

CPSC 340: Machine Learning and Data Mining

PageRank

Fall 2019

Web Search before Google

Multi Search university [Next! \[national parks\]](#)

10 results

Query: university
11 Results Returned
Showing Results From 0 to 10

Stanford University Homepage
http://www.stanford.edu/
74.79% 4K - 2591993 - 0103997

Stanford University Portfolio Collection
http://www.stanford.edu/home/administration/portfolio.html
65.78% 3K - 2591993 - 0103997

University of Illinois at Urbana-Champaign
http://www.uiuc.edu/
73.26% 15K - 2203096 - 0103997

Indiana University
http://www.indiana.edu/
68.38% 1K - 0903096 - 0103997

University of California, Irvine
http://www.uci.edu/
68.07% 3K - 2203096 - 0103997

University of Minnesota
http://www.umn.edu/
67.05% 8K - 2216996 - 0103997

Iowa State University Homepage
http://www.iastate.edu/
66.66% 3K - 2216996 - 0103997

The University of Michigan
http://www.umich.edu/
66.35% 1K - 2591993 - 0103997

Mississippi State University
http://www.msstate.edu/
66.35% 3K - 2591993 - 0103997

Northwestern University NUInfo
http://www.nyu.edu/
66.15% 3K - 2216996 - 0103997

next 10

Optical Physics at the University of Oregon
Oregon Center for Optics in Science and Technology. Department of Physics, University of Oregon, Eugene OR 97403. Research Groups: Carmichael Group....
<http://optich.uoregon.edu/> - size 1K - 16 Dec 96

Carnegie Mellon University - Campus Networking
Departments. Data Communications. Data Communications is responsible for installing and maintaining all on campus networking equipment and all of...
<http://www.net.cmu.edu/> - size 4K - 19 Aug 95

Wesleyan University Computer Science Group Home Page
Computer Science Group. Wesleyan University. Welcome to the home page of the Computer Science Group at Wesleyan University. We are administratively within.
<http://www.cs.wesleyan.edu/> - size 2K - 15 Apr 96

Keio University Shonan Fujisawa Campus (SFC)
B\$3\$N\$E2IEF\$F\$B\$6\$-9c9e9Q99 (B(SFC) \$B\$N (BWWW \$B99 \$B\$CmOU=q\$- (B \$B\$FIS\$G\$/\$@5\$5\$# (B. Nihongo | English. SFC \$B>pJs (B. | \$B9e9G9\$9\$*9e9? | *...
<http://www.sfc.keio.ac.jp/> - size 3K - 5 Feb 97

School of Chemistry, University of Sydney
The School of Chemistry. School of Chemistry, University of Sydney, NSW 2006 Australia International Phone: +61-2-9351-4504 Fax: +61-2-9351-3329 Australia.
<http://www.chem.su.oz.au/> - size 4K - 25 Feb 97

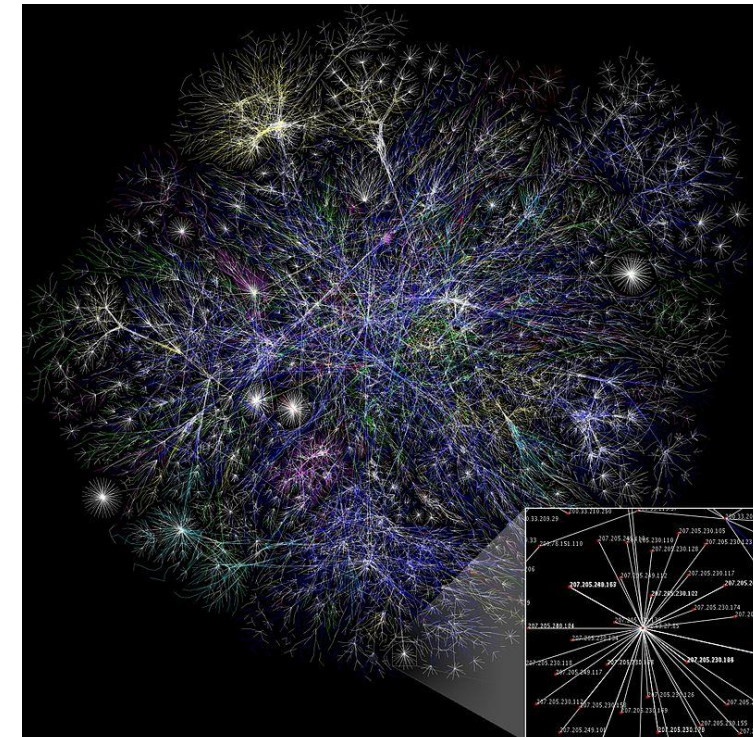
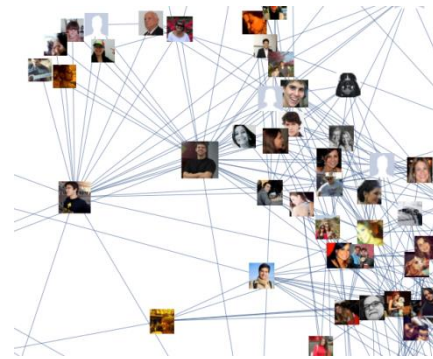
Mankato State University
The Campus Athletics, Campus Tour, Bookstore, Maps, Current Events... Admission & Registration Admissions, Financial Aid, Registrar's, Graduate...
<http://www.mankato.msus.edu/> - size 3K - 27 Nov 96

