Mathematics Everywhere:

from Google to Sudoku to NCAA to Netflix

Amy Langville
Mathematics Department
College of Charleston

MAA-SC 3/20/2009

Mathematics Everywhere:

from Google to Sudoku to NCAA to Netflix

(how an OR analyst thinks and works)

Amy Langville
Mathematics Department
College of Charleston

MAA-SC 3/20/2009

Mathematics Everywhere:

from Google to Sudoku to NCAA to Netflix

(how an OR analyst thinks and works)

(my life as an OR analyst)

Amy Langville
Mathematics Department
College of Charleston

MAA-SC 3/20/2009

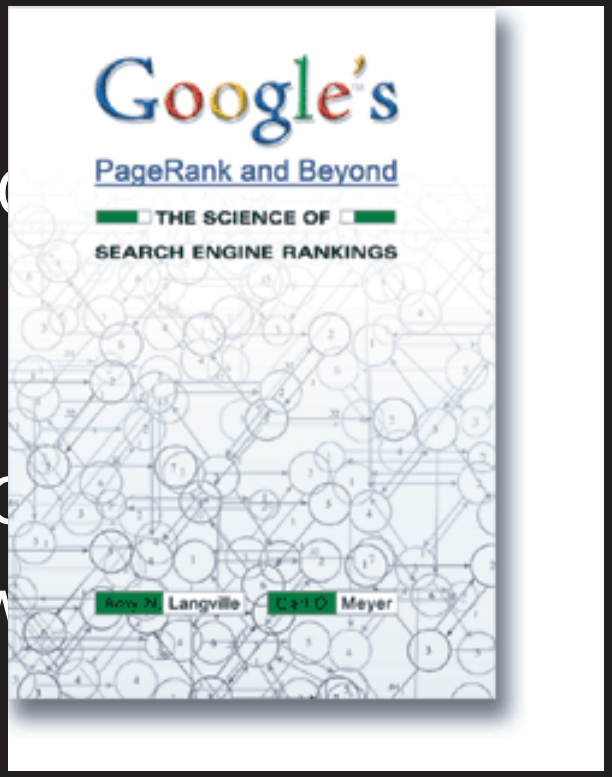
Brief History

- 1997, B.S. Mathematics, Mt. St. Mary's College
- 2002, Ph.D. Operations Research, N. C. State University
- 2005, Postdoc Applied Mathematics, N. C. State University
  - information retrieval, text mining, webpage ranking

Brief History

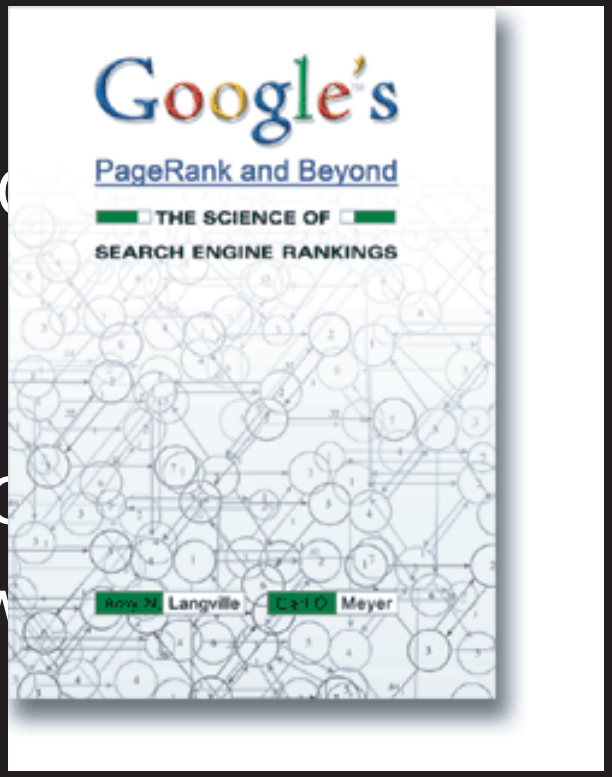
- 1997, B.S. Mathematics, Mt. St. Mary's College
- 2002, Ph.D. Operations Research, N. C. State University
- 2005, Postdoc Applied Mathematics, N. C. State University
  - information retrieval, text mining, webpage ranking
  - Carl Meyer

