The Edge

Dimitris Bertsimas
MIT

Reflections from abroad

- Greece is in the middle of one of the most significant economic crisis of its history that threatens the very way of life of a large portion of the Greek population.
- There has traditionally been significant unemployment of young Greeks and especially of university graduates. The economic crisis has worsened the situation significantly.
- Young people in Greece do not see a path towards a life that is better than their parents.
- The fundamental problem of Greece is that there is no significant engine of growth.

Some personal history

- I had and continue to have two parallel careers in my 25 years in the US: A professor of Operations Research at MIT and an entrepreneur in analytics.
- When I received the invitation to give a talk, I naturally thought to talk about my experiences of my first career; I suspect this is why you invited me on the first place.
- After reflecting on the current situation in Greece that I summarized earlier, I decided to talk about a key lesson I learned from my second career, that I feel is quite relevant to the issues that Greece and Greek universities are facing.
- Most importantly, I feel this lesson has the potential of creating an engine of growth for Greece that can sustain our way of life.

Some personal history

- The first company I started was Dynamic Ideas in 1998. It built quantitative models for asset management. Sold to American Express in 2002; The group became the quantitative division of Riversource Investments, and in 2010 managed $12 billion.

- The second company I was involved was D2 Hawkeye in 2001. It is a medical data mining company that developed algorithms for predicting the health risk of individuals based on claims data. The company was sold to Verisk in 2009. It now manages the medical records of more than 30 million people.

Key lesson from my Experience

- Unprecedented availability of **Data**
- Opportunities to build **Models** based on this data
- This leads to better **Decisions**
- In many cases, the **DMD**approach doesn't just help at the margin. It creates a key competitive **Edge** that is very difficult for incumbents to counter.
- When properly implemented the **Edge** can lead to dominant companies, significant wealth generation, a engine for growth and often a change in the world.
- The **DMD**approach can dramatically "move the needle" in the real world.

Key lesson from my Experience

- As the volume, variety and availability of data continues to increase,
  - the potential to create DMD-powered businesses will continue to increase
  - the market value of skilled and experienced “analytic practitioners” will increase as well

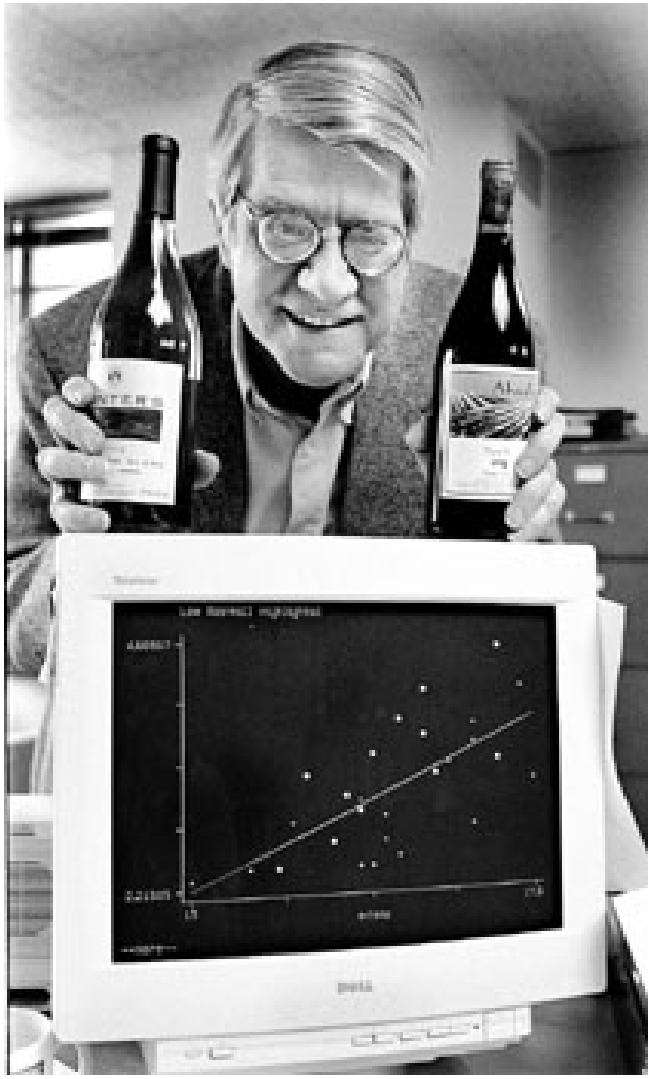
Aspirations of the Lecture

- Illustrate the DMD approach and its**Edge**.
- In one Industry, start with a real world company and illustrate the core thinking and methodology that created the edge.
- Most importantly, inspire especially the younger members of the audience to start their own analytics-powered businesses and change the world.
- Propose a path for AUEB to contribute to the economic growth of Greece in a period that Greece absolutely needs it.

