# Analysis of Large Graphs: Link Analysis, PageRank

Mining of Massive Datasets Leskovec, Rajaraman, and Ullman Stanford University



### New Topic: Graph Data!

High dim.

Locality sensitive hashing

Clustering

Dimensional ity reduction

Graph data

**PageRank,** SimRank

Community Detection

Spam Detection

Infinite data

Filtering data streams

Web advertising

Queries on streams

Machine learning

**SVM** 

Decision Trees

Perceptron, kNN

Apps

Recommen der systems

Association Rules

Duplicate document detection

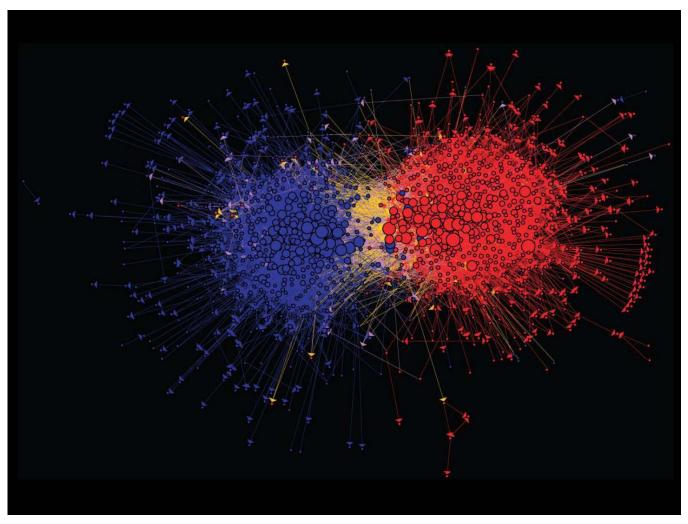
### **Graph Data: Social Networks**



#### Facebook social graph

4-degrees of separation [Backstrom-Boldi-Rosa-Ugander-Vigna, 2011]