Brief History



- 1997, B.S. Mathematics, Mt. St. Mary's C
- 2002, Ph.D. Operations Research, N. C.
- 2005, Postdoc Applied Mathematics, N. C
  - information retrieval, text mining, w
  - Carl Meyer
  - *Google's PageRank and Beyond: The Science of Search Engine Rankings*

Brief History



- 1997, B.S. Mathematics, Mt. St. Mary's C
- 2002, Ph.D. Operations Research, N. C.
- 2005, Postdoc Applied Mathematics, N. C
  - information retrieval, text mining, w
  - Carl Meyer
  - *Google's PageRank and Beyond: The Science of Search Engine Rankings*

Unexpected Outcomes

+ invitations to speak/travel

+ consult for tech and legal firms

Ranking webpages

Sign in

Web Images Video News Maps more »

Google

business intelligence

Search

Advanced Search
Preferences

Web

Results **1** - **10** of about **122,000,000** for **business intelligence**. (**0.10** seconds)

SAS **Business Intelligence** Sponsored Link
www.SAS.com Get Better Answers Faster w/ SAS' Award-winning BI Software. Get Info

Business intelligence - Wikipedia, the free encyclopedia
Business intelligence (**BI**) is a **business** management term which refers to applications and technologies which are used to gather, provide access to, ...
en.wikipedia.org/wiki/**Business_intelligence** - 43k - Cached - Similar pages

Business Intelligence .com :: The Resource for **Business Intelligence**
The **Business Intelligence** resource for **business** and technical professionals covering a wide range of topics including Performance Management, Data Warehouse ...
www.**businessintelligence**.com/ - 74k - Apr 15, 2007 - Cached - Similar pages

Business Intelligence and Performance Management Software ...
Business intelligence and **business** performance management software. Reporting, analytics software, budgeting software, balanced scorecard software, ...
Stock quote for COGN
www.cognos.com/ - 32k - Cached - Similar pages

Oracle **Business Intelligence** Solutions
The First Comprehensive, Cost-Effective BI Solution Only Oracle delivers a complete, pre-integrated technology foundation to reduce the cost and complexity ...
www.oracle.com/solutions/ **business_intelligence**/index.html - 55k - Cached - Similar pages

Business Intelligence - Management Best Practice Reports
Business Intelligence: Providers of independent reports containing best practice advice, proprietary research findings and case studies for senior managers ...
www.**business-intelligence**.co.uk/ - 18k - Cached - Similar pages

Intelligent Enterprise: Better Insight for Business Decisions

Sponsored Links

SQL Database Management
Enterprise Data Mgmt Solutions From Dell™. Find Out More Here
www.dell.com

Business Intelligence
See what **business intelligence** can do for you (free interactive demo).
www.InformationBuilders.com

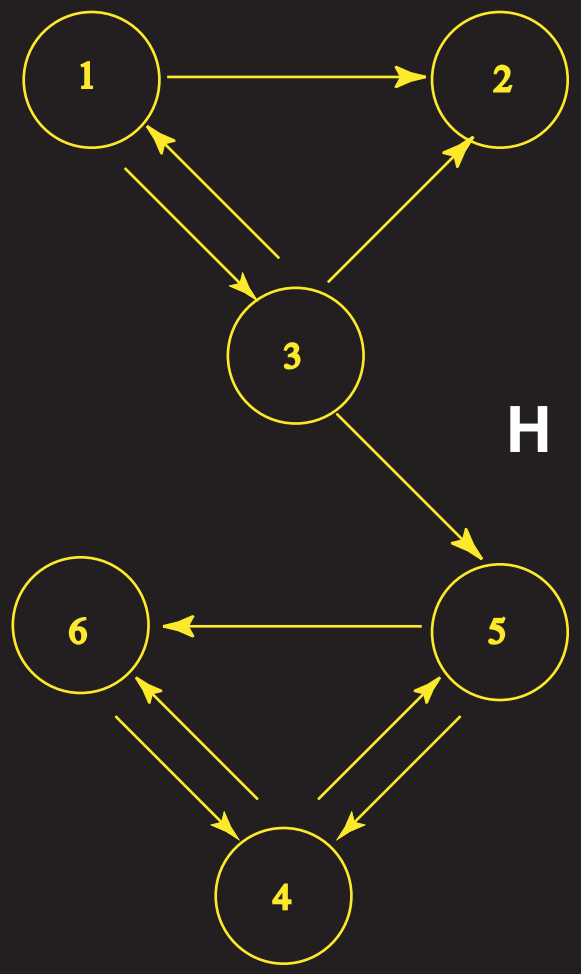
MCITP: BI Cert Boot Camp
9-Day MCITP Certification Boot Camp **Business Intelligence** All Inclusive
www.mcseclasses.com