Outline

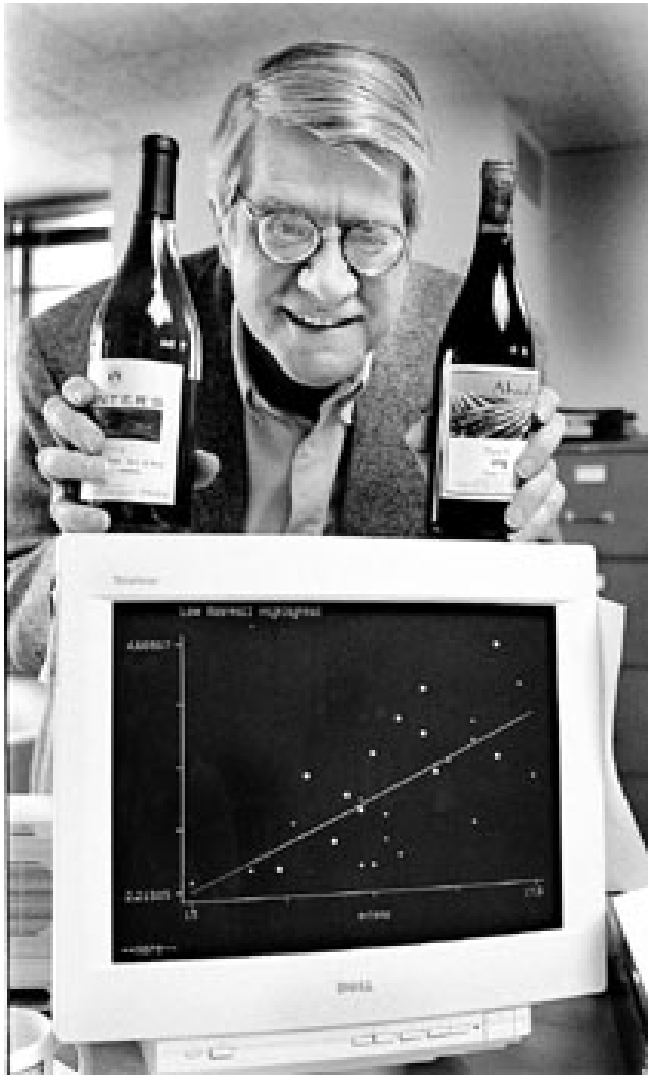
- The Edge in Wines
- The Edge in the Internet
- The Edge in Finance
- Other Edges
- A proposal for the Athens University of Economics and Business

The Edge in Wines



The Edge in Wines

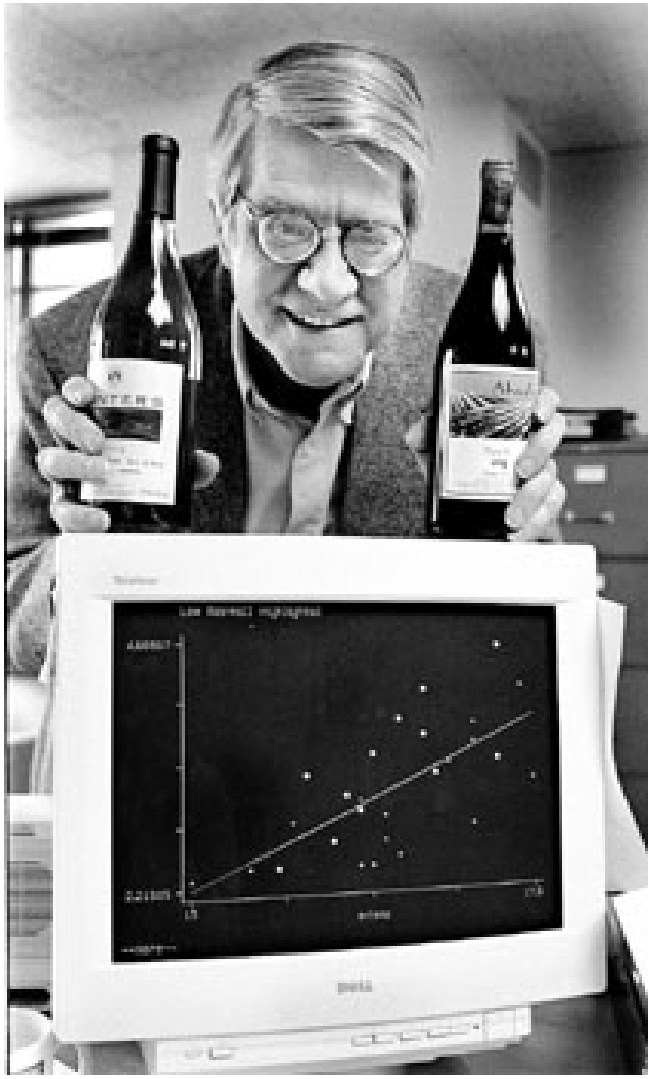
- Orley Ashenfelter, a Princeton professor, decided he could do better than the ``swishing and spitting" approach of wine experts.



The Edge in Wines

- Orley Ashenfelter, a Princeton professor, decided he could do better than the ``swishing and spitting" approach of wine experts.

- He analyzed some data sets and came up with a formula for predicting wine quality.

