Chapter 1:

Analysis of Large Graphs: Link Analysis

崔金华
电子邮箱：jhcui@hust.edu.cn
个人主页: https://csjhcui.github.io/

致谢:课件主要参考《Mining of Massive Datasets》(Second Edition) J. Leskovec, A. Rajaraman, J. Ullman的相关课件, 特此致谢!!!

Graph data	High dim. data	Stream data	Machine learning	Apps
PageRank	Locality sensitive hashing	Queries on streams	SVM	Recommender systems
TrustRank	Clustering	Filtering data streams	Decision Trees	Association Rules
HITS	Dimensionality reduction	Counting elements	kNN	MapReduce

Contents


Blazers
Magic
Wizards
Lakers
Spurs
Heat
Clippers
Bucks
Pelicans
Timberwolves
Knicks
Kings
Raptors
Pacers
76ers
Grizzlies
Rockets
Nets
Cavaliers
Warriors
Mavericks
Suns
Hawks
Hornets
Bulls
Jazz
Celtics
Thunder
Pistons

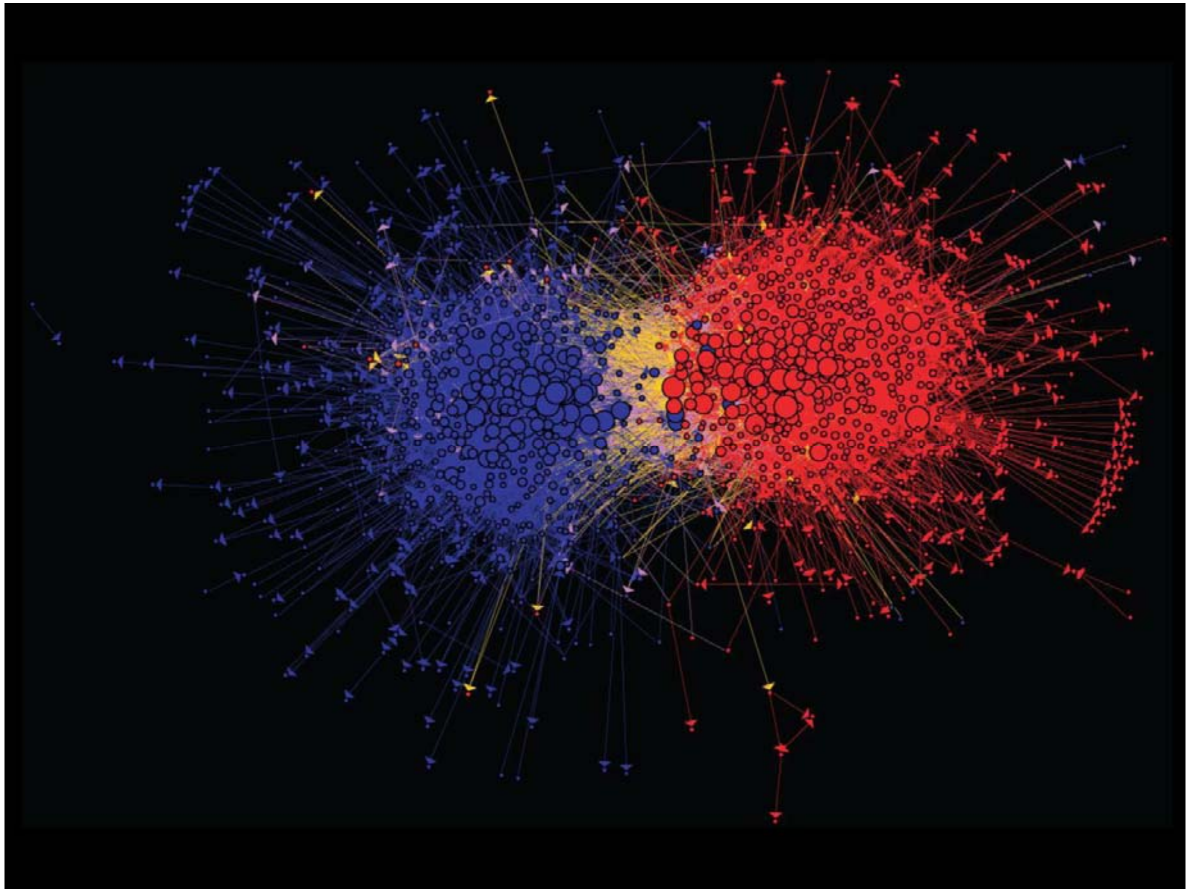
Section 1.1 Graph Data Background

1.1.1 Graph Data: Social Networks

Facebook social graph

4-degrees of separation (四度分离理论) [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

