Lecture VIII – Link Analysis

Marina Meilă

Department of Statistics
University of Washington

STAT 548/CSE 547
Winter, 2022

1. (Directed) networks

2. Finding important pages on the web. PageRank

3. Personalized PageRank

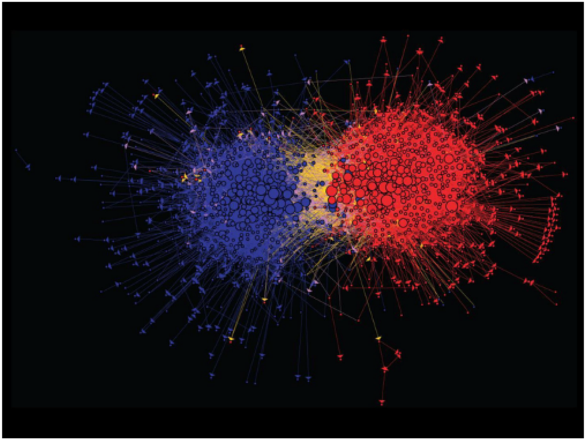
Graph Data: Social Networks

Facebook social graph

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

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 3

Graph Data: Media Networks



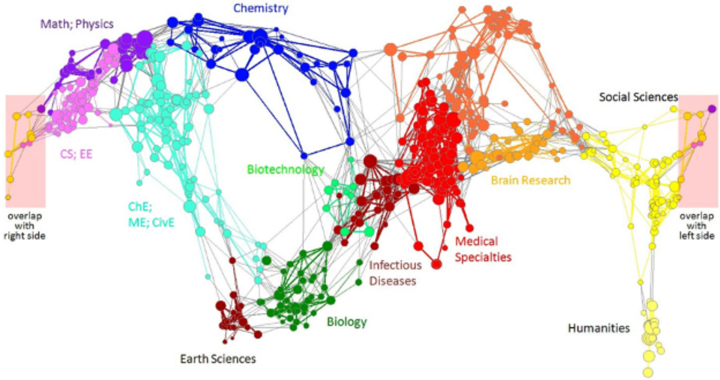
Connections between political blogs

Polarization of the network [Adamic-Glance, 2005]

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

4

Graph Data: Information Nets

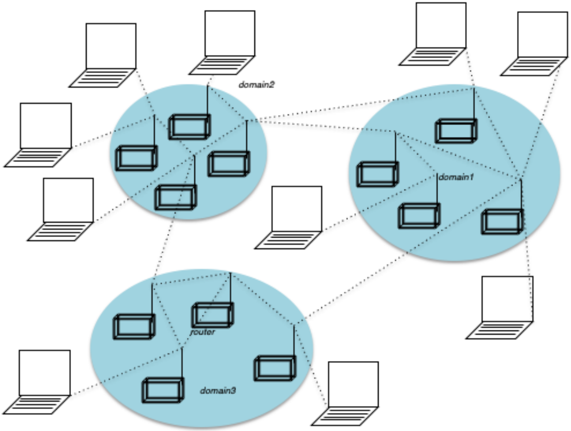


Citation networks and Maps of science

[Börner et al., 2012]

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 5

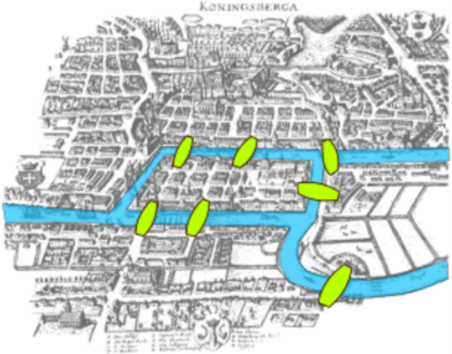
Graph Data: Communication Nets



Internet

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 6

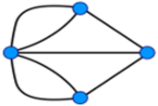
Graph Data: Technological Networks



Seven Bridges of Königsberg

[Euler, 1735]

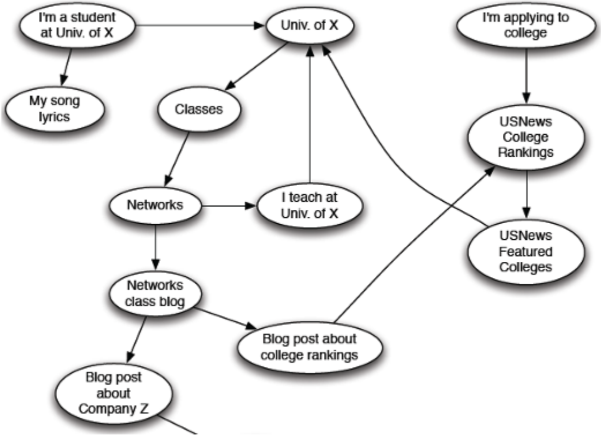
Return to the starting point by traveling each link of the graph once and only once.