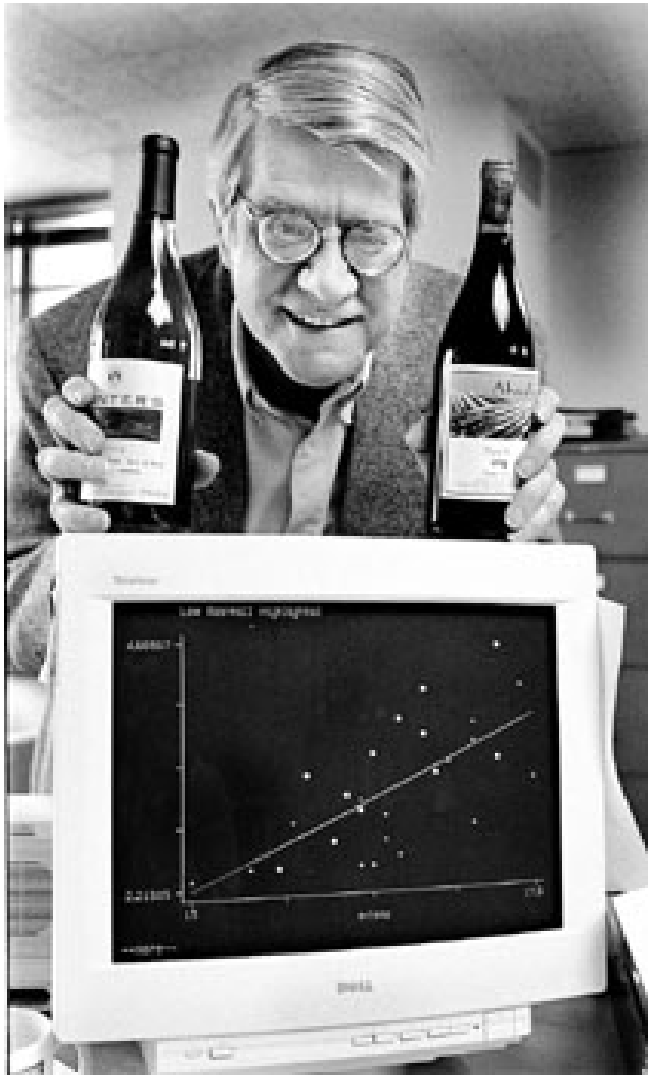


The Edge in Action: DMD and Wines

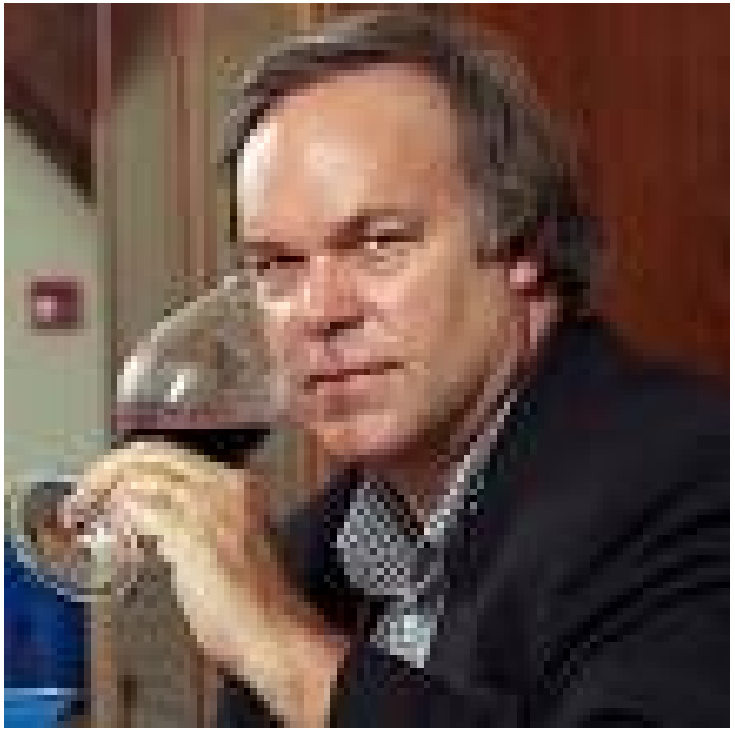
- Orley Ashenfelter, a Princeton professor, decided he could do better than the ``swishing and spitting" approach of wine experts.

- He analyzed some data sets and came up with a formula for predicting wine quality.

- His formula was based on just three variables: winter rainfall, average growing season temperature and harvest rainfall.



The Edge in Action: DMD and Wines



- Britain's Wine magazine sniffed: “the formula's self-evident silliness invites disrespect”.
- Robert Parker, the world's most influential writer and publisher of the Wine Advocate: “Ashenfelter is an absolute total sham”.
- Parker said Ashenfelter was “rather like a movie critic who never goes to see the movie but tells you how good it is based on the actors and the director”.

The Edge in Action: DMD and Wines

- Parker rated the 1986 Bordeaux “very good to sometimes exceptional”.

The Edge in Wines

- Parker rated the 1986 Bordeaux “very good to sometimes exceptional”.

