Goals of the course

- Presentation of some recent research on search engines and high dimensional data analysis.
- Not an overview: small number of topics for which we understand the mathematics and the computational methods.

Comparison to other topics/courses:

- Machine learning: amount of data makes a difference, goals are different: indexing, searching more modest than machine learning.
- Statistics:
  - similar goals: what is interesting about the dataset.
  - large amount of data.
  - Question: less sophisticated.
- Graph theory
  - Graph: mathematical model to represent relationships.
Example of basic questions:

1. Find structure in the collection of documents on the web.
2. Find topics: words have the same meaning.
3. Related/similar documents

Basic representation:

\[ A(d, w) = \begin{cases} 1 & \text{if document } d \text{ contains word } w \end{cases} \]

Market basket data

\[ \text{customer } i \xrightarrow[]{\text{product}} \begin{pmatrix} A(i,j) \\ 1 \\ 1 \end{pmatrix} \]

\( A(i,j) = 1 \text{ if customer } i \text{ purchased product } j \)

Example: recommendation system:

Netflix, Amazon, etc.

Gene expression data:

Column = gene
Row = disease/mouse

\[ A(i,j) \text{: expression level of gene } j \text{ in mouse } i \]
Web usage data:

Column $j$: web page
Row $i$: IP/USER/SESSION

$A(i,j) = 1$ if IP $i$ accessed page $j$

Web structure data:

Column, row: page
$A(i,j) = 1$ if page $i$ points to page $j$

Not a symmetric matrix!

How do we analyze this data?

- Eigenvalues of $A$
- PageRank: Brin & Page
- Kleinberg's HITS (hubs & authorities)
Introduction to search engines

- Steps for search:
  1) Crawl the web, create database
  2) Answer query
  3) Two ideas to rank answers came from exploiting link structure:
     - pagerank
     - hits

- pagerank: give a score to every page.