05/17/2022

Week 8 Section Review of PCA

Zichen Chen

Content

- Covariance
- Principal Component Analysis

- Demo

Covariance

- Correlated Variables
  - 2 variables are highly correlated —> Pearson's Correlation Coefficient to measure

- Variance
  - A measure of how spread out the data is
  - In 2D, you can measure variance in x-dim (x-variance) and in y-dim (y-variance)

Covariance cont'd

- How much one column (i.e. vector) of numbers varies with another
- Similar to average of the sum of the squares of the coordinates
- Correlation measures [-1, 1]
- Covariance (-∞, +∞)

single observed value of dependent variable

mean of all values of independent variable

single observed value of independent variable

total count of sample values

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

mean of all values of independent variable

population count minus one (Bessel's Correction)

α

Calculating the Covariance

- Step 1: Calculate the Sample Mean
  - for both **variable**

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{3.15 + 3.47 + \ldots + 6.16}{10} = \boxed{4.62}$$

$$\bar{x} = \frac{10 + 11 + 12 + \ldots + 14}{10} = \boxed{5.50}$$

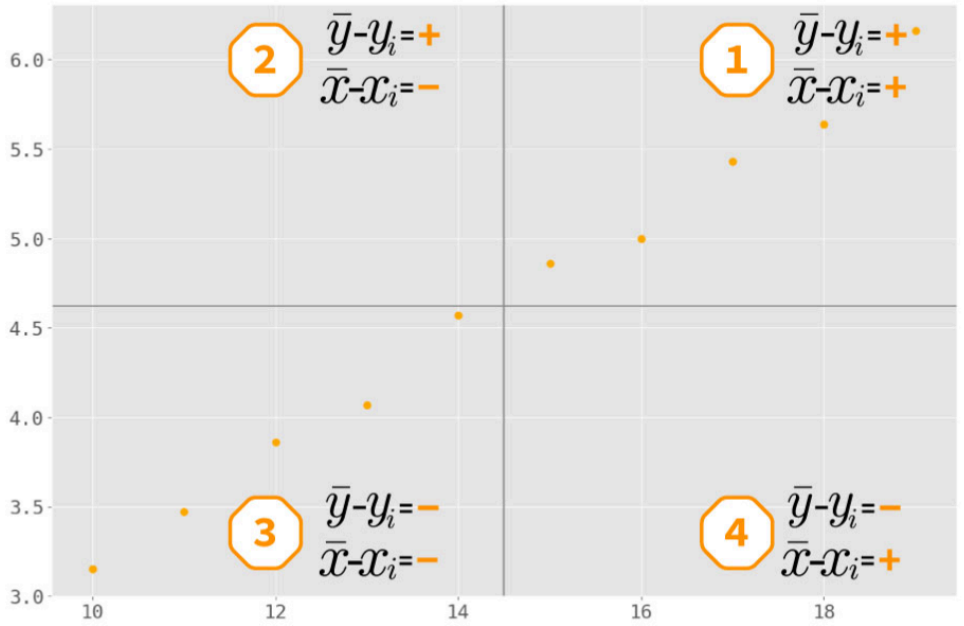
Observation Number	Predictor (X)	Response (Y)
1	10	3.15
2	11	3.47
3	12	3.86
4	13	4.07
5	14	4.57
6	15	4.86
7	16	5
8	17	5.43
9	18	5.64
10	19	6.16

Calculating the Covariance

- Step 2 (optional): Calculate **Signs** of Sample Mean Relationships
  - calculating the sign (positive vs. negative) of the relationship between each of our sample **variables** and their respective **mean**
  - describes whether the observed value of one variable might **increase** vs. **decrease** in relation to the other.

Calculating the Covariance

- Step 2 (optional): Calculate **Signs** of Sample Mean Relationships
  - X and Y is a positive linear relationship



Quadrant	$y_i - \bar{y}$	x_i-x	$(y_i - \bar{y})(x_i$-x)
1	+	+	+
2	+	-	-
3	-	-	+
4	-	+	-

Calculating the Covariance

- Step 3: Calculate the Covariance
  - a **positive linear relationship** between the values of x and y.
  - summing the results of the $(y_i - \bar{y})(x_i - \bar{x})$ column values —> positive value of ~26.61