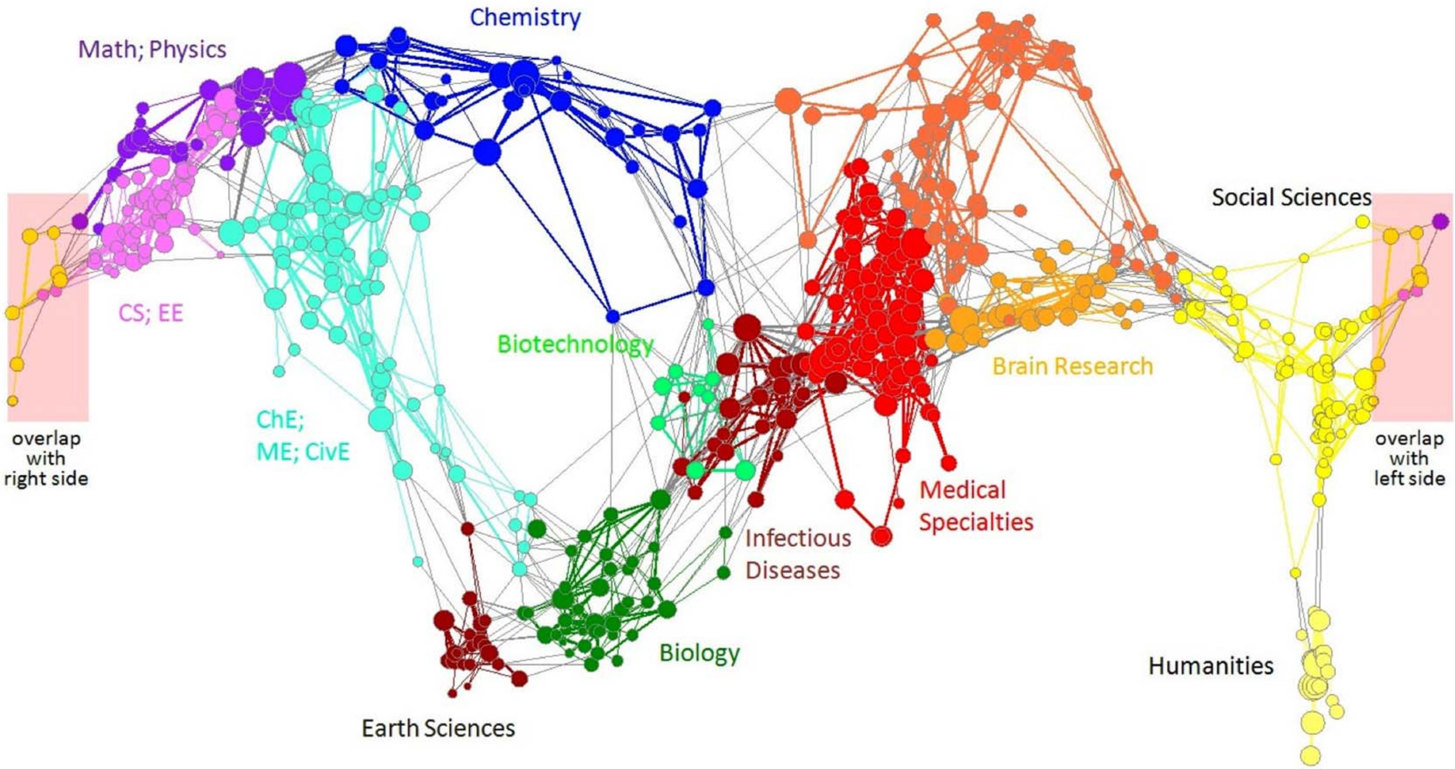
1.1.1 Graph Data: Media Networks



Connections between political blogs(民主党,共和党等)

Polarization of the network [Adamic-Glance, 2005]