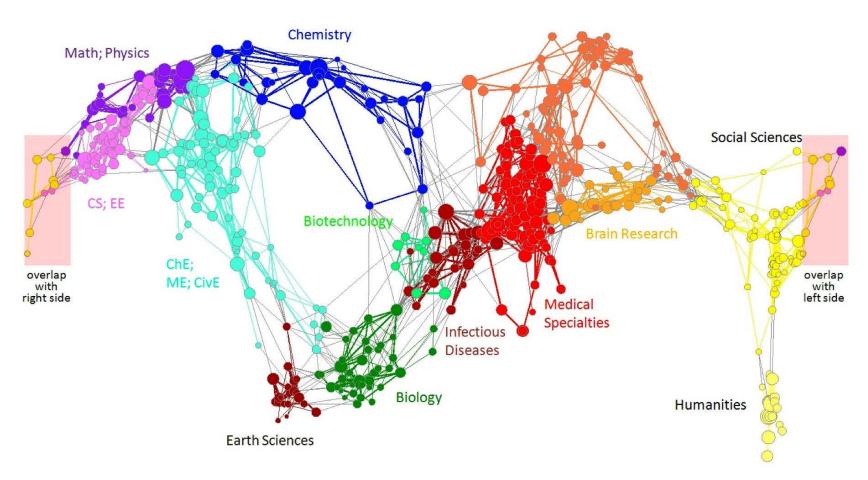
### Graph Data: Media Networks



Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

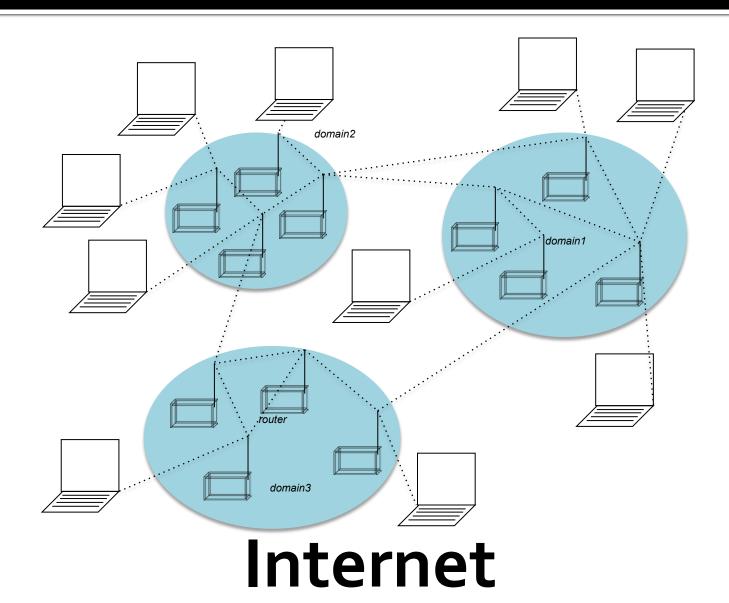
### **Graph Data: Information Nets**



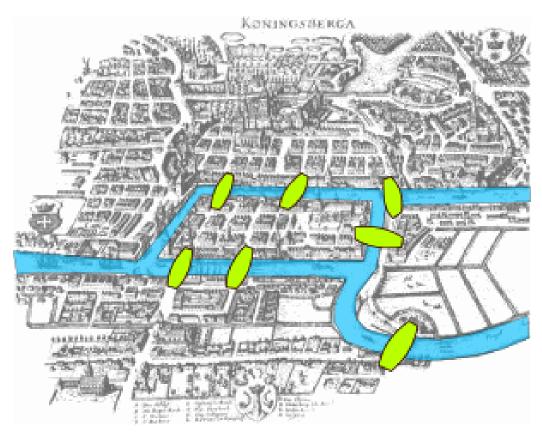
#### Citation networks and Maps of science

[Börner et al., 2012]

### **Graph Data: Communication Nets**



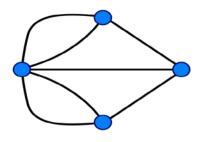
## Graph Data: Technological Networks



#### Seven Bridges of Königsberg

[Euler, 1735]

Return to the starting point by traveling each link of the graph once and only once.



### Web as a Graph

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

I teach a class on Networks.

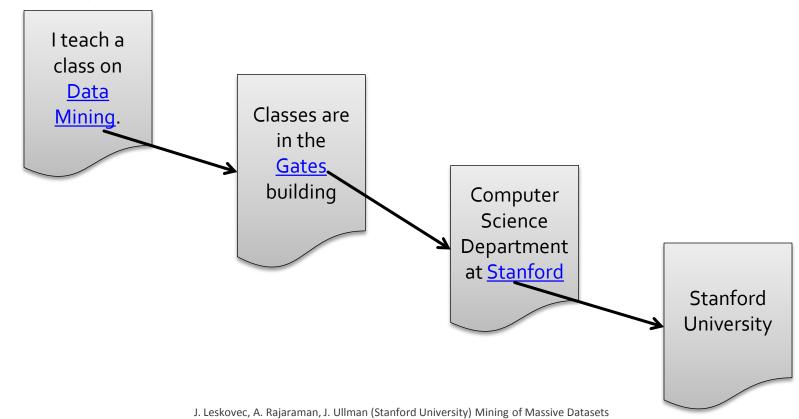
CS224W: Classes are in the Gates building

Computer
Science
Department
at Stanford

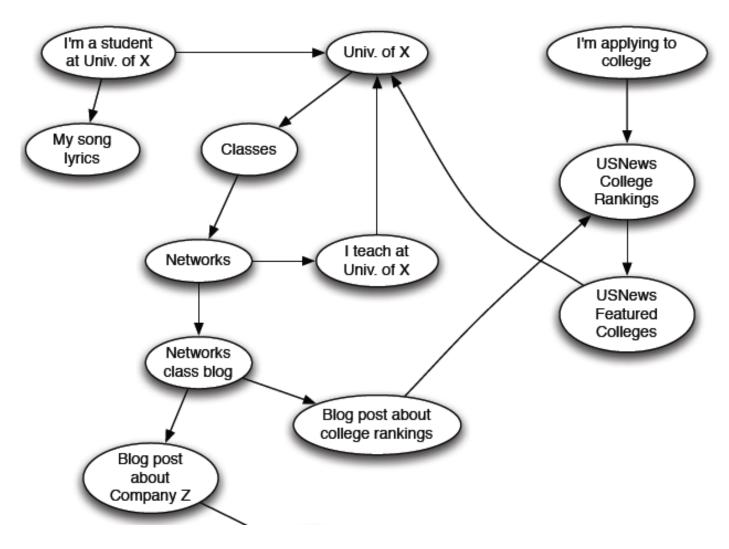
Stanford University

### Web as a Graph

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



### Web as a Directed Graph



### **Broad Question**

- How to organize the Web?
- First try: Human curated
   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.



### Web Search: 2 Challenges

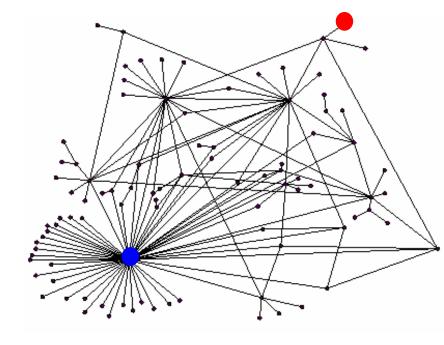
- 2 challenges of web search:
- (1) Web contains many sources of information Who to "trust"?
  - Trick: Trustworthy pages may point to each other!
- (2) What is the "best" answer to query "newspaper"?
  - No single right answer
  - Trick: Pages that actually know about newspapers might all be pointing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

### 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!



### Link Analysis Algorithms

- We will cover the following Link Analysis approaches for computing importances of nodes in a graph:
  - Page Rank
  - Hubs and Authorities (HITS)
  - Topic-Specific (Personalized) Page Rank
  - Web Spam Detection Algorithms