Calculating the Covariance

- n=10, observations
- Product of each observed value relative to sample mean

$$Cov(x,y) = \frac{\Sigma (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

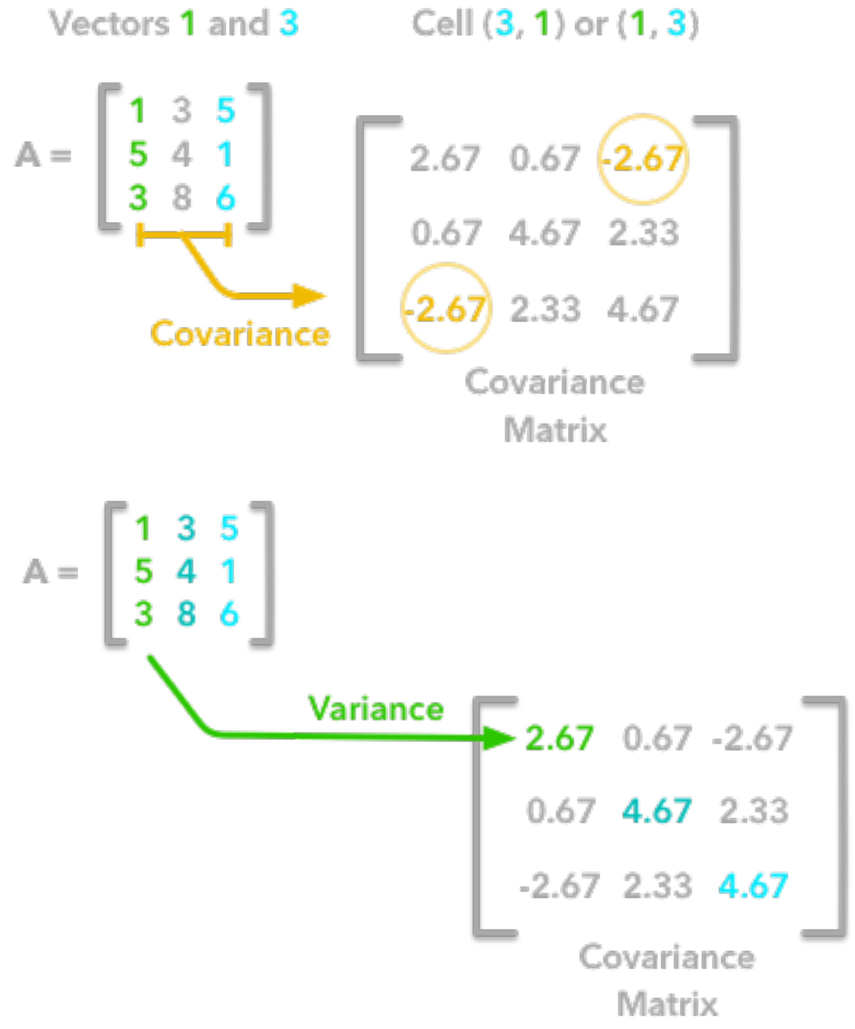
Observation Number	Predictor (X)	$x_i-\bar{x}$	Response (Y)	$y_i - \bar{y}$	$(y_i - \bar{y})(x_i-\bar{x})$
1	10	4.5	3.15	-1.471	-6.6195
2	11	5.5	3.47	-1.151	-6.3305
3	12	6.5	3.86	-0.761	-4.9465
4	13	7.5	4.07	-0.551	-4.1325
5	14	8.5	4.57	-0.051	-0.4335
6	15	9.5	4.86	0.239	2.2705
7	16	10.5	5	0.379	3.9795
8	17	11.5	5.43	0.809	9.3035
9	18	12.5	5.64	1.019	12.7375
10	19	13.5	6.16	1.539	20.7765

- covariance of x and y ~2.96

Covariance Matrix

- cov(x, y) —> The covariance of (column) vectors xi and xj

$$\begin{array}{cc} & \begin{array}{ccc} x & \quad y & \quad z \end{array} \\ \begin{array}{c} x \\ y \\ z \end{array} & \begin{bmatrix} var(x) & cov(x,y) & cov(x,z) \\ cov(x,y) & var(y) & cov(y,z) \\ cov(x,z) & cov(y,z) & var(z) \end{bmatrix} \end{array}$$



Principle Component Analysis (PCA)

- The process of finding the principal components of a set of data (matrix) and using only the first few principal components to explain the data outcomes and ignoring the rest (i.e. variable reduction)

- The principal components are **eigenvectors** of the data's **covariance** matrix