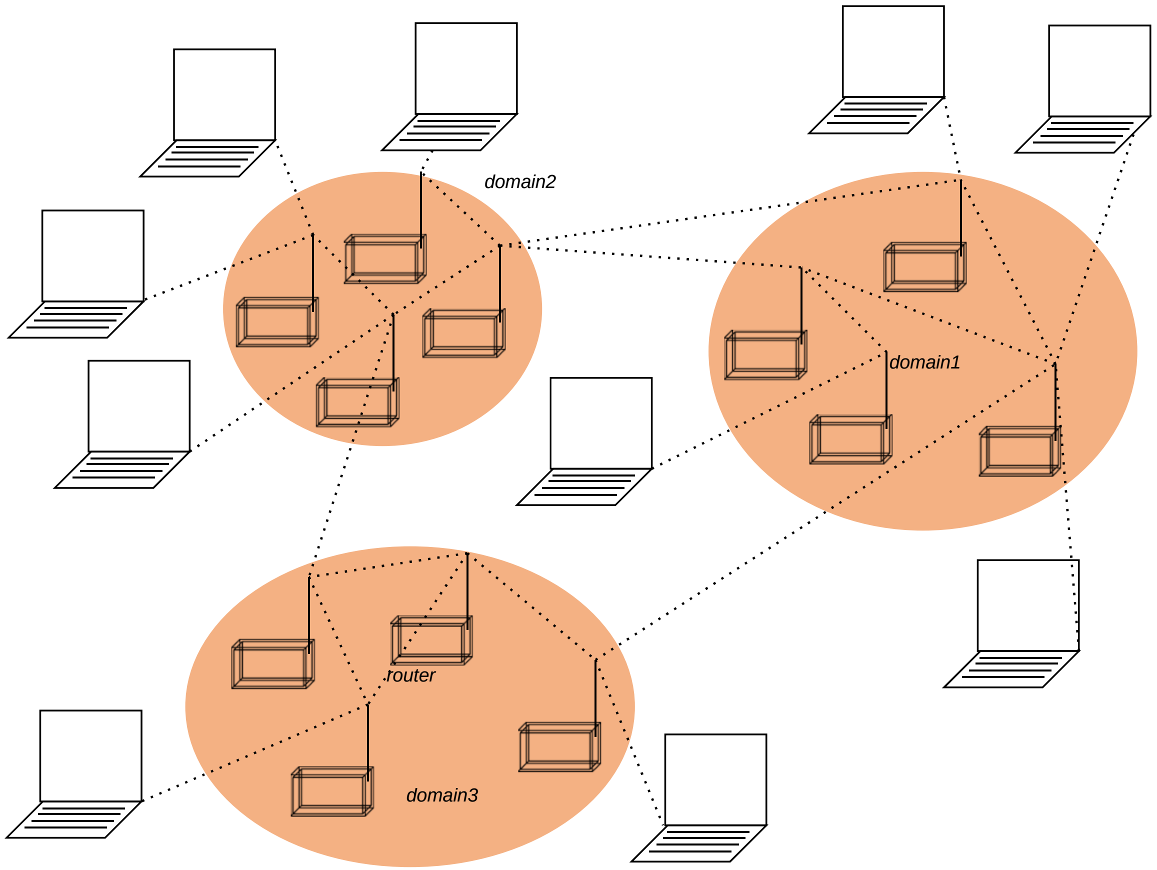
1.1.1 Graph Data: Information Nets



Citation networks and Maps of science

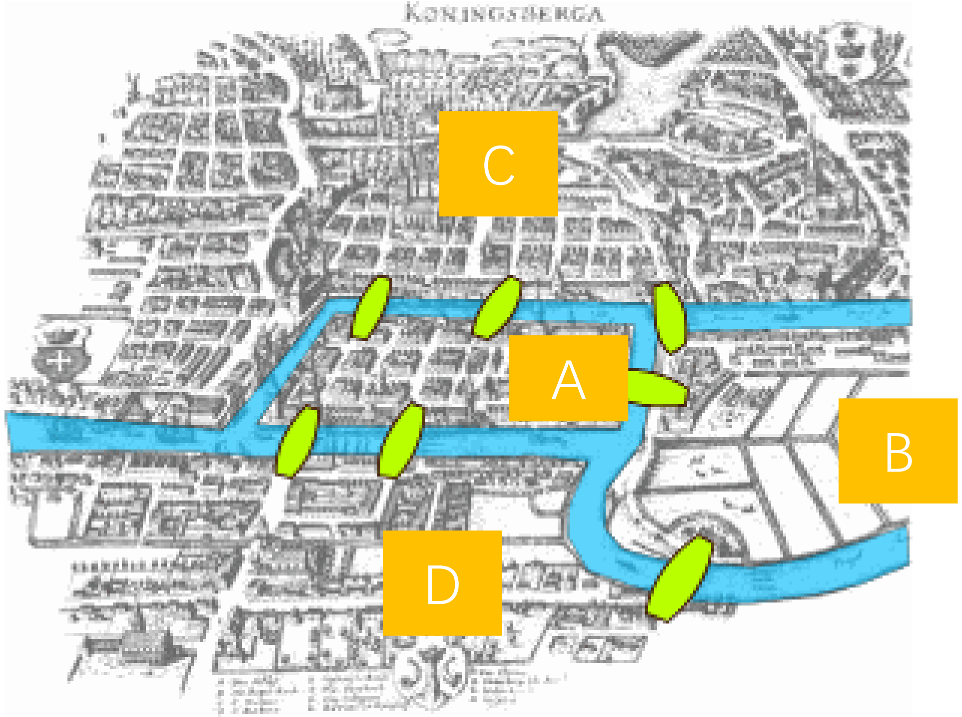
[Börner et al., 2012]