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

7

Web as a Directed Graph



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 10

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.

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 11

Web Search: 2 Challenges

2 challenges of web search:

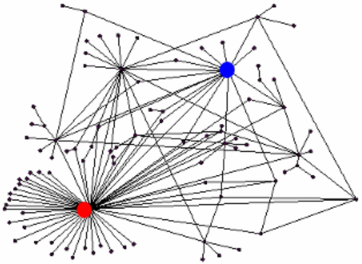
- **(1) Web contains many sources of information**
  Who to "trust"?
  - **Trick:** Trustworthy pages may point to each other!
- **(2) What is the "best" answer to query "newspaper"?**
  - No single right answer
  - **Trick:** Pages that actually know about newspapers might all be pointing to many newspapers

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 12

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!



- Hubs and authorities

J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org 13

PageRank idea

nitions

- A asymmetric adjacency matrix, $A_{ij} = 1$ is link $i \to j$
- $d_i = \sum_j A_{ij}$ **out-degree** of page i
- $r_i > 0$ the **importance/prestige** of page i (higher is better)

- PageRank principle: the importance of page j comes from the pages pointing to it

$$r_j = \sum_i A_{ij} \frac{r_i}{d_i} \tag{1}$$

- Division by d_i because prestige of i is equally divided between the pages i points to

PageRank idea

nitions

- A asymmetric adjacency matrix, $A_{ij} = 1$ is link $i \rightarrow j$
- $d_i = \sum_j A_{ij}$ **out-degree** of page i
- $r_i > 0$ the **importance/prestige** of page i (higher is better)

- PageRank principle: the importance of page j comes from the pages pointing to it

$$r_j = \sum_i A_{ij} \frac{r_i}{d_i} \quad (1)$$

- Division by d_i because prestige of i is equally divided between the pages i points to

- In matrix form $A \in \{0,1\}^{n \times n}$, $r \in [0, \infty)^n$

$$P = D^{-1}A \quad \textbf{transition matrix} \qquad r = P^T r \quad (2)$$

- r is eigenvector of P^T!

Markov chains recap

- P transition matrix, is stochastic matrix
- $P\mathbf{1} = \mathbf{1}$ hence $\mathbf{1}$ is (left) e-vector with e-value $\lambda_1 = 1$
- $\lambda_1 = 1$ is the largest e-value of P
- What is the **left e-vector** corresponding to $\lambda_1 = 1$?
- $\pi^T = \pi^T P$ the **stationary distribution** of P
- π always exists, always $\pi \succ 0$, π unique when P ergodic

- Mixture of random walks P, P', with probability $\beta \in (0, 1)$: $P^{mix} = \beta P + (1 - \beta) P'$

Algorithms for finding r

- Eigen-solver, firnd 1st principal e-vector of P^T
- **Power iteration**

  Input A
  Init $r_i = 1/n$

  1. Compute $d_{1:n}$
  2. Repeat until convergence
     1. $r_i \leftarrow r_i/d_i$ for all i $\mathcal{O}(n)$
     2. $r_j \leftarrow \sum_{i \to j} r_i$ for all j $\mathcal{O}(|E|)$

Problem with directed graphs

- strongly connected component (SCC)
- ergodic set

Teleporting Random Walk

- Solution: have a single SCC
- **Teleporting random walk** transition matrix

$$P^{tele} = \beta P + (1-\beta)\frac{1}{n}\mathbf{1}_{n\times n}$$

  with $\beta \in (0,1)$ (in practice $\beta = 0.8 - 0.9$)
- From each node i transition
  - on an outgoing edge uniformly at random, w.p. β (original r.w.)
  - to an arbitrary page, selected uniformly at random
- Sinks are special case, $\beta = 0$ for sink.
- Power iteration (simplified)
  1. $r \leftarrow \beta P^T r$ (original Power Iteration, rescaled by β)
  2. $r \leftarrow r + \frac{1-\beta}{n}$

Prestige of pages w.r.t a topic

- Problem: find pages most relevant to a topic
- Define topic by (small) set S of known pages on the topic
- S is **seed set**
- Teleporting $Uniform(S)$ instead of $Uniform$(all pages)
  - $Uniform(S)$ can be replaced with other distribution over S

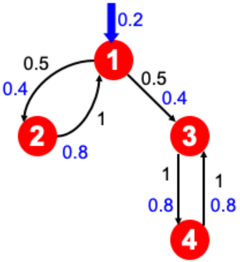
$$P^{PPR} \ = \ \beta P + (1-\beta)\frac{1}{|S|}\mathbf{1}_{[S]}$$

- Power iteration works similarly

Example PPR vs PageRank

Example: Topic-Specific PageRank

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]