Business Intelligence
Improve information integrity with real-time data integration software
www.DataMirror.com

Love Data?
Empower yourself with MS BI Tools via SetFocus' Master's Program
www.SetFocus.com

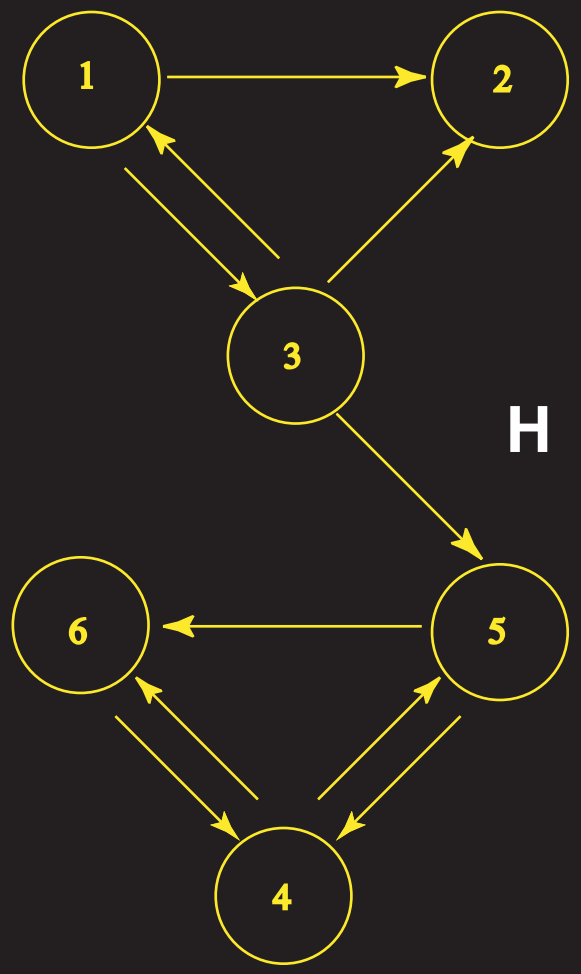
Business Intelligence
Conquer DW/BI Slowdown. Get Faster Queries & Performance - Learn How.
www.Sybase.com

Tiny Web



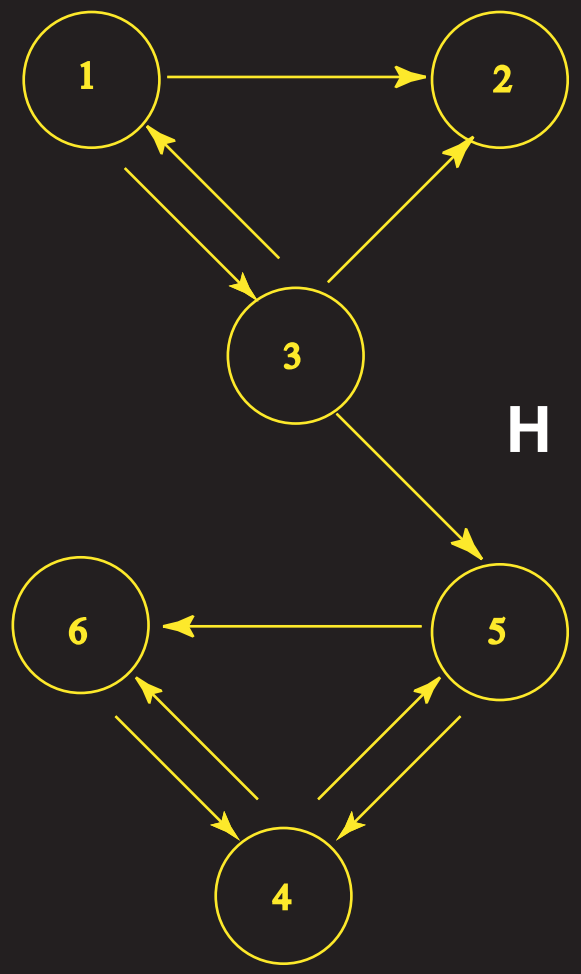
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