- Ashenfelter disagreed:
  - He claimed that below average growing season temperature and above average harvest rainfall doomed this vintage to mediocrity.
  - In 1989, he predicted that the 1989 Bordeaux will be “the wine of the century”.
  - In 1990, he predicted that the 1990 vintage would be even better!

The Edge in Wines

- So who won?

The Edge in Wines

- So who won?
  - In wine auctions, the 89s sold for more than twice the price of 86s and the 90s sold for even higher prices!
  - The humble regression formula that beat Robert Parker:
    - Wine Quality= 12.145+0.00117 winter rainfall
      +0.0614 average growing season temperature
      -0.00386 harvest rainfall
    - See http://www.liquidasset.com/ for more.

Lecture Outline

- The Edge in Wines
- The Edge in the Internet
- The Edge in Finance
- Other Edges
- A proposal for the Athens University of Economics and Business

The Edge in the Internet

- What was the landscape in Internet search fifteen years ago?
- Why Microsoft offered $45 billion for Yahoo?
- Why Google is the dominant player today?

The Edge in the Internet: Billions of Dollars Later …

Larry Page and Sergey Brin, Google

Searching the Web: Why is it hard?

- The web is huge and growing: Google's index of web pages crossed the 1 trillion page mark in July 2008!
- Dynamic: 40% of all webpages change within a week and 23% of .com pages change daily.
- Self-organized: any one can post a webpage and link away at will.
- Data is heterogeneous, and exists in multiple formats, languages and alphabets.

Searching the Web: Building blocks

- Query module
  - converts natural language to numbers and consults various indexes to answer queries from users
  - module consults content index and its inverted file to find pages that use the query terms

- Ranking module
  - **takes relevant pages and ranks them**
  - the outcome: ordered list of webpages. This is the most important component of the search process.

Page Rank

- Think of a hyperlink as a recommendation or a vote: a hyperlink from my page to yours is an endorsement of your webpage.
- A page with more recommendations (i.e., inbound links) is more important. As of April 5, 2010:
  - www.mit.edu had 46,400 inbound links
  - www.harvard.edu had 9,320 inbound links

Page Rank

- Think of a hyperlink as a recommendation or a vote: a hyperlink from my page to yours is an endorsement of your webpage.
- A page with more recommendations (i.e., inbound links) is more important. As of April 5, 2010:
  - www.mit.edu had 46,400 inbound links
  - www.harvard.edu had 9,320 inbound links
- But the status of the recommender matters as well.

Page Rank

- An endorsement from Warren Buffet probably does more to strengthen a job application than 20 endorsements from 20 unknown teachers and colleagues.
- However, if the interviewer knows that Warren Buffet is very generous with his praise and has written 20,000 recommendations, then the endorsement drops in weight.

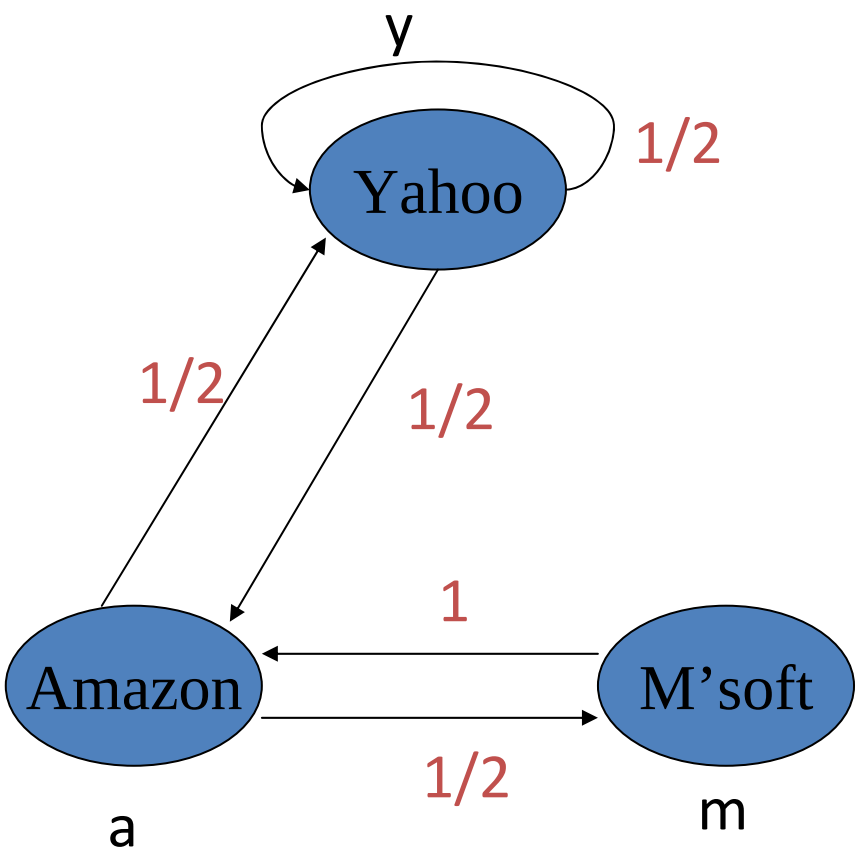
Key Idea

- Consider the web as a graph and imagine a random surfer of the web, who is visiting random links moving from page to page.
- How likely it is that the random surfer will be on a particular page after a long period of time?
- For the probabilists in the audience, what is the stationary probability distribution of the Markov chain defined by the hyperlink structure of the web?