St. Ambrose University
Main Index: Academic Departments. Administrative Services. Campus News. Computing Services. Galvin Fine Arts Center. Internet Connections. Library...
<http://www.sau.edu/> - size 2K - 4 Feb 97

University of Washington ECSEL Projects

Unsupervised Graph-Based Ranking

- We want to rank “importance” based on graph between examples.
 - Every webpage is a node, and every web-link is an edge.
 - Every paper is a node, and every citation is an edge.
 - Every Facebook user is a node, and every “friendship” is an edge.



Unsupervised Graph-Based Ranking

- We want to rank “importance” based on graph between examples.
 - Every webpage is a node, and every web-link is an edge.
 - Every paper is a node, and every citation is an edge.
 - Every Facebook user is a node, and every “friendship” is an edge.
- Key idea: use links (edges) to predict importance of nodes.
- Many link analysis methods, usually with recursive definitions:
 - A journal is “influential” if it is cited by “influential” journals.
- We will discuss PageRank, Google’s original ranking algorithm.

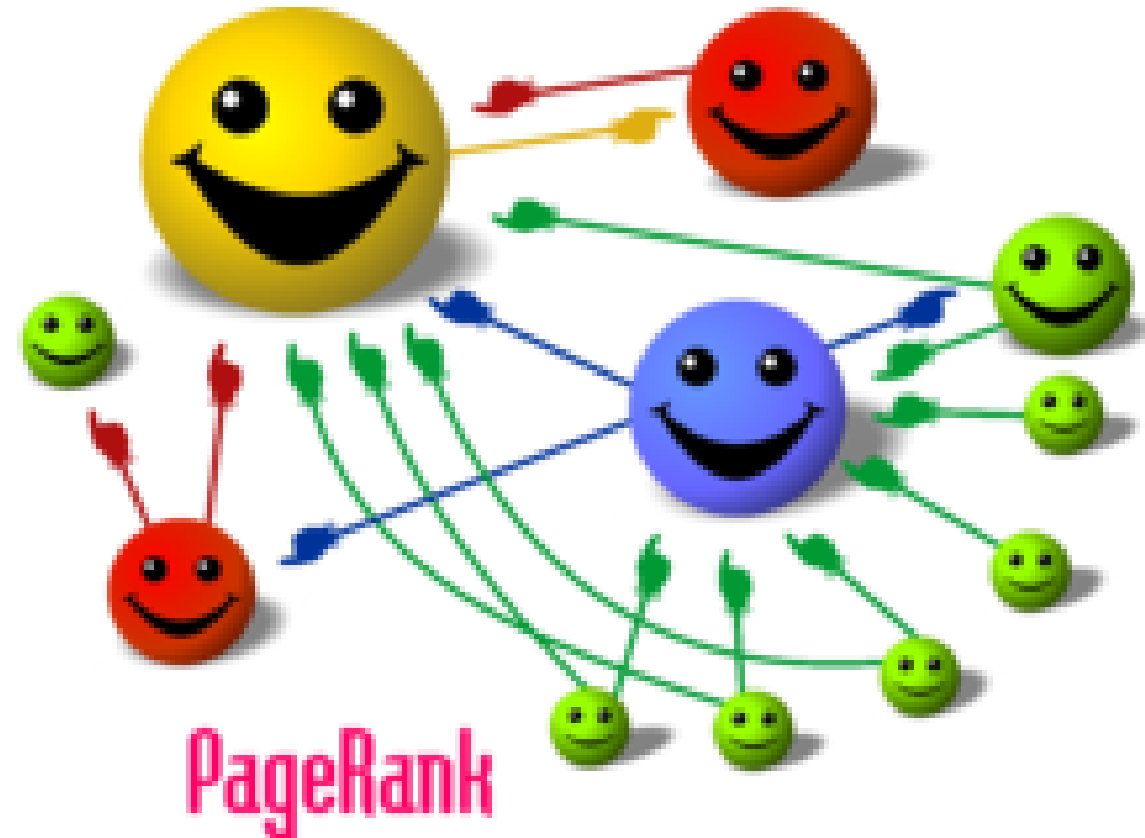
PageRank

- Wikipedia's cartoon illustration of PageRank:

- Large face => higher rank.