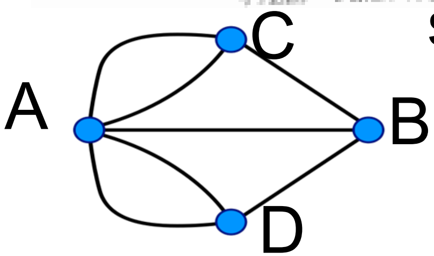
1.1.1 Graph Data: Communication Nets



Internet

1.1.1 Graph Data: Technological Networks



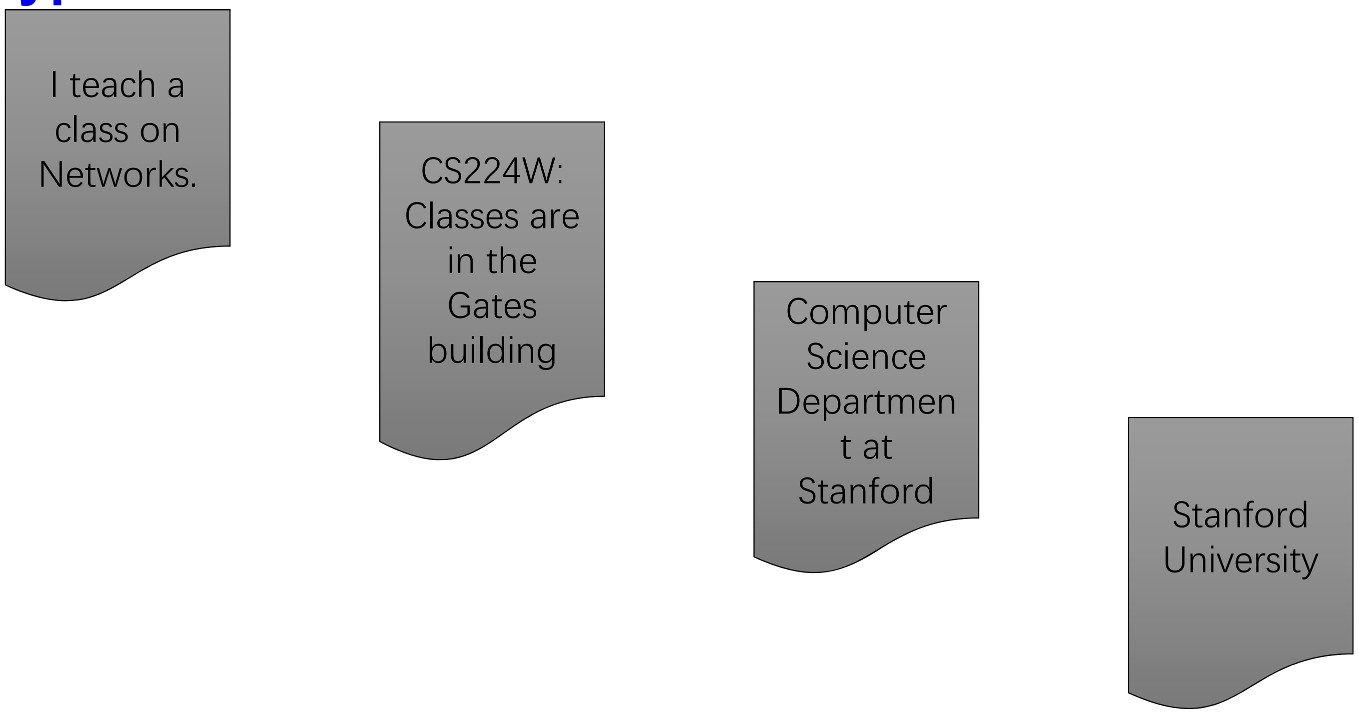


Seven Bridges of Königsberg
(哥尼斯堡七桥问题)
[Euler, 1735]
Return to the starting point by traveling each link of the graph once and only once.