The math underlying PageRank

- Each link's vote is proportional to the importance of its source page
- If page P with importance x has n outlinks, each link gets x/n votes
- Page P's own importance is the sum of the votes on its inlinks

Simple model



$$y = y/2 + a/2$$

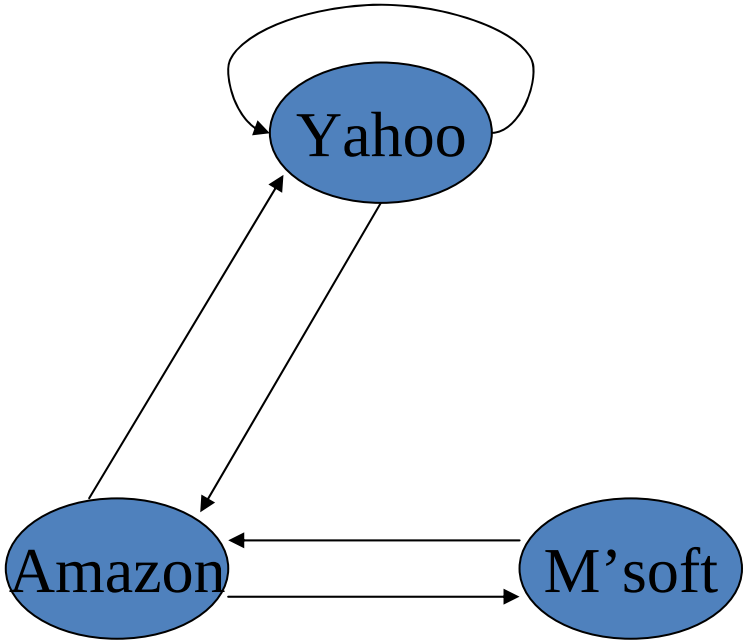
$$a = y/2 + m$$

$$m = a/2$$

Solving the equations

- 3 equations, 3 unknowns, no constants
  - No unique solution
  - All solutions equivalent modulo scale factor
- Additional constraint forces uniqueness
  - y+a+m = 1
  - y = 2/5, a = 2/5, m = 1/5
- Gaussian elimination method works for small examples, but we need a better method for large graphs

Example



$$y = y/2 + a/2$$
$$a = y/2 + m$$
$$m = a/2$$

	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

r = Mr

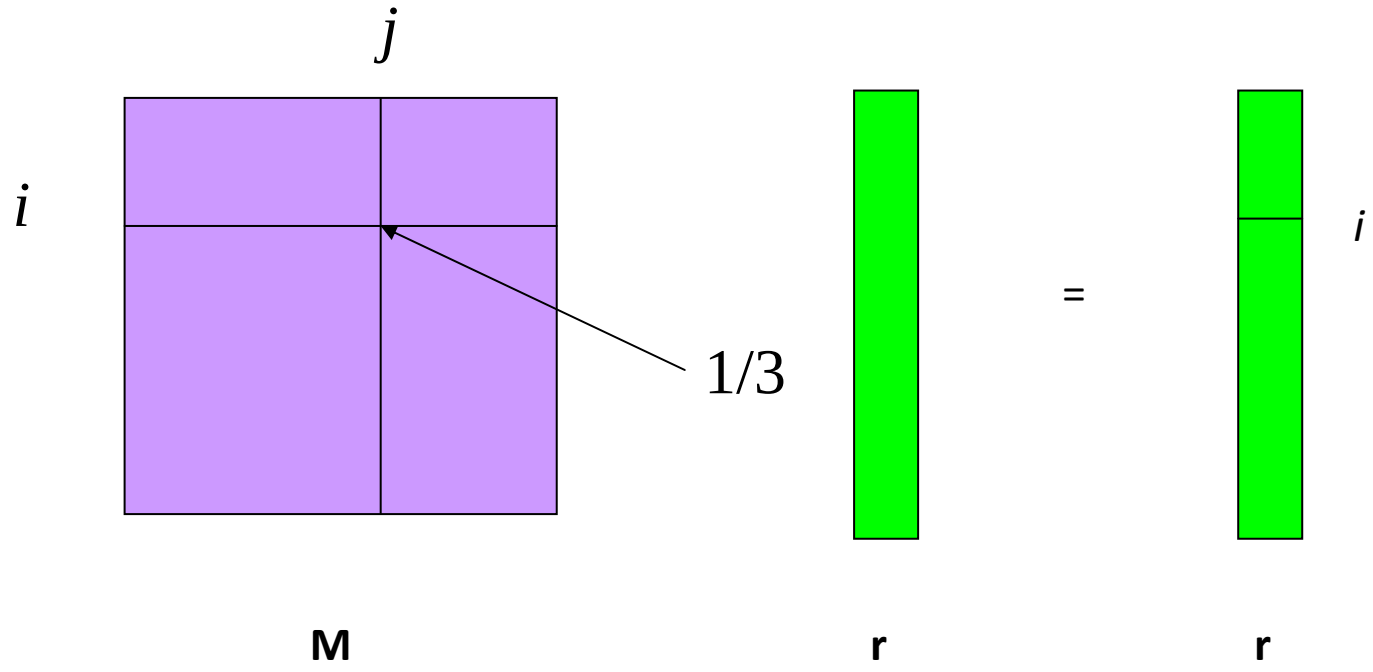
$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