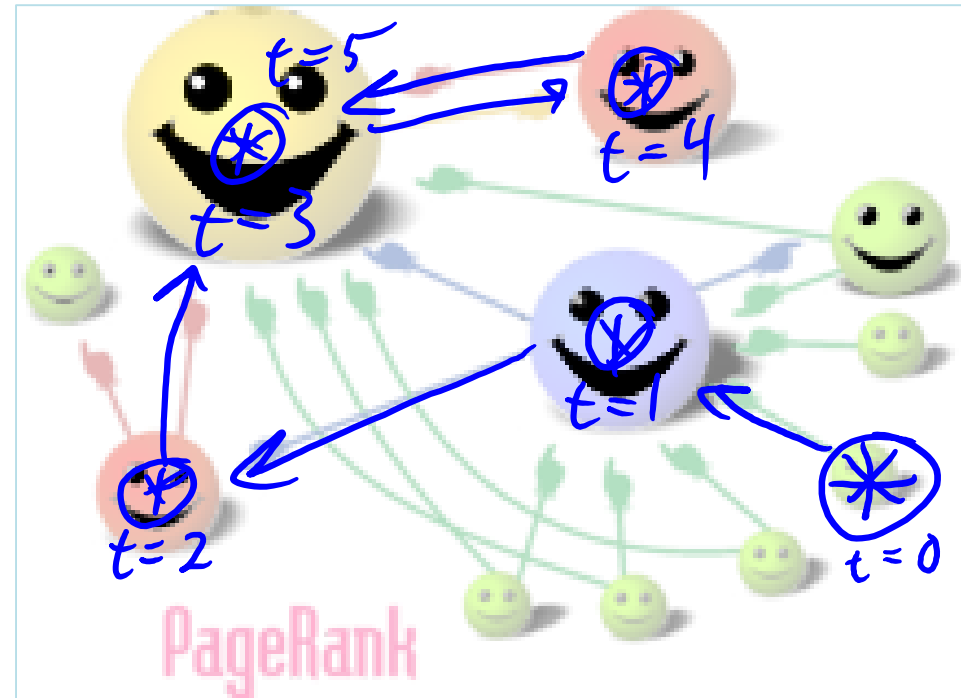
- Key ideas:

- Important webpages are linked from other important webpages.
 - Link is more meaningful if a webpage has few links.



Random Walk View of PageRank

- PageRank algorithm can be interpreted as a **random walk**:
 - At time $t=0$, start at a random webpage.
 - At time $t=1$, follow a random link on the current page.
 - At time $t=2$, follow a random link on the current page.
 -
- **PageRank**:
 - Probability of landing on page as $t \rightarrow \infty$.
- **Obvious problem**:
 - Pages with no in-links have a rank of 0.
 - Algorithm can get “stuck” in part of the graph.



Random Walk View of PageRank

- Fix: add **small probability of going to a random webpage** at time 't'.
- **Damped PageRank** algorithm:
 - At time $t=0$, start at a random webpage.
 - At time $t=1$:
 - With probability α (like 10%): go to a random webpage.
 - With probability $(1 - \alpha)$: follow a random link on the current page.
 - At time $t=2$, follow a random link on the current page.
 - With probability α : go to a random webpage.
 - With probability $(1 - \alpha)$: follow a random link on the current page.
- **PageRank**:
 - Probability of landing on page as $t \rightarrow \infty$.

PageRank Computation

- “Monte Carlo” method for computing PageRank:
 - Just run the random walk algorithm a really long time.
 - Count the number of times you visit each webpage.
 - Maybe include a “burn in” time at the start where you don’t count pages.
 - Can parallelize by using ‘m’ independent surfers.
 - Intuitive but **slow**.
- It can also be solved analytically with SVD:
 - But $O(n^3)$ for ‘n’ webpages.
- Google’s approach is the **power method**:
 - Repeated multiplication by transition matrix: $O(n\text{Links})$ per iteration.

Application: Game of Thrones

- PageRank can be used for other applications.
- “Who is the main character in the Game of Thrones books?”