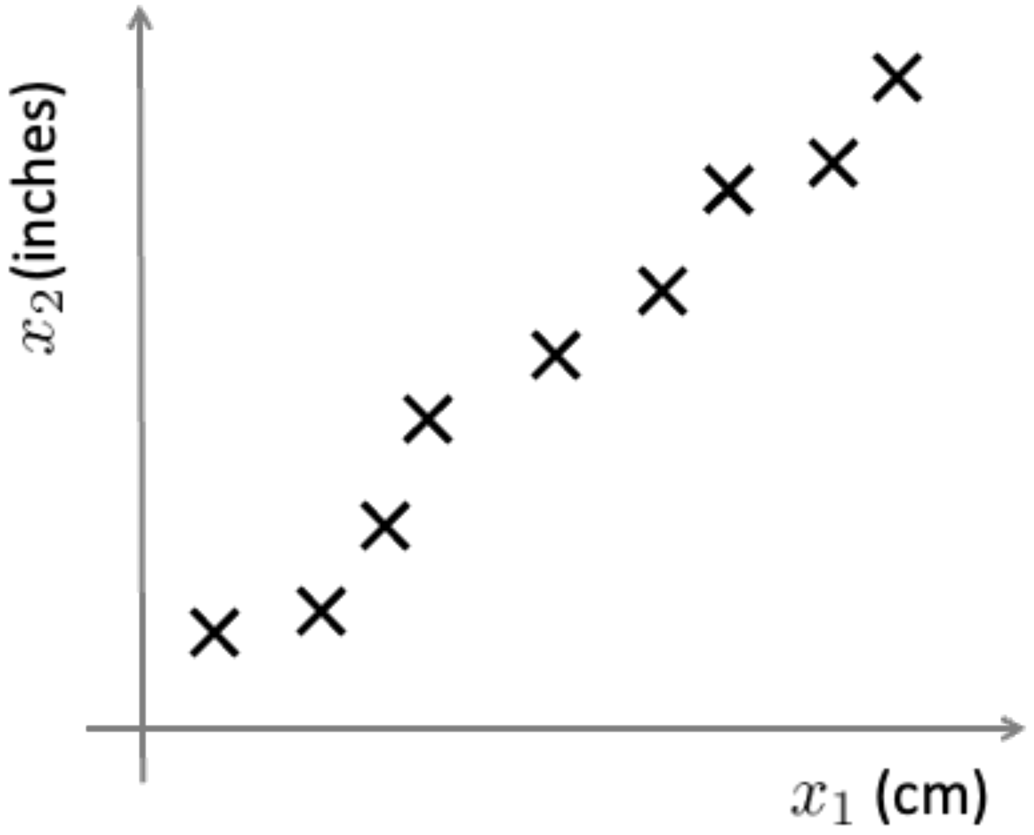
Apply PCA …

- **Clustering**
  - One way to summarize a complex real-valued data point with a single categorical variable
- **Dimensionality reduction**
  - Another way to simplify complex high-dimensional data
  - Summarize data with a lower dimentional real valued vector

Given data points in d dimensions
Convert them to data points in r<d dimensions
With minimal loss of information