Matrix formulation

- Matrix **M** has one row and one column for each web page
- Page j has noutlinks
  - If jnot i, then $M_{ij}=1/n$
  - Else $M_{ij}=0$
- Columns of **M** sum to 1
- **r**is a vector with one entry per web page
  - r_i is the importance score of page i
  - Call it the rank vector
  - $|\mathbf{r}| = 1$

Example

Suppose page *j* links to 3 pages, including *i*

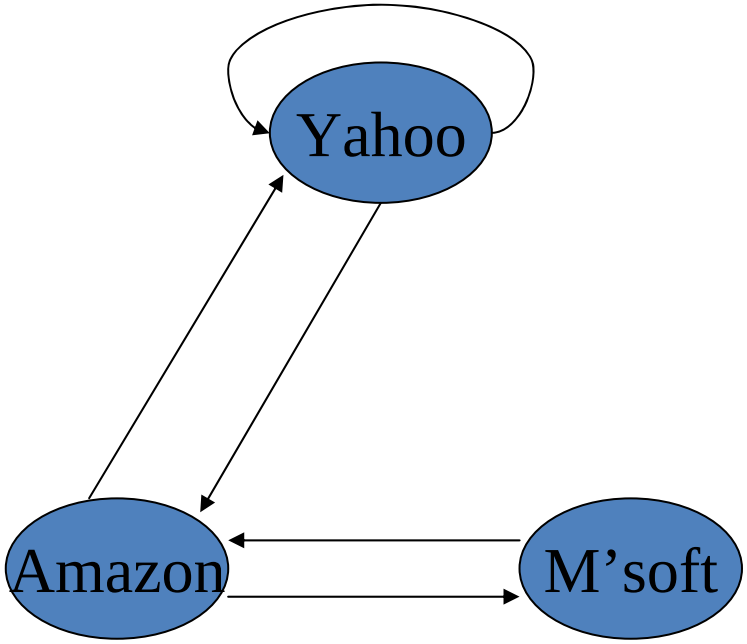


The equations can be compactly written ***r= Mr***

Power Iteration method

- Simple iterative scheme
- Suppose there are N web pages
- Initialize: $\mathbf{r}^0 = [1/N,\ldots.,1/N]^T$
- Iterate: $\mathbf{r}^{k+1} = \mathbf{M}\mathbf{r}^k$
- Stop when $|\mathbf{r}^{k+1} - \mathbf{r}^k|_1 < \varepsilon$

Power Iteration Example



	y	a	m
y	1/2	1/2	0
a	1/2	0	1
m	0	1/2	0

y		1/3	1/3	5/12	3/8		2/5
a	=	1/3	1/2	1/3	11/24	. . .	2/5
m		1/3	1/6	1/4	1/6		1/5

Will the rank vector always make sense?

A central result from the theory of random walks (aka Markov processes):

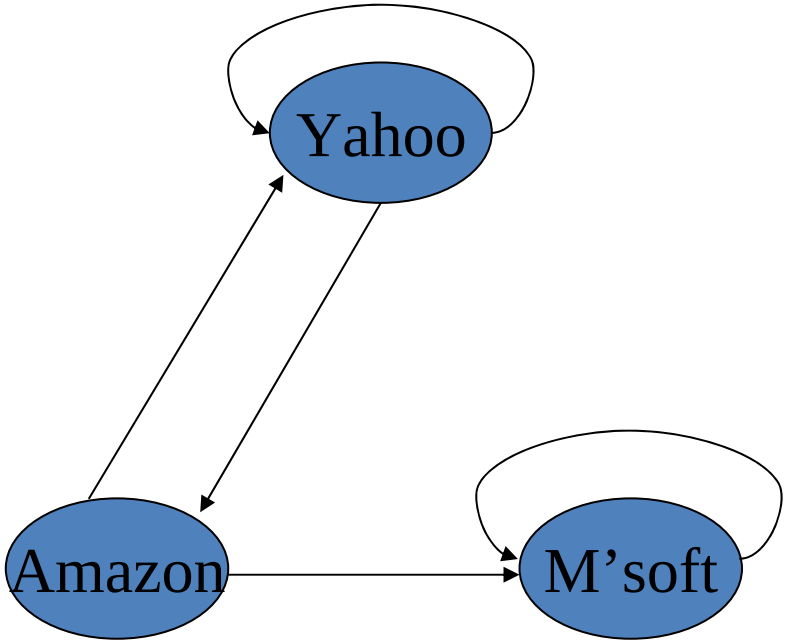
For graphs that satisfy certain conditions, the stationary distribution is unique and eventually will be reached no matter what the initial probability distribution.