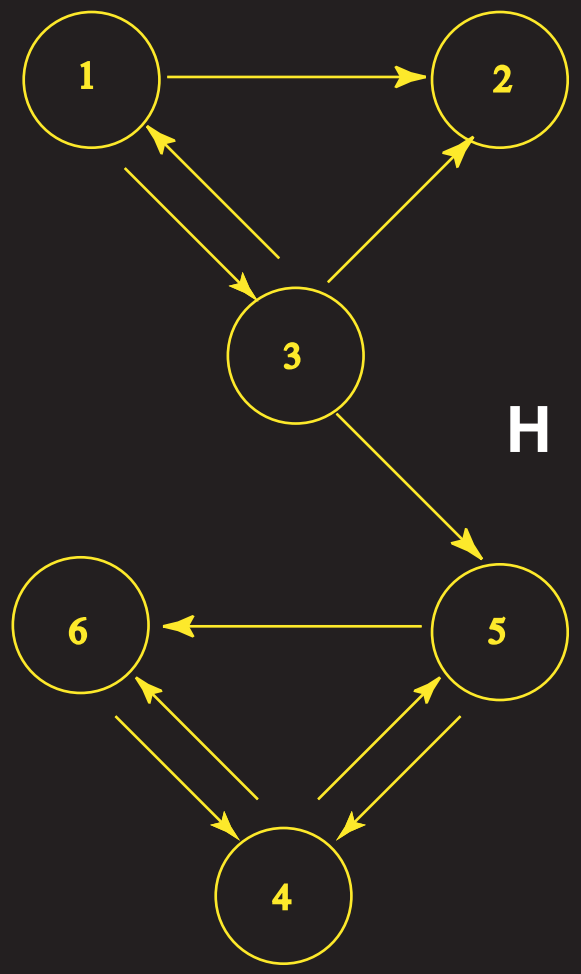
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



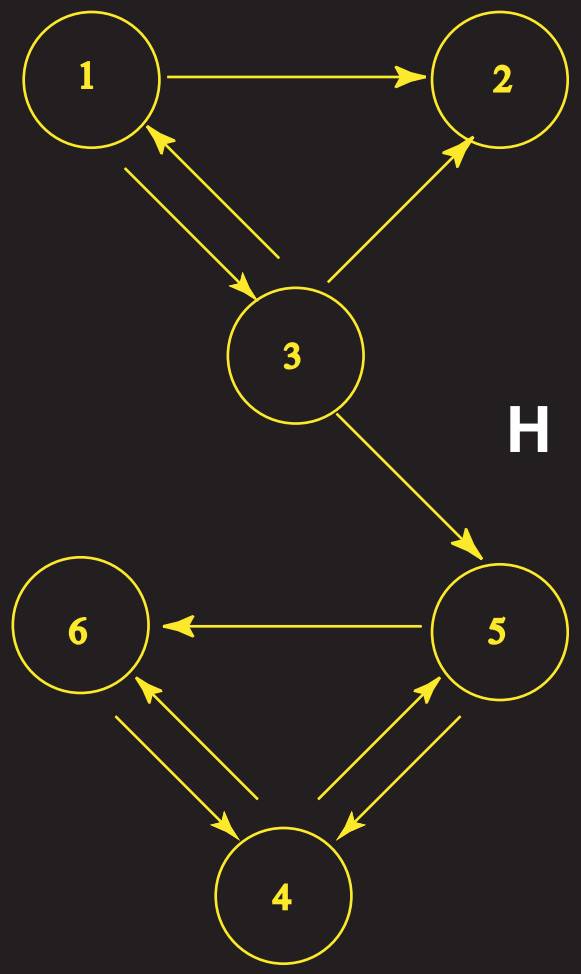
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