Data Compression

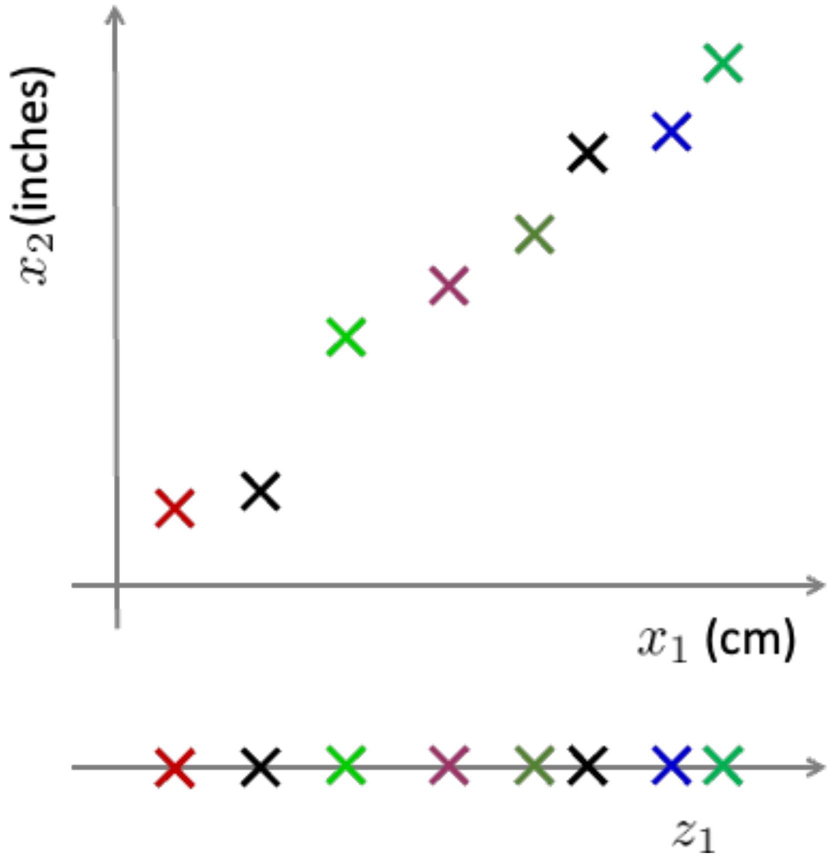
- Reduce data from 2D to 1D



Data Compression

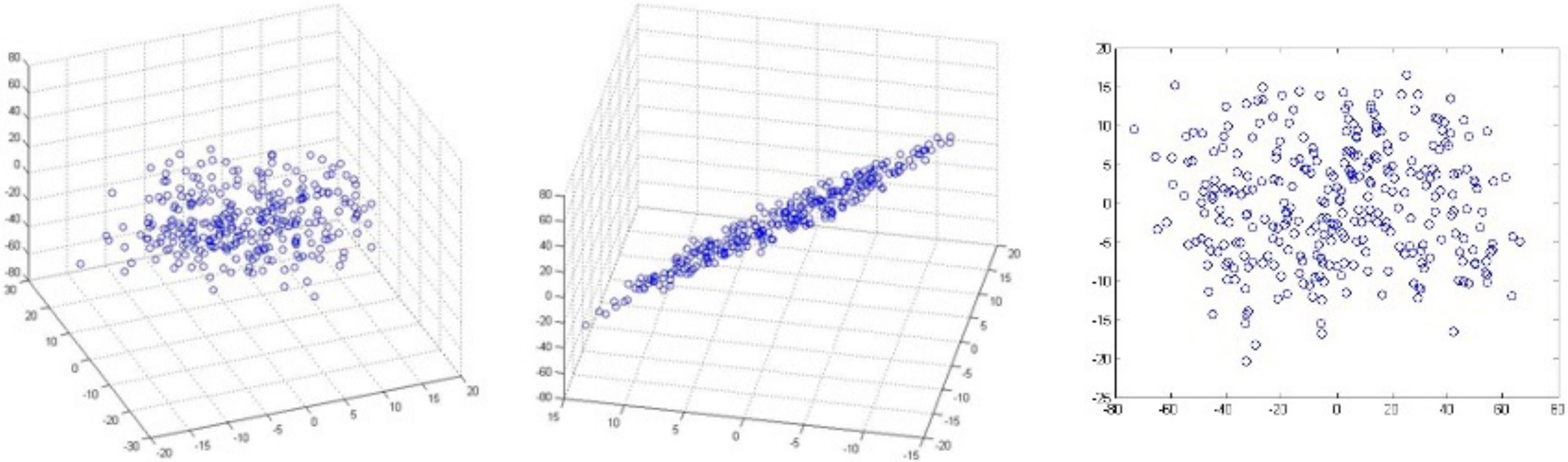
- Reduce data from 2D to 1D

$$
\begin{aligned}
x^{(1)} &\rightarrow z^{(1)} \\
x^{(2)} &\rightarrow z^{(2)} \\
&\vdots \\
x^{(m)} &\rightarrow z^{(m)}
\end{aligned}
$$