“Spider traps” and “dead ends”

- Spider traps
  - A group of pages is a spider trap if there are no links from within the group to outside the group
  - Spider traps violate the conditions needed for thePageRank algorithm to converge

- Dead ends: pages with no outlinks

Let's say the Microsoft site is a spider trap



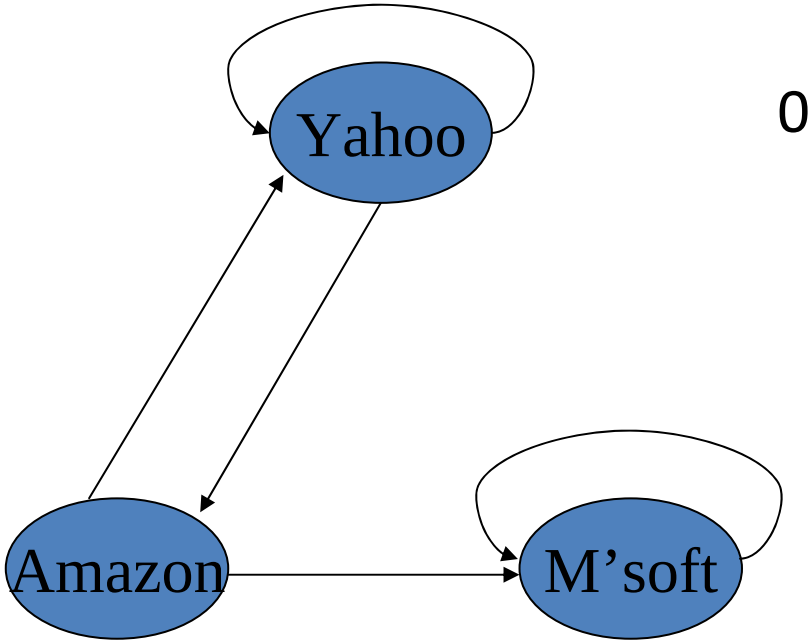
	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	1

y		1	1	3/4	5/8		0
a	=	1	1/2	1/2	3/8	. . .	0
m		1	3/2	7/4	2		3

The Google spider trap antidote

- Suppose there are N pages
  - Consider a page j, with set of outlinksO(j)
  - We have $M_{ij} = 1/|O(j)|$ when j not iand $M_{ij} = 0$ otherwise
  - The “spider trap antidote” is equivalent to
    - adding afake linkfrom j to every other page with weight $(1-\beta)/N$
    - reducing the weight of each outlink from $1/|O(j)|$ to $\beta/|O(j)|$
    - Equivalent: tax each page a fraction $(1-\beta)$ of its score and redistribute evenly

The Google spider trap antidote



$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 1 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 13/15 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} = \begin{matrix} 1 & 1.00 & 0.84 & 0.776 & & 7/11 \\ 1 & 0.60 & 0.60 & 0.536 & \ldots & 5/11 \\ 1 & 1.40 & 1.56 & 1.688 & & 21/11 \end{matrix}$$

Page Rank

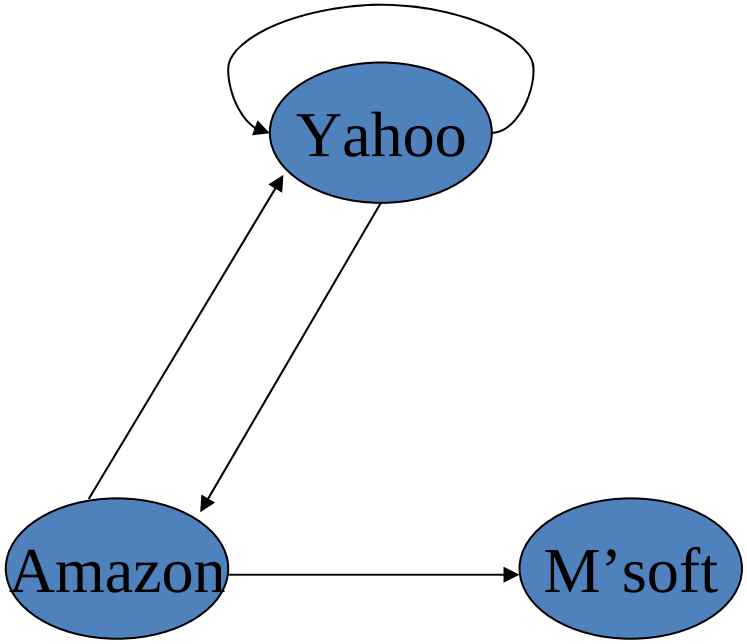
- Construct the NxNmatrix **A** as follows
  - $A_{ij} = \beta M_{ij} + (1-\beta)/N$
- Verify that**A's** columns add to 1
- **A** is called the "Google matrix"
- The page rank vector**r**satisfies the equations**r = Ar**

“Spider traps” and “dead ends”

- Spider traps
  - A group of pages is a spider trap if there are no links from within the group to outside the group
  - Spider traps violate the conditions needed for thePageRank algorithm to converge