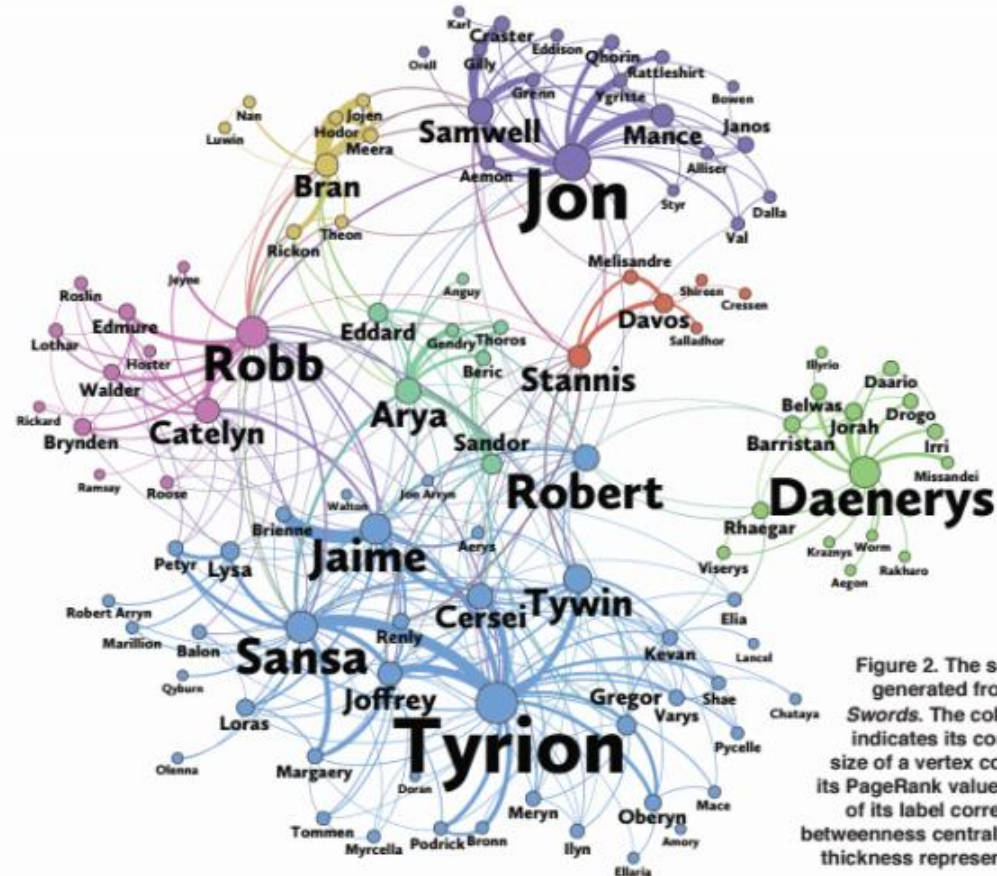


Figure 2. The social network generated from *A Storm of Swords*. The color of a vertex indicates its community. The size of a vertex corresponds to its PageRank value, and the size of its label corresponds to its betweenness centrality. An edge's thickness represents its weight.