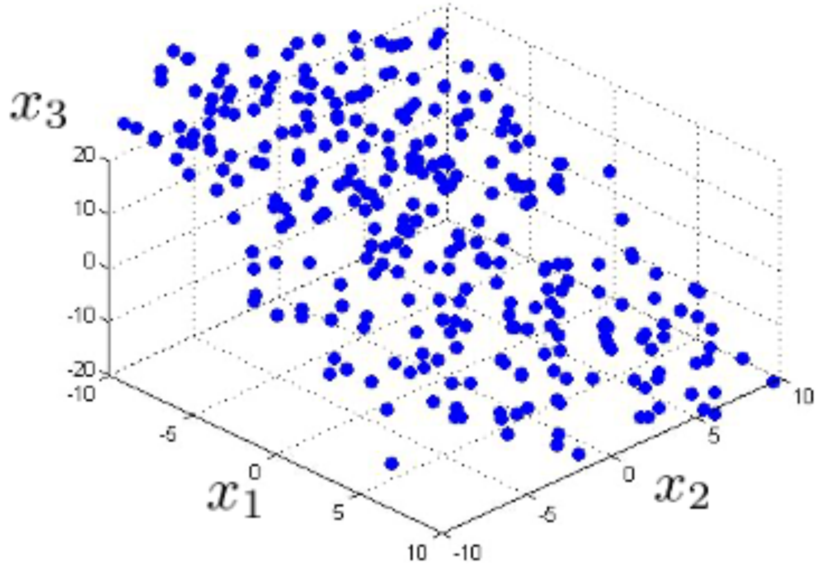


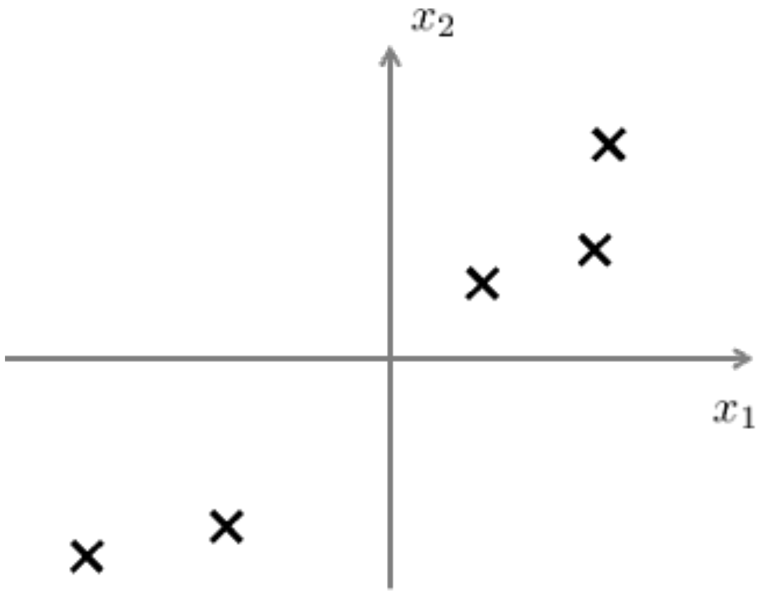
Data Compression

- Reduce data from 3D to 2D



Problem formulation





- Reduce from n-dimension to k-dimension: Find k vectors $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ (directions) onto which to project the data so as to minimize the projection error

PCA process

Goal: Find r-dim projection that best preserves variance

1. Compute mean vector μ and covariance matrix Σ of original points
2. Compute eigenvectors and eigenvalues of Σ
3. Select top r eigenvectors
4. Project points onto subspace spanned by them:

$$y = A(x - \mu)$$

   where y is the new point, x is the old one, and the rows of A are the eigenvectors

Application

PageRank - THE $25,000,000,000* EIGENVECTOR

Pagerank

- Ranking for search results: bring order to the web
- New challenges for information retrieval on the World Wide Web.
  - Huge number of web pages: 150 million by 1998
  - Diversity of web pages: different topics, different quality, etc.

- A method for rating the importance of web pages objectively and mechanically using the link structure of the web.

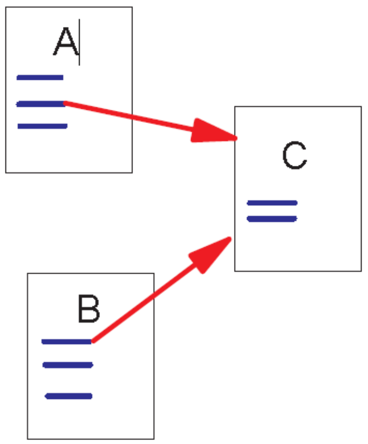
History

- PageRank was developed by Larry Page (hence the name *Page*-Rank) and Sergey Brin.
- It is first as part of a research project about a new kind of search engine.
- That project started in 1995 and led to a functional prototype in 1998.
- Shortly after, Page and Brin founded Google.

- Now, SEO use different trick methods to make a web page more important under the rating of PageRank.

Link Structure of the Web

- 150 million web pages →1.7 billion links
- Intuitively, a webpage is important if it has a lot of backlinks.



Backlinks and Forward links:

- A and B are C's backlinks
- C is A and B's forward link

Simple Version of PageRank

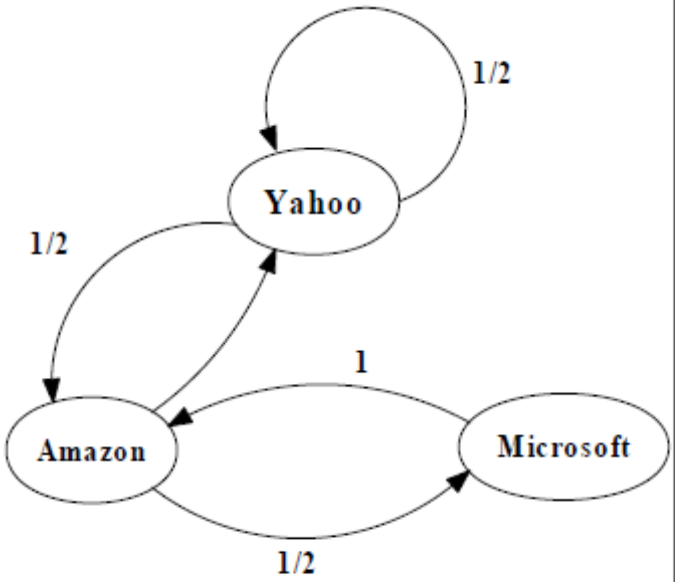
- u: a web page
- Bu: the set of u's backlinks
- Nv: the number of forward links of page v
- c: the normalization factor to make ||R||L1 = 1 (||R||L1= |R1 + … + Rn|)