供水网络
学习网络结构, 检测故障, 检测疾病爆发或污染等

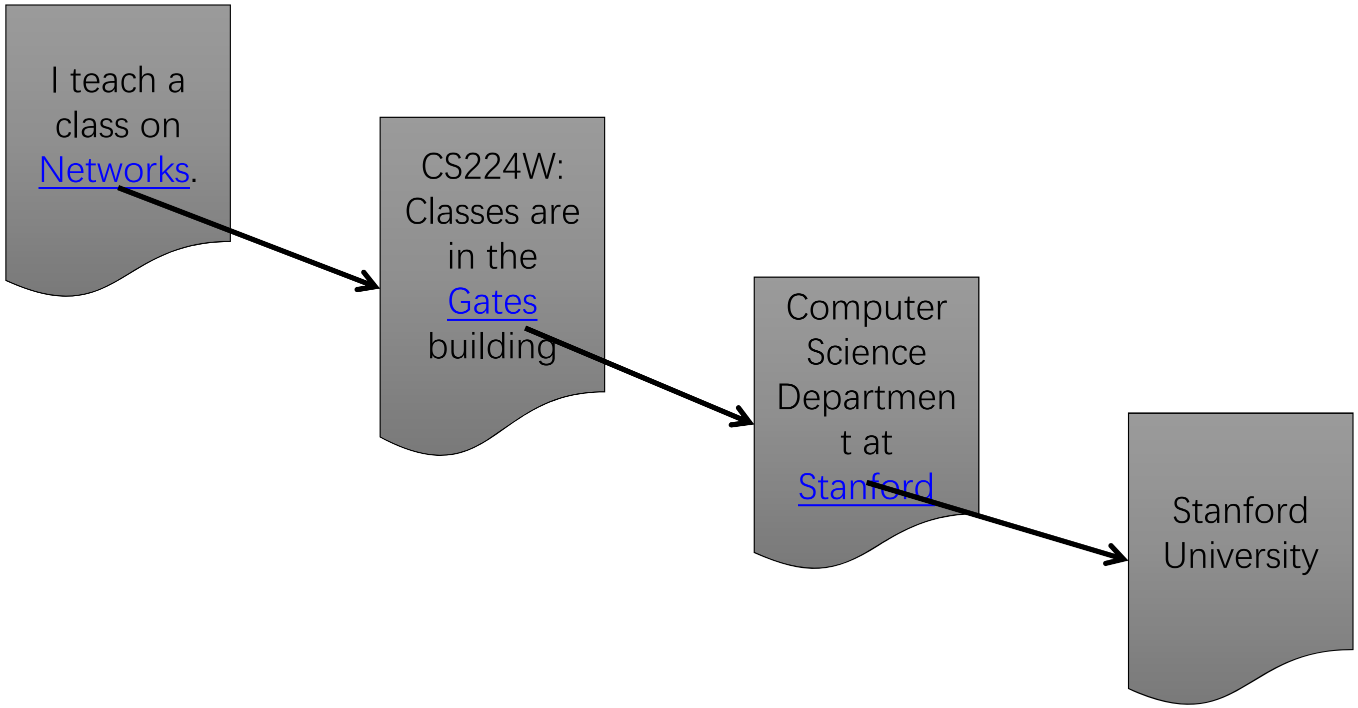
1.1.1 Web as a Graph

- **Web as a directed graph:**
  - **Nodes: Webpages**
  - **Edges: Hyperlinks**

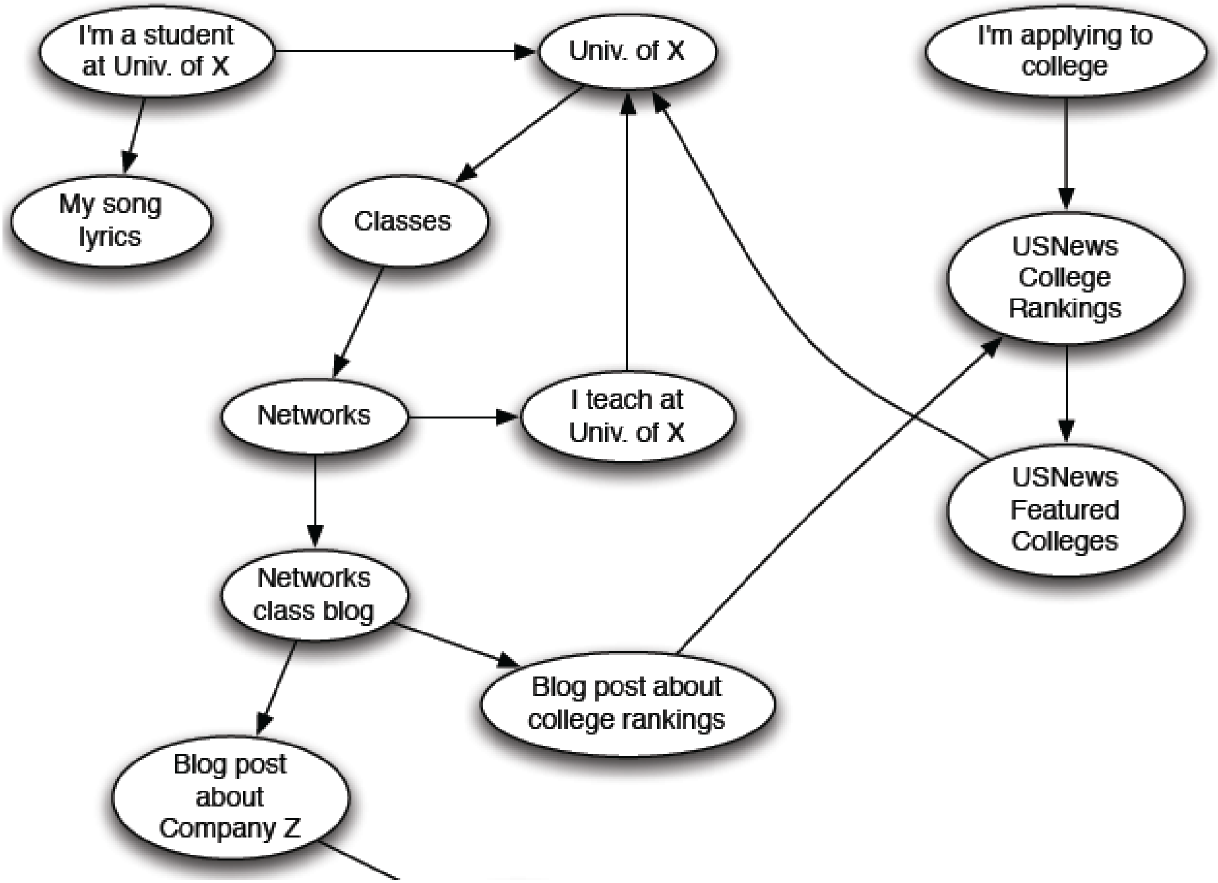


1.1.1 Web as a Graph

- **Web as a directed graph:**
  - **Nodes: Webpages**
  - **Edges: Hyperlinks**



1.1.1 Web as a Directed Graph



1.1.2 Broad Question

- **How to organize the Web?**
- **First try: Human created web directories**
  - Yahoo, DMOZ, LookSmart

- **Second try: Web Search**
  - Information Retrieval investigates:
    Find relevant docs in a small and trusted set