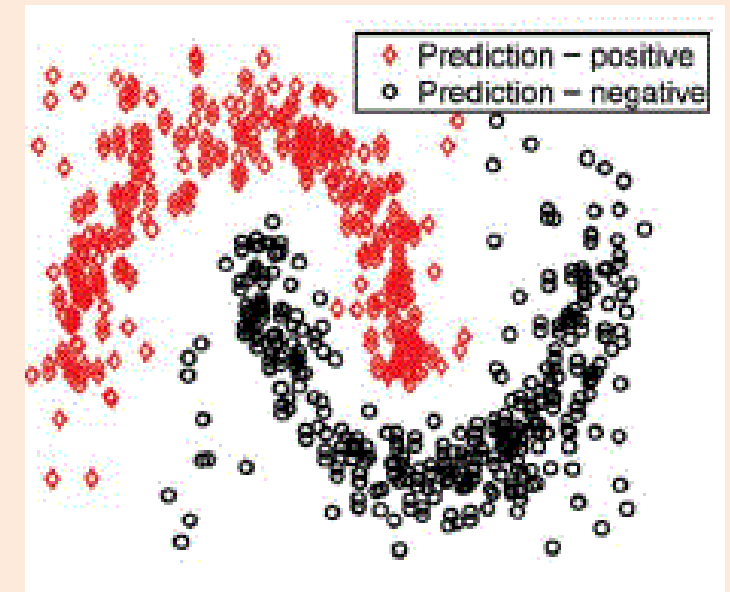
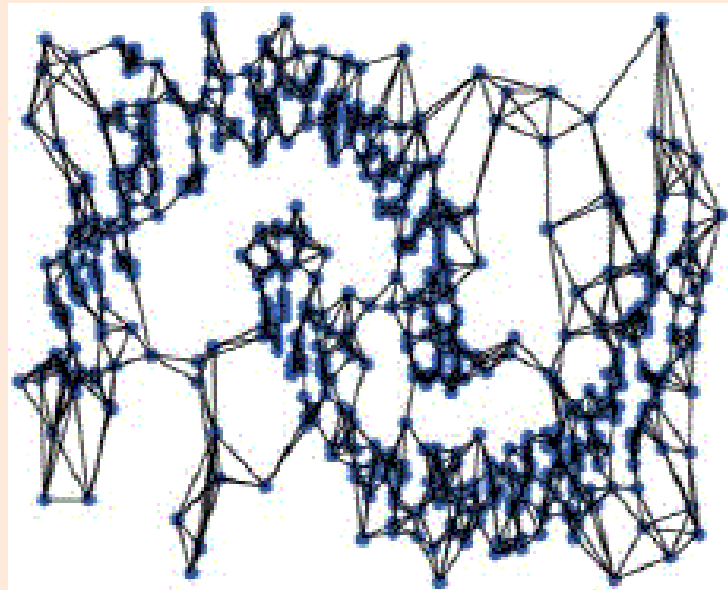
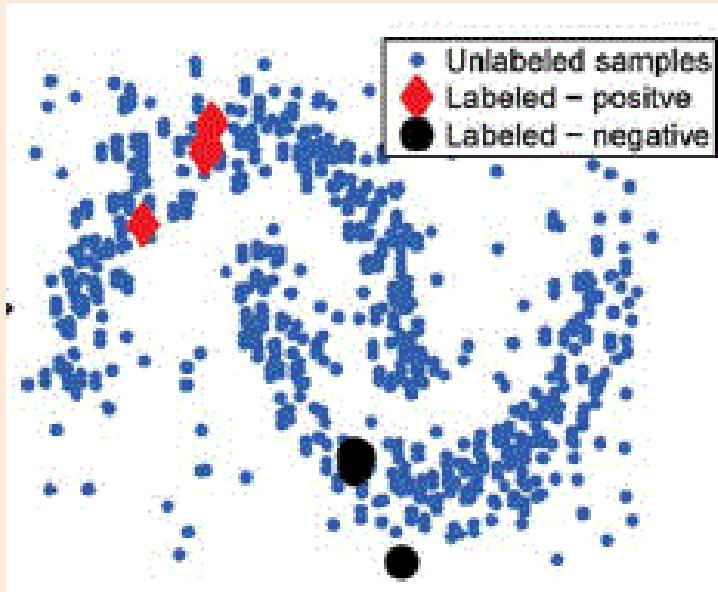
Ranking Discussion

- Modern ranking methods are more advanced:
 - Guarding against methods that exploit algorithm.
 - Removing offensive/illegal content.
 - Supervised and personalized ranking methods.
 - Take into account that you often only care about top rankings.
 - Also work on diversity of rankings:
 - E.g., divide objects into sub-topics and do weighted “covering” of topics.
 - Persistence/freshness as in recommender systems (news articles).

(pause)

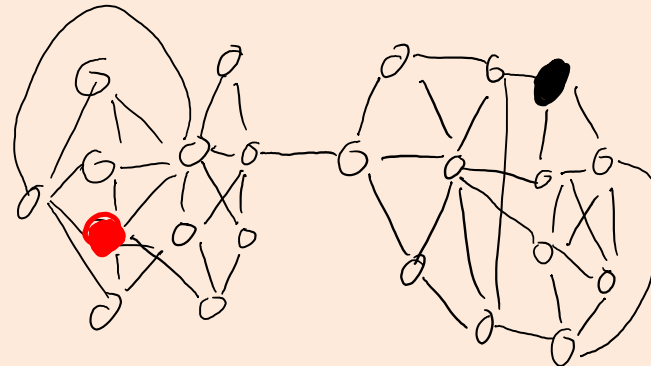
Previously: Graph-Based Semi-Supervised Learning

- Graph-based semi-supervised learning:
 - Define weighted graph on training examples:
 - For example, use KNN graph or points within radius ' ϵ '.
 - Weight is how 'important' it is for nodes to share label.



PageRank, Label Propagation, and Random Walks

- Standard graph-based SSL also has a **random walk** interpretation:
 - At time $t = 0$, set your state to the node you want to label.
 - At time $t > 0$, **move to a random neighbor**.
 - With probability proportional to w_{ij} (how much we want them to be similar).
 - If you land on a labeled node, choose that label for this “round”.
- Final **predictions are probabilities of outputting each label**.