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

-

Simple Version of PageRank

- PageRank Calculation: first iterations



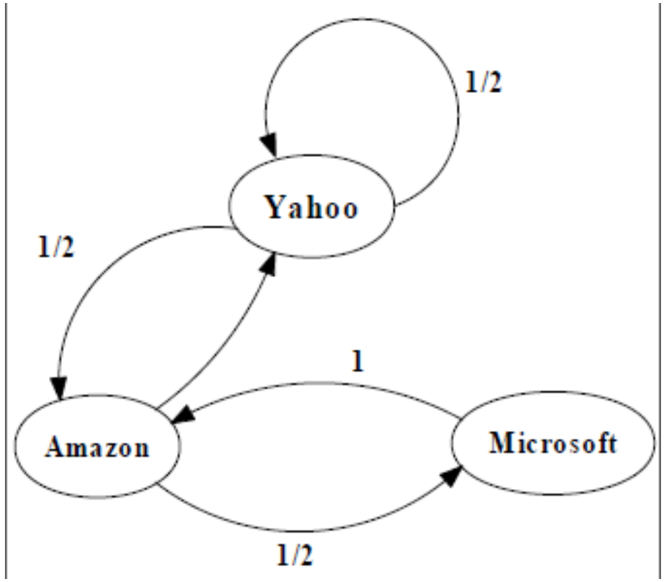
$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Simple Version of PageRank

- PageRank Calculation: second iterations



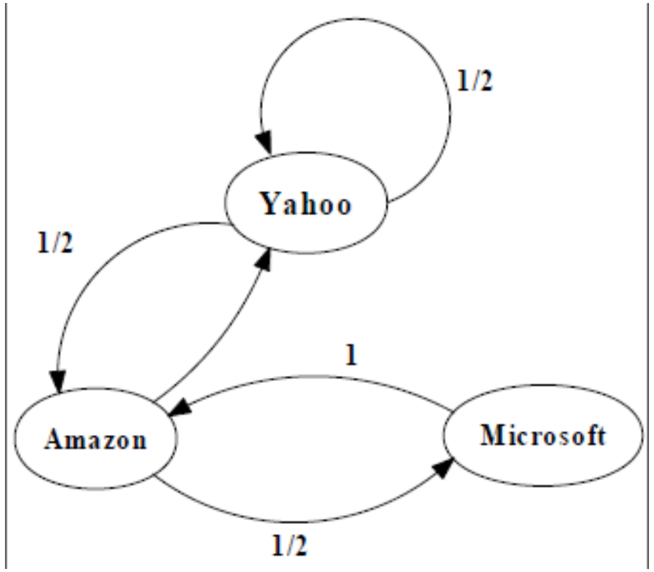
$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/2 \\ 1/6 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Simple Version of PageRank