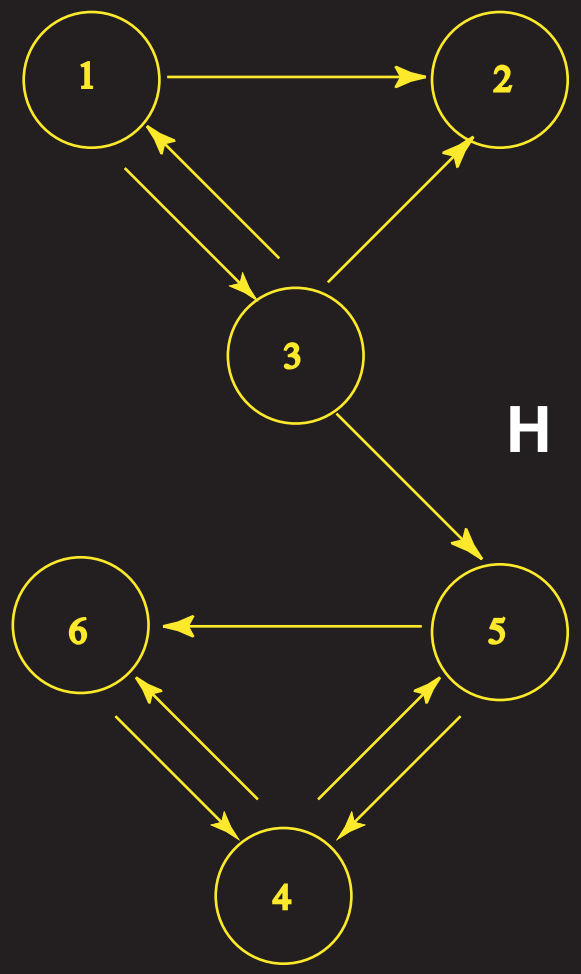
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



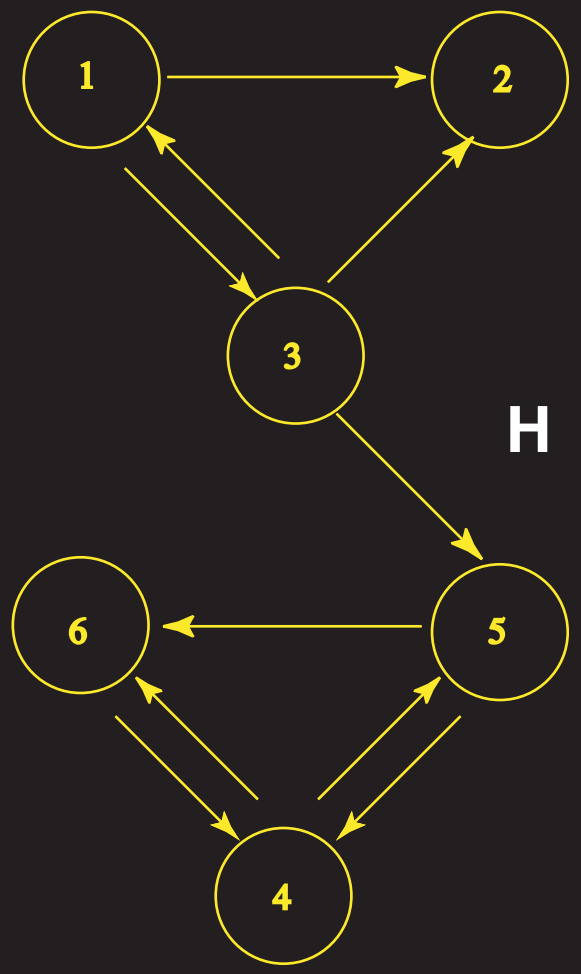
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ & & & & & \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