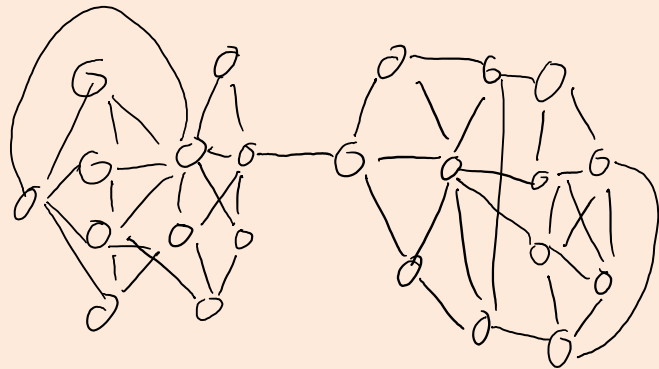


What else can we do with random walks?

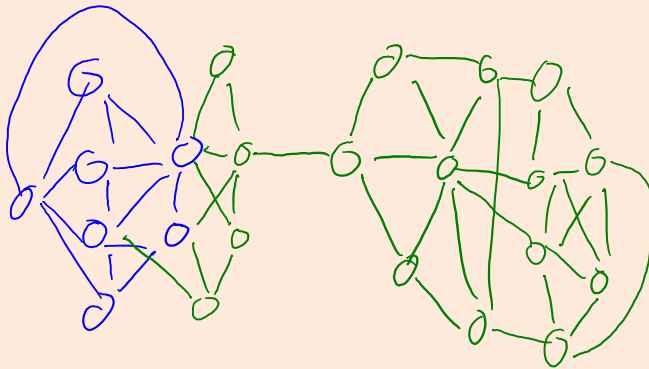
- We've discussed **random walks** for ranking and SSL.
 - Useful for problems defined on graphs.
 - We can convert from features to graphs using things like KNN graphs.
- Random walks for other tasks:
 - **Outlier detection** with **outrank**:
 - Examples with low PageRank are considered outliers (can **detect outlier clusters**).

What else can we do with random walks?

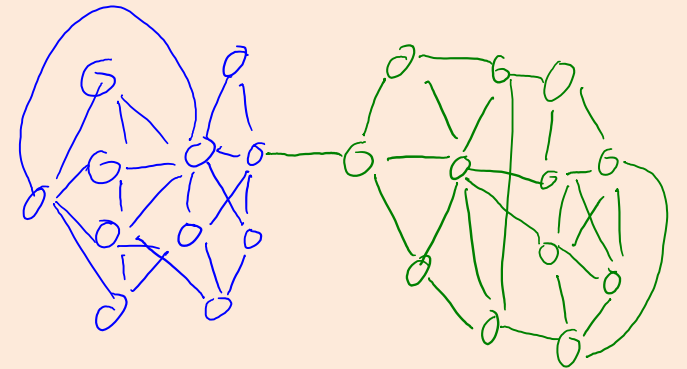
- We've discussed **random walks** for ranking and SSL.
 - Useful for problems defined on graphs.
 - We can convert from features to graphs using things like KNN graphs.
- Random walks for other tasks:
 - **Clustering** with **spectral clustering** (and “spectral graph theory”):
 - “If we start in cluster ‘c’, **random walk should tend to stay in cluster ‘c’**”.