    - Newspaper articles, Patents, etc.
  - But: Web is huge, full of untrusted documents, random things, web spam, etc.

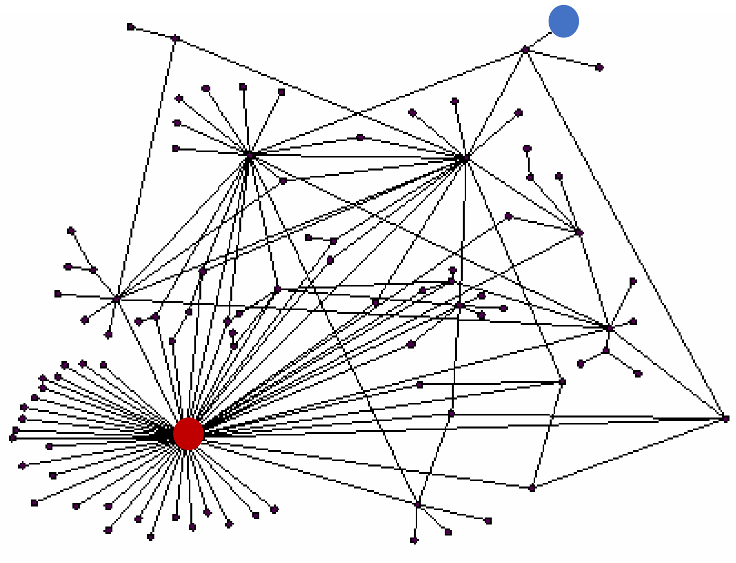
⟹ Need to find relevant and trusted webs!

1.1.2 Web Search: Two Challenges

- **Two challenges of web search:**
- **1) Web contains many sources of information.**
  - Who to “trust”?
  - Trick: Trustworthy pages may point to each other!
- **2) What is the “best” answer to query “newspaper”?**
  - No single right answer
  - Trick: Pages that actually know about newspapers might all be pointing to many newspapers