- PageRank Calculation: converge after some iterations



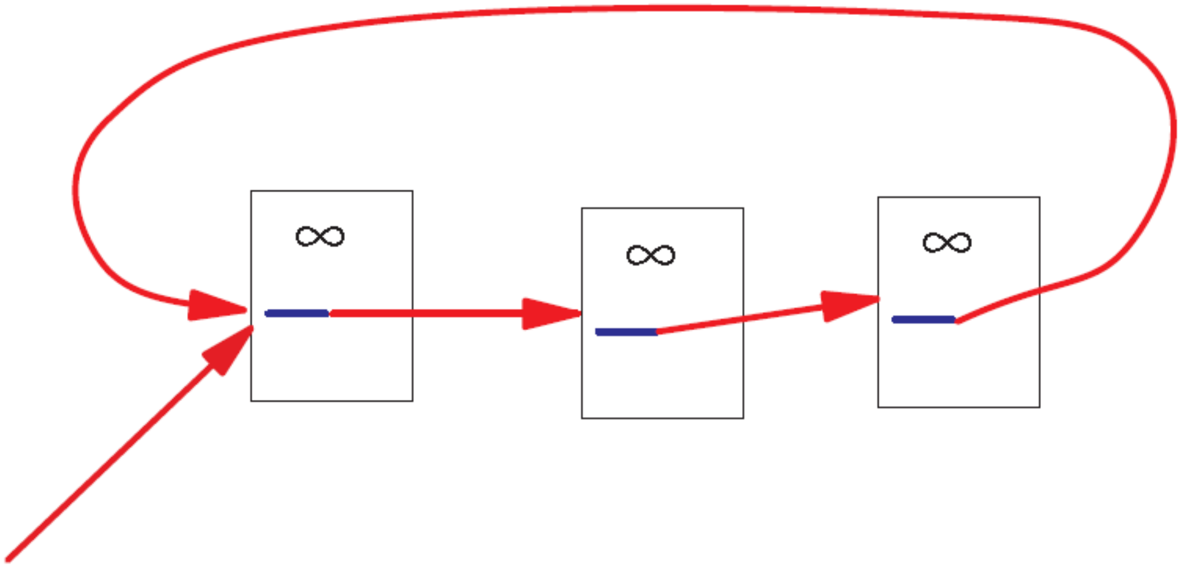
$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 3/8 \\ 11/24 \\ 1/6 \end{bmatrix} \begin{bmatrix} 5/12 \\ 17/48 \\ 11/48 \end{bmatrix} \ldots \begin{bmatrix} 2/5 \\ 2/5 \\ 1/5 \end{bmatrix}$$

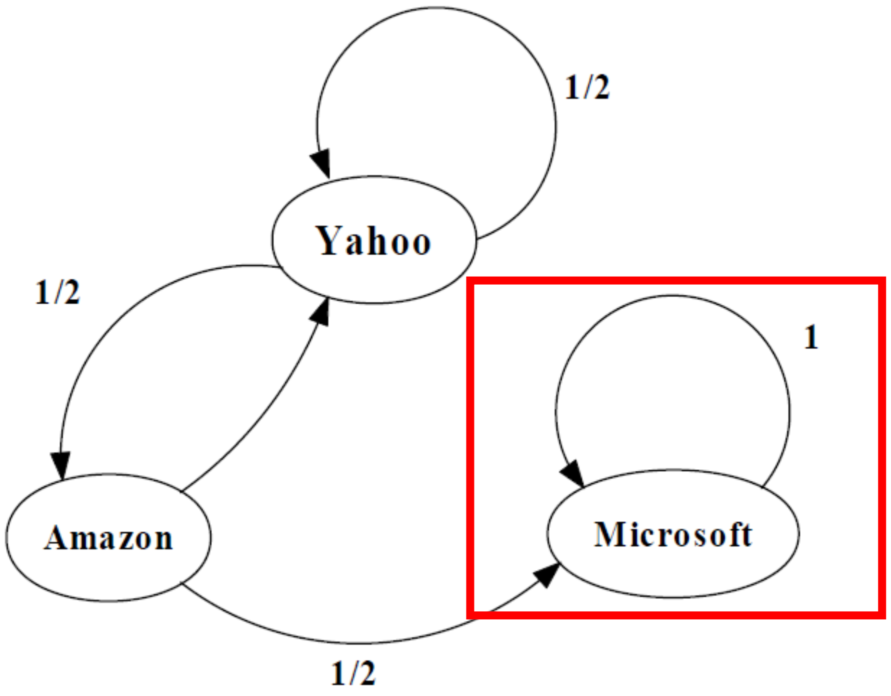
Simple Version of PageRank

- A loop



- During each iteration, the loop accumulates rank but never distributes rank to other pages

