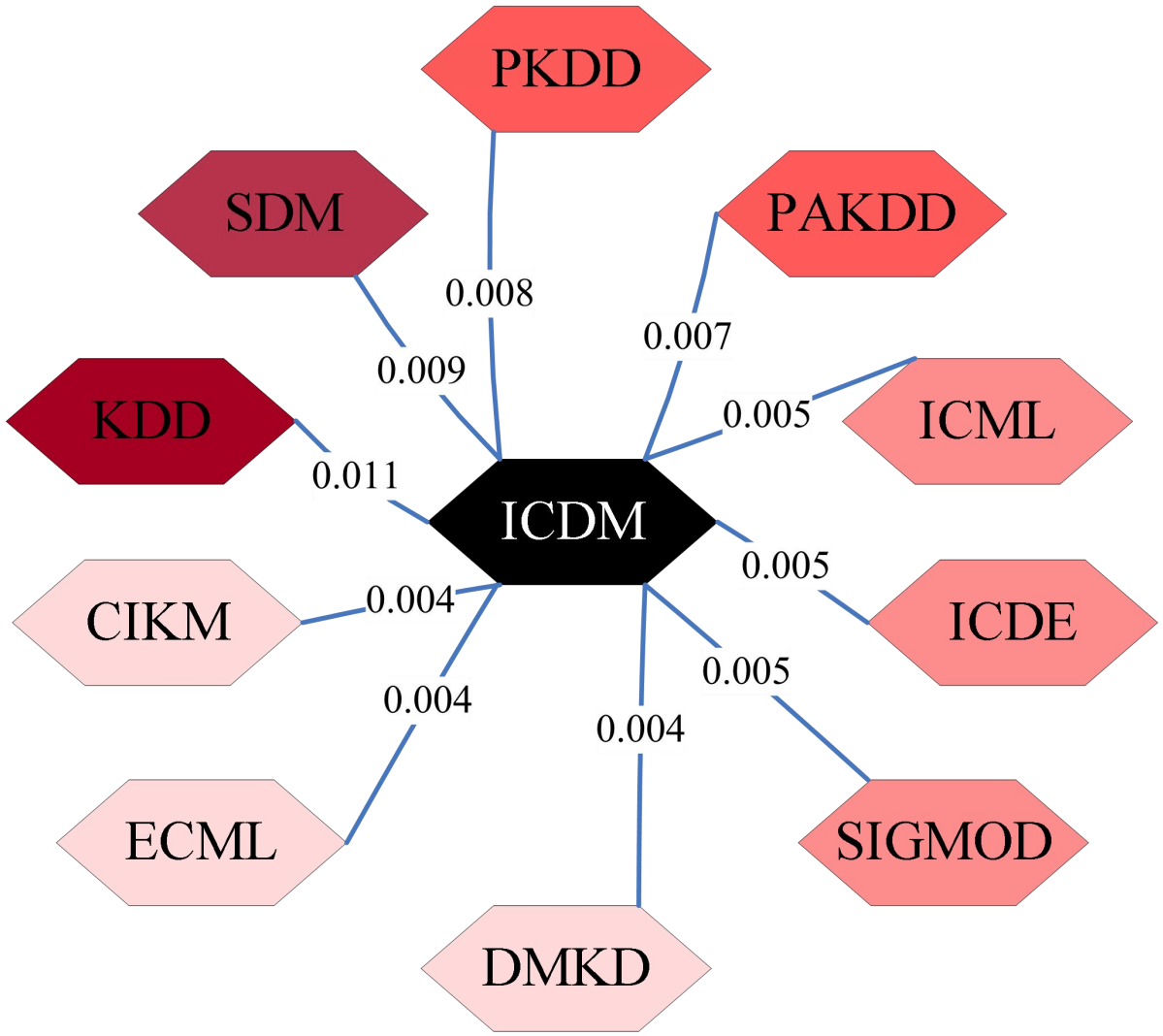
1.5.3 Example



Q: What is most related conference to **ICDM**?