1.1.2 Ranking Nodes on the Graph

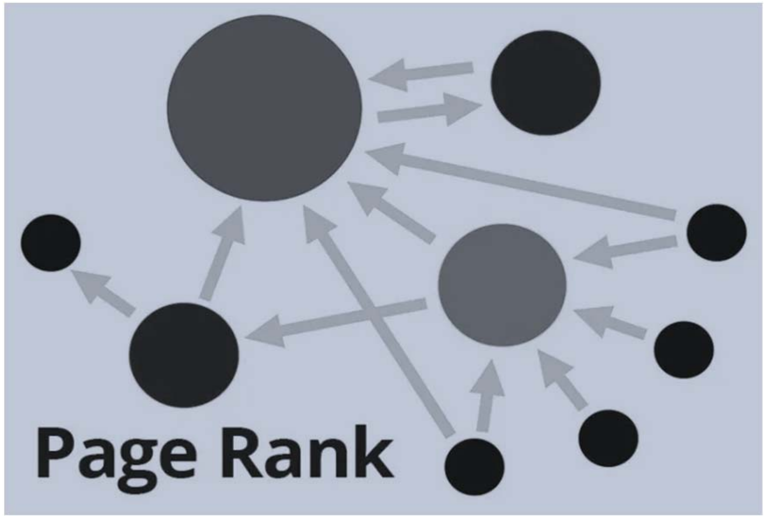
- **All web pages are not equally "important"**
  - www.joe-schmoe.com vs. www.stanford.edu



- **There is large diversity in the web-graph node connectivity. Let's rank the pages by the link structure!**

1.1.2 Link Analysis Algorithms

- **We will cover the following Link Analysis approaches for computing importance's of nodes in a graph:**
  - Section 1.2-1.4, PageRank
  - Section 1.5, Topic-Specific (Personalized) PageRank
  - Section 1.6, TrustRank
  - Section 1.7, Hubs and Authorities (HITS)



Section 1.2: PageRank,

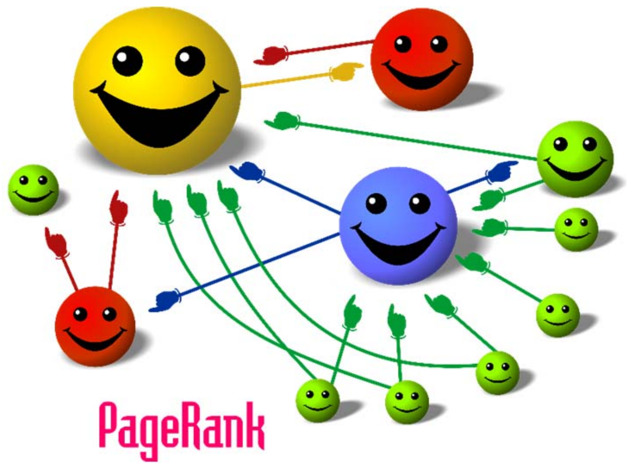
The “Flow” Formulation

Content

1.2.1 Links as Votes

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

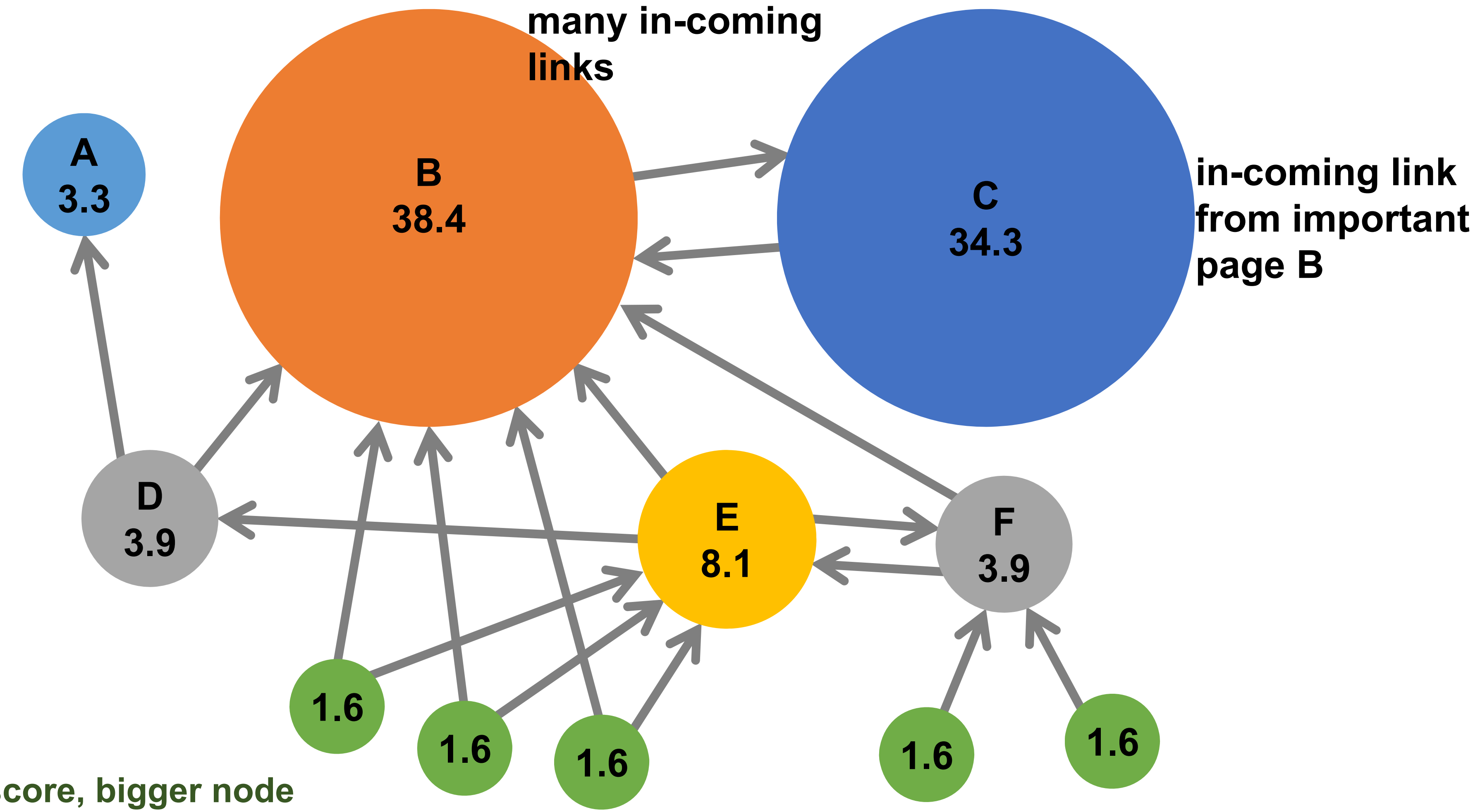
- **Idea: Links as votes**
  - **Page is more important if it has more links.**
  - **So In-coming links? Out-going links?**