Graph representation of data

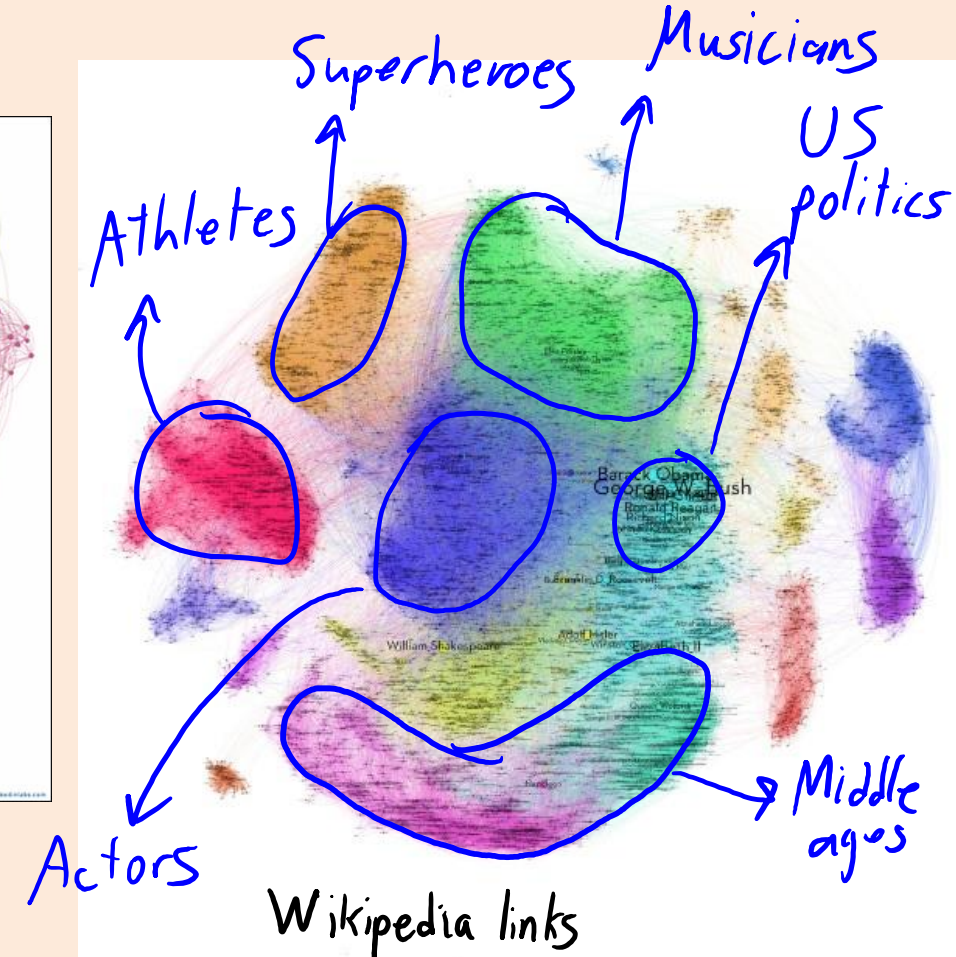
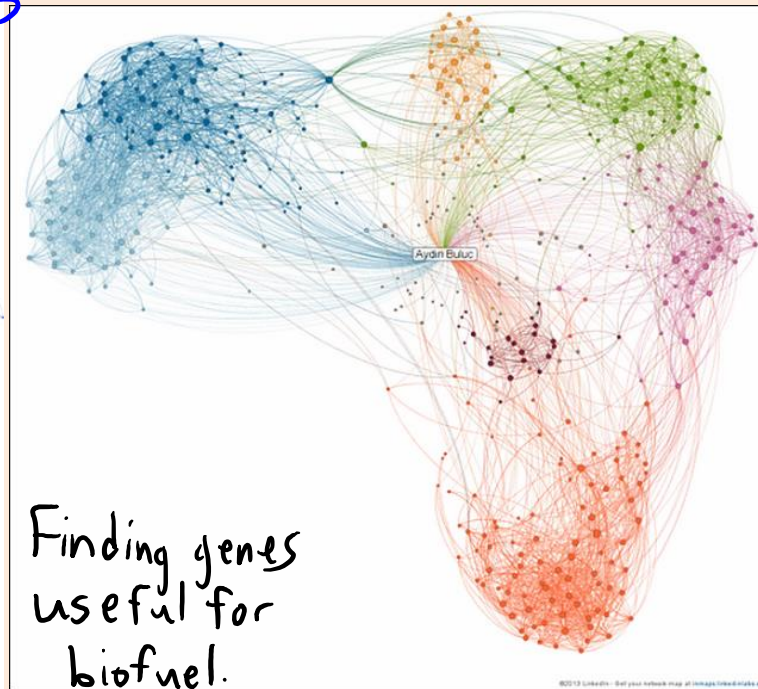
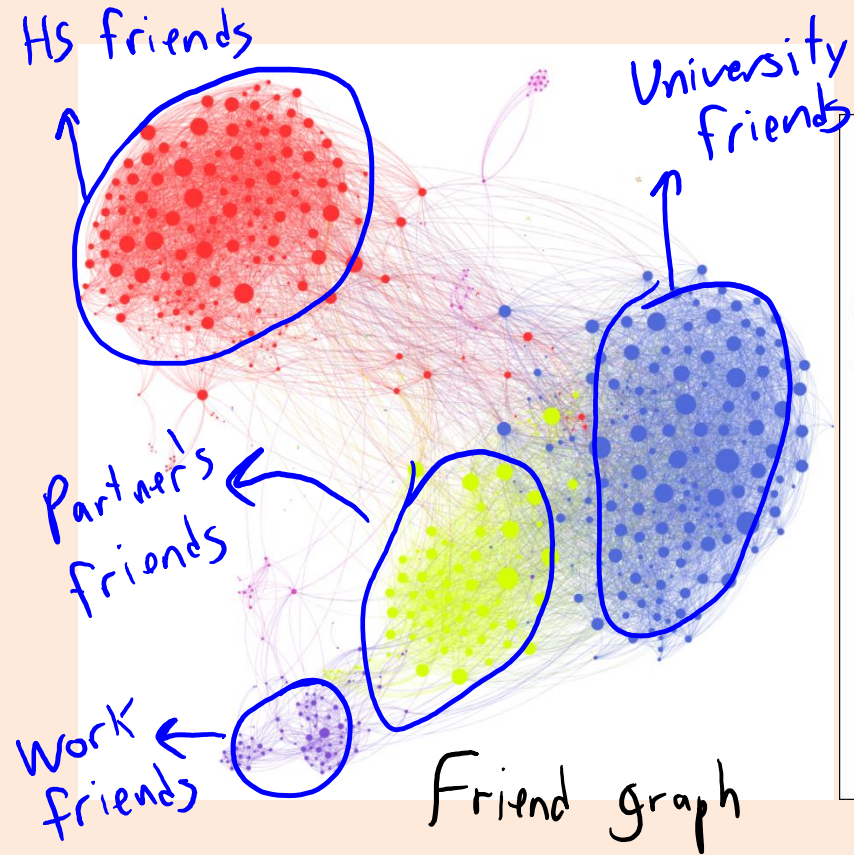