A: Topic-Specific PageRank with teleport set S={ICDM}

1.5.3 Example


PKDD
SDM
PAKDD
0.008
0.009
0.007
KDD
0.005
ICML
0.011
ICDM
0.005
CIKM
0.004
ICDE
0.004
0.005
0.004
ECML
SIGMOD
DMKD

1.5.3 PageRank: Summary

- **"Normal" PageRank:**
  - Teleports uniformly at random to any node
  - All nodes have the same probability of surfer landing there: **S =** [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]

- **Topic-Specific PageRank:**
  - Teleports to a topic specific set of pages
  - Nodes can have different probabilities of surfer landing there: **S =** [0.1, 0, 0, 0.2, 0, 0, 0.5, 0, 0, 0.2]

- **Random Walk with Restarts:**
  - Topic-Specific PageRank where teleport is always to the same node. **S=**[0, 0, 0, 0, **1**, 0, 0, 0, 0, 0, 0]

Section 1.6: TrustRank, Combating the web spam

Content

1.6.1 What is Web Spam?

- **Spamming:**
  - Any deliberate action to boost a web page's position in search engine results, incommensurate with page's real value
- **Spam:**
  - Web pages that are the result of spamming
- This is a very broad definition
  - Search engine optimization (SEO, 搜索引擎优化) industry might disagree!
- Approximately **10-15%** of web pages are spam

1.6.1 Web Search

- **Early search engines:**
  - Crawl the Web
  - Index pages by the words they contained
  - Respond to search queries (lists of words) with the pages containing those words
- **Early page ranking:**
  - Attempt to order pages matching a search query by “importance”
  - **First search engines considered:**
    - **(1)** Number of times query words appeared
    - **(2)** Prominence of word position, e.g. title, header

1.6.1 First Spammers

- As people began to use search engines to find things on the Web, those with commercial interests tried to **exploit search engines** to bring people to their own site – whether they wanted to be there or not
- **Example:**
  - Shirt-seller might pretend to be about “movies”
- **Techniques for achieving high relevance/importance for a web page**

1.6.1 First Spammers: Term Spam

- **How do you make your page appear to be about movies?**
  - **(1)** Add the word movie 1,000 times to your page. Set text color to the background color, so only search engines would see it
  - **(2)** Or, run the query “movie” on your target search engine. See what page came first in the listings. Copy it into your page, make it “invisible”
- **These and similar techniques are term spam (词项作弊)**