- **Think of in-links as votes:**
  - www.stanford.edu has 23,400 in-links
  - www.joe-schmoe.com has 1 in-link

- **Are all in-links are equal?**
  - **Links from important pages count more**
  - Recursive question!

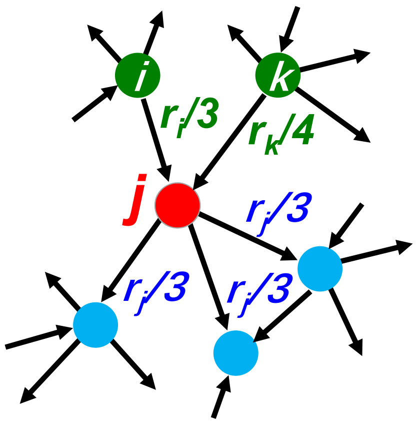
1.2.1 Example: PageRank Scores



Note: higher score, bigger node

1.2.1 Simple Recursive Formulation

- **Each link's vote is proportional to the importance of its source page**
  - If page j with importance r_j has n out-links, each link gets r_j/n votes
  - Page j's own importance is the sum of the votes on its in-links



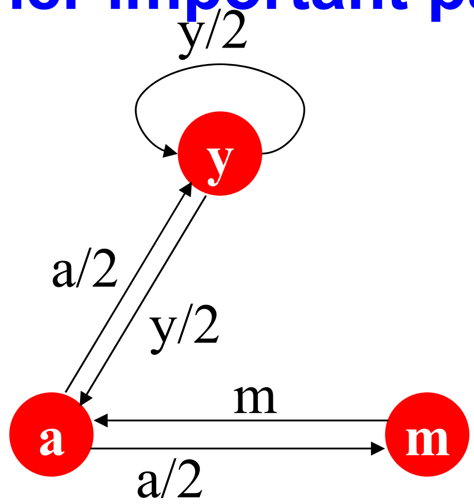
So, $r_j = r_i/3 + r_k/4$

1.2.1 PageRank: The “Flow” Model

- **A “vote” from an important page is worth more**
- **A page is important if it is pointed to by other important pages**
- **Define a “rank” r_j for page j**

$$r_j = \sum_{i \to j} \frac{r_i}{d_i}$$

d_i … out-degree of node i



The web in 1839

“Flow” equations:

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

1.2.1 Solving the Flow Equations

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

Flow equations:

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m$$
$$r_m = r_a/2$$

- **3 equations, 3 unknowns, no constants**
  - No unique solution
  - All solutions equivalent modulo the scale factor
- **Additional constraint forces uniqueness:**
  - $\boldsymbol{r_y + r_a + r_m = 1}$
  - **Solution:** $\boldsymbol{r_y = \frac{2}{5},\ r_a = \frac{2}{5},\ r_m = \frac{1}{5}}$
- **Gaussian elimination method (高斯消元法/高斯消去法) works for small examples, but we need a better method for large web-size graphs** ⇨ **We need a new formulation!**