Simple Version of PageRank

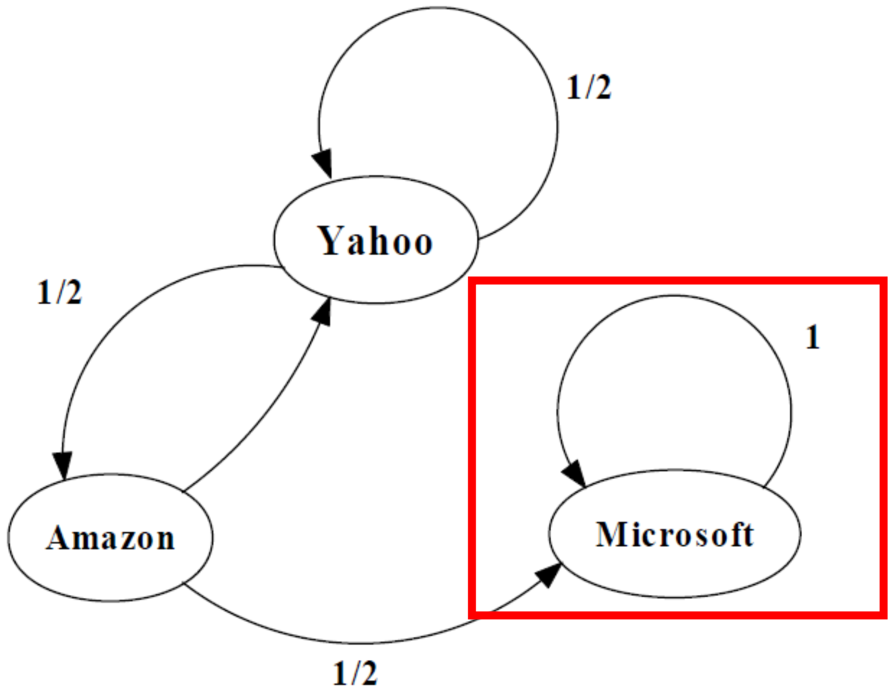


$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

Simple Version of PageRank

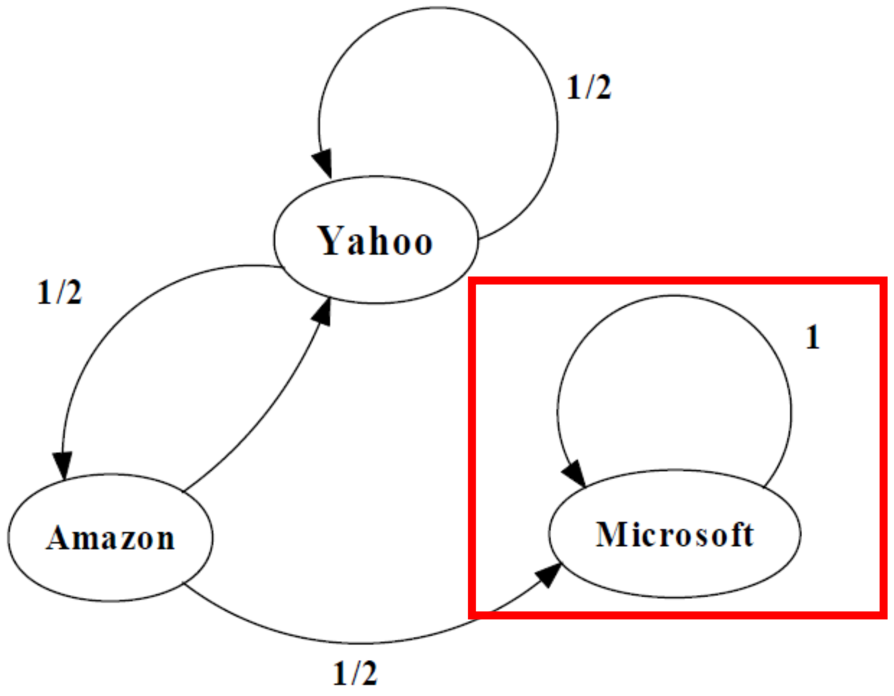


$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 \\ 1/6 \\ 7/12 \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/6 \\ 1/2 \end{bmatrix}$$

Simple Version of PageRank



$$M = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix}$$

$$\begin{bmatrix} \text{yahoo} \\ \text{Amazon} \\ \text{Microsoft} \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$\begin{bmatrix} 5/24 \\ 1/8 \\ 2/3 \end{bmatrix} \begin{bmatrix} 1/6 \\ 5/48 \\ 35/48 \end{bmatrix} \cdots\cdots \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Random Walks in Graphs

- The Random Surfer Model
  - The simplified model: the standing probability distribution of a random walk on the graph of the web. simply keeps clicking successive links at random
- The Modified Model
  - The modified model: the “random surfer” simply keeps clicking successive links at random, but periodically “gets bored” and jumps to a random page based on the distribution of E

$$R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u)$$

E(u): a distribution of ranks of web pages that “users” jump to when they “gets bored” after successive links at random.

Random Walks in Graphs

- The Web is an expander-like graph
  - Theory of random walk: a random walk on a graph is said to be rapidly-mixing if it quickly converges to a limiting distribution on the set of nodes in the graph. A random walk is rapidly-mixing on a graph if and only if the graph is an expander graph.
  - Expander graph: every subset of nodes S has a neighborhood (set of vertices accessible via outedges emanating from nodes in S) that is larger than some factor a times of |S|. A graph has a good expansion factor if and only if the largest eigenvalue is sufficiently larger than the second-largest eigenvalue.

Conclusion

- PageRank is a global ranking of all web pages based on their locations in the web graph structure
- PageRank uses information which is external to the web pages – backlinks
- Backlinks from important pages are more significant than backlinks from average pages
- The structure of the web graph is very useful for information retrieval tasks.