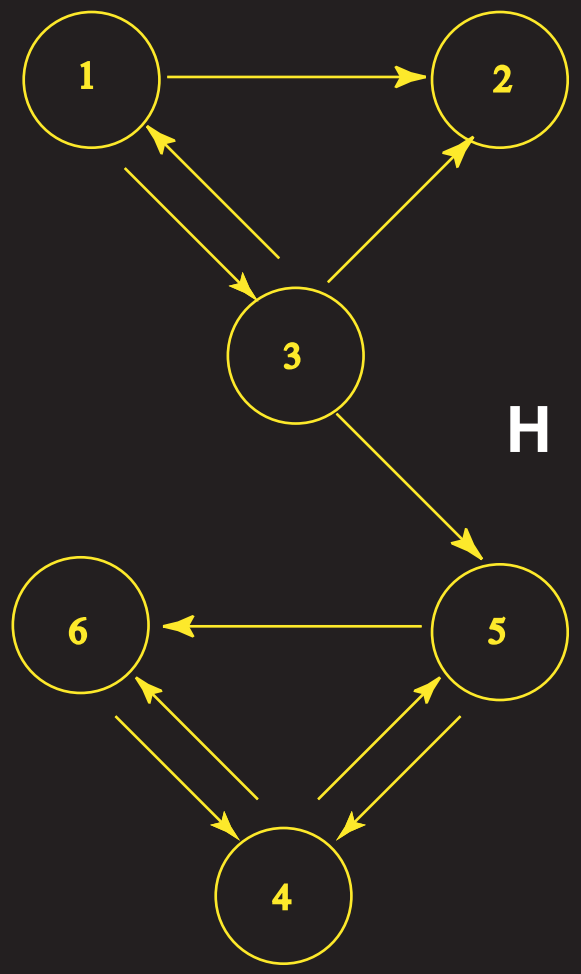
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ & & & & & \end{pmatrix} \end{array}$$

Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

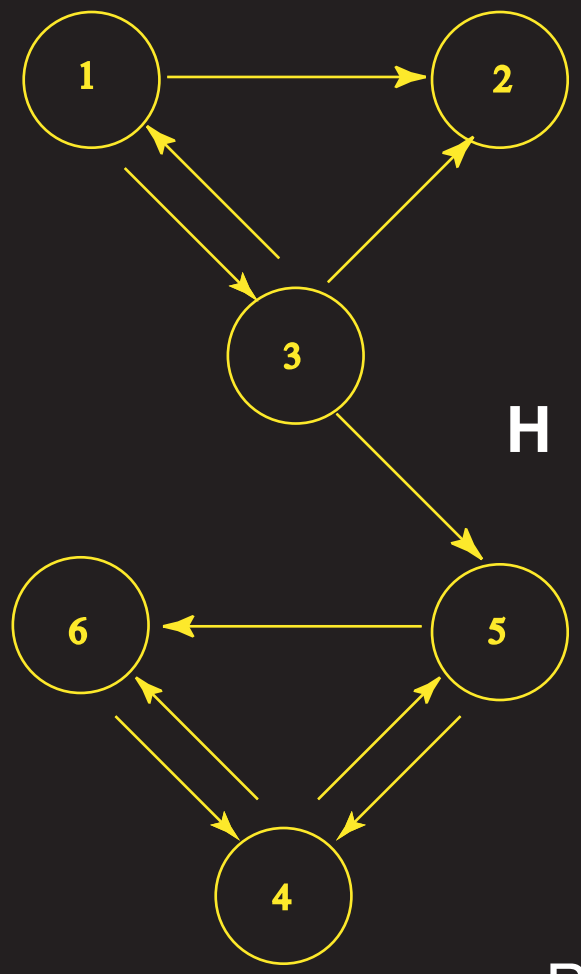
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph

Tiny Web