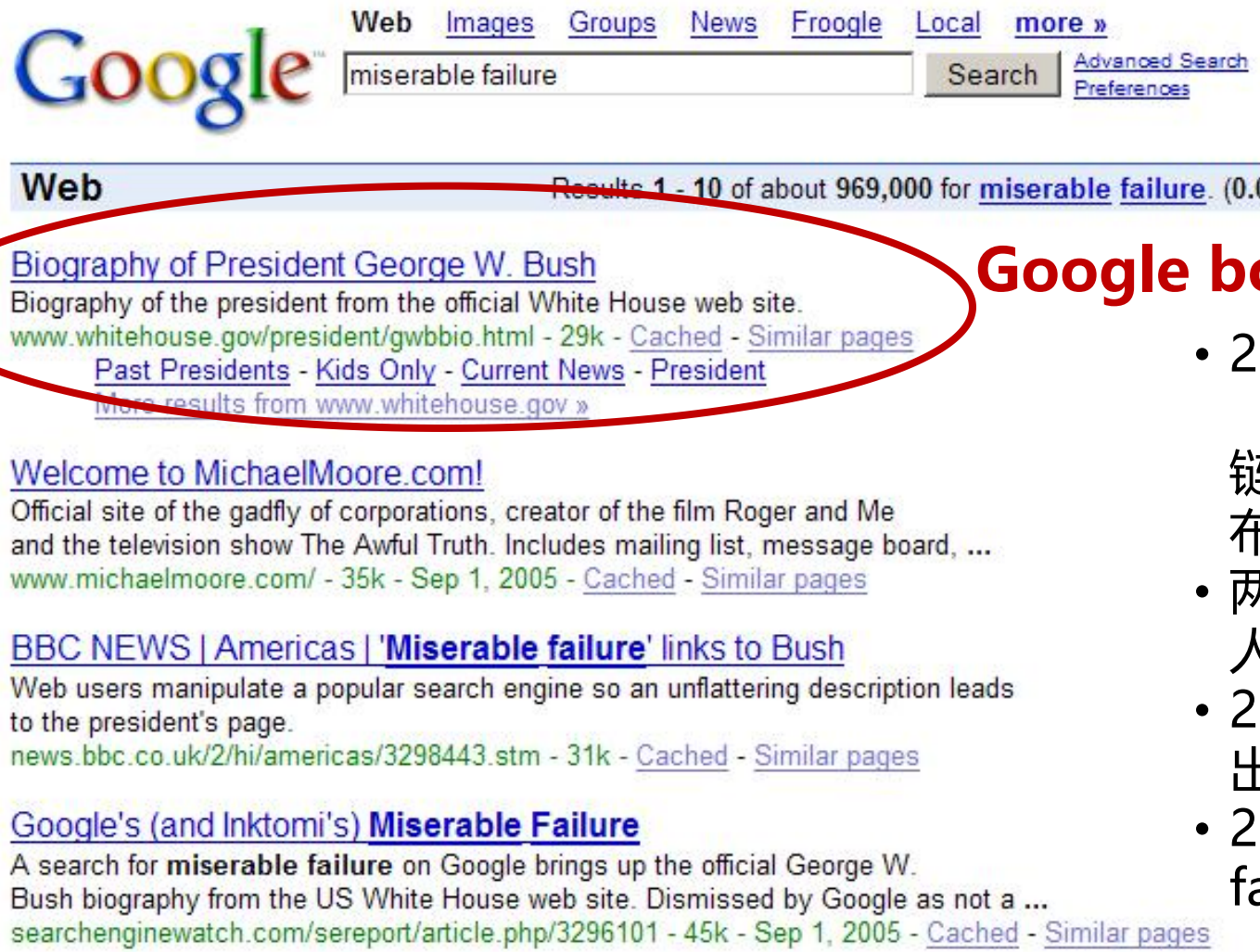
1.6.1 Google's Solution to Term Spam

- **Believe what people say about you, rather than what you say about yourself**
  - Use words in the anchor text (words that appear underlined to represent the link) and its surrounding text
- **PageRank** as a tool to measure the "importance" of Web pages

1.6.1 Why It Works?

- **Our hypothetical shirt-seller looses**
  - Saying he is about movies doesn't help, because others don't say he is about movies
  - His page isn't very important, so it won't be ranked high for shirts or movies
- **Example:**
  - Shirt-seller creates 1,000 pages, each links to his with "movie" in the anchor text
  - These pages have no links in, so they get little PageRank
  - So the shirt-seller can't beat truly important movie pages, like IMDB

1.6.1 Why it does not work?



Google bomb (谷歌炸弹, 或称谷歌轰炸)

- 2003.10,乔治·约翰逊号召人们用“miserable failure”(惨败)关键词做为链接文字, 链接到当时的美国白宫网站的布什总统的个人介绍页.
- 两个月后, Google上搜索该词, 布什的个人介绍页上升到第一位.
- 2007.1, 以色列总理奥尔默特的简历页也出现在该搜索页面中.
- 2009.7, 奥巴马政府出现在“worst ever failure”（史上最失败事件）搜索结果中.

1.6.1 Why it does not work?

1.6.2 Google vs. Spammers: Round 2!

華中科技大學
计算机科学与技术学院
School of Computer Science &Technology，HUST

- Once Google became the dominant search engine, spammers began to work out ways to fool Google

- **Spam farms (垃圾农场)** were developed to concentrate PageRank on a single page

- **Link spam (链接作弊):**
  - Creating link structures that boost PageRank of a particular page

1.6.2 Link Spamming

- Three kinds of web pages from a spammer's point of view
  - 1, Inaccessible pages (不可达网页)
  - 2, Accessible pages (可达网页)
    - e.g., blog comments pages
    - spammer can post links to his pages
  - 3, Owned pages (自有网页)
    - Completely controlled by spammer
    - May span multiple domain names