Bad clustering



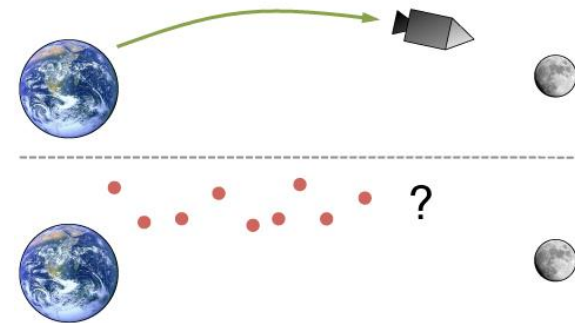
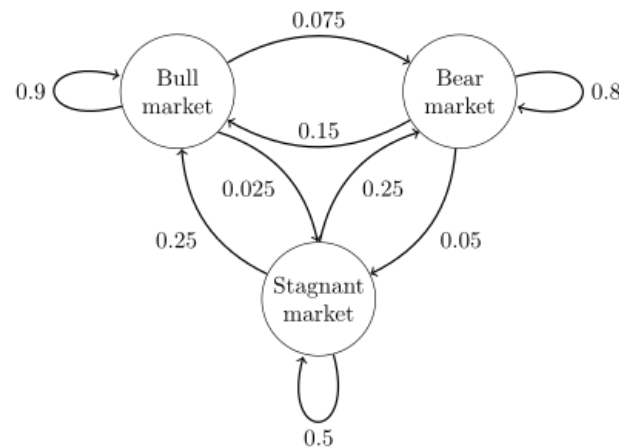
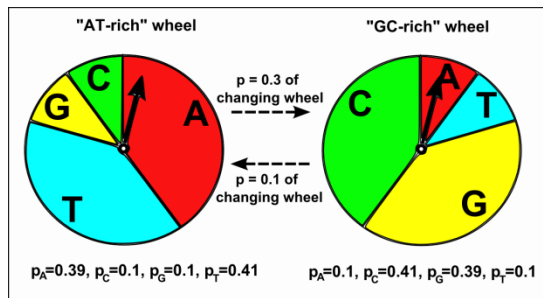
Good clustering.

Graph-Based Clustering Methods



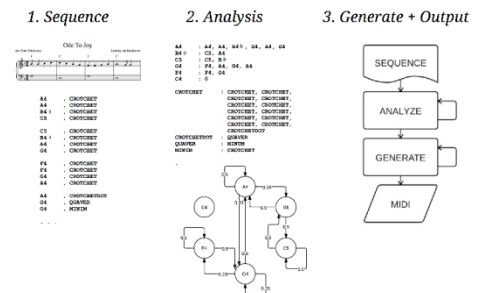
Markov Chains

- These **random walk** algorithms are special cases of **Markov chains**:
 - Most common **framework for modeling sequences**.
 - Bioinformatics, physics/chemistry, speech recognition, predator-prey models, language tagging/generation, computing integrals, economic models, flying airplanes, tracking missiles/players, modeling music.



Melody Generator

Generates a random melody using Markov Chains built from states and transitions extracted from an analysis of existing songs.



Summary

- Graph-based ranking uses links to solve ranking queries.
 - PageRank is based on a model of a random web user.