- Dead ends: pages with no outlinks

Microsoft becomes a dead end



$$0.8 \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \end{bmatrix} + 0.2 \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$\begin{matrix} y \\ a \\ m \end{matrix} \begin{bmatrix} 7/15 & 7/15 & 1/15 \\ 7/15 & 1/15 & 1/15 \\ 1/15 & 7/15 & 1/15 \end{bmatrix}$$

Doesn't add up to 1!

$$\begin{matrix} y \\ a \\ m \end{matrix} = \begin{matrix} 1 \\ 1 \\ 1 \end{matrix} \quad \begin{matrix} 1 \\ 0.6 \\ 0.6 \end{matrix} \quad \begin{matrix} 0.787 \\ 0.547 \\ 0.387 \end{matrix} \quad \begin{matrix} 0.648 \\ 0.430 \\ 0.333 \end{matrix} \quad \begin{matrix} \\ \ldots \\ \end{matrix} \quad \begin{matrix} 0 \\ 0 \\ 0 \end{matrix}$$

Dealing with dead-ends

- Simpler version of “spider trap antidote”
- Preprocess the graph to eliminate dead-ends

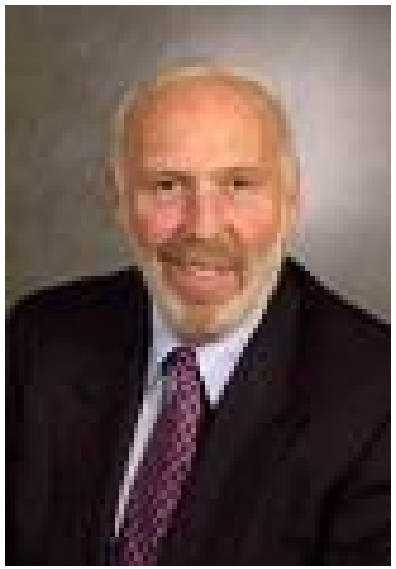
Where is the Edge?

- Simple but powerful mathematics.
- Implemented brilliantly.
- Distinguished between free search and sponsored links.
- Created a very powerful brand.
- Would there be a business without the math?

Lecture Outline

- The Edge in Wines
- The Edge in the Internet
- The Edge in Finance
- Other Edges
- A proposal for the Athens University of Economics and Business

The Edge in Finance: Billions of Dollars Later …



James Simons, Renaissance Technologies

The Edge in Finance

- High frequency trading
- Importance of quantitative alpha models for prediction of stock price movement tick by tick
- Importance of optimization of transaction costs as the benefit of the alpha models is comparable to the transaction costs
- Importance of implementation
- Renaissance Technologies, DE Shaw are examples of such firms that exhibited an edge that made their founders billionaires.

Lecture Outline

- The Edge in Wines
- The Edge in the Internet
- The Edge in Finance
- Other Edges
- A proposal for the Athens University of Economics and Business

Edge in Pricing

- The Edge in retail: ProfitLogic and optimizing prices of fashion goods

- The Edge in retail (continued): Optimizing prices of staple goods

- The Edge in consumer finance: Nomis Solutions and pricing consumer loans

nomis solutions™ | Pricing and Profitability Management for Financial Services

Edge in Healthcare

- The Edge in health care: Informatics by Humedica

- The Edge in pharma: MarketRx and optimizing salesforce effectiveness

- The Edge in medicine: Memorial Sloan-Kettering and optimizing cancer treatment

Edge in Online Advertising

- The Edge in Search/ Computational Advertising

- The Edge in paid search advertising: Efficient Frontier and optimizing SEM

- The Edge in display advertising: DataXu and optimizing ad impressions

Edge in the Internet

- Mining the Long Tail: Amazon, NetFlix and recommender systems

- The Edge in social media: Real-time Analytics with Twitter

- The Edge in social media: Social Graph Analytics

Other Edges

- The Edge in B2B CRM: Lattice Engines

- The Edge in internet auctions: eBay

- The Edge in gaming: Revenue Management at Harrah's Casinos

Key concepts

- Data mining methods
- Optimization
- Probabilistic modeling
- Most importantly: Approach the world quantitatively

Lecture Outline

- The Edge in Wines
- The Edge in the Internet
- The Edge in Finance
- Other Edges
- A proposal for the Athens University of Economics and Business

A proposal for AUEB

- Focus on Analytics as a strategic area
- Create a new undergraduate and Masters program of analytics
- Hire professors who have quantitative backgrounds, and especially entrepreneurship experience in Analytics;
- The government funds Small Business grants for Analytics and obtains 25% ownership.
- When successes occur, the 25% ownership can fund the new companies of the future.

Why does it have a chance?

- Contrast with the growth of the High Tech industry in **Israel.**
- Israel is smaller country than Greece in constant war with its neighbors.
- What is a significant engine of growth in the US: Its research universities in collaboration with the VC community. Here the government becomes the VC.
- So, we have an existence proof.
- Whether we have an efficient, constructive algorithm of this plan really depends **on you.**
- Having done it twice, I can tell you it is among the most creative, challenging and fun experiences of my life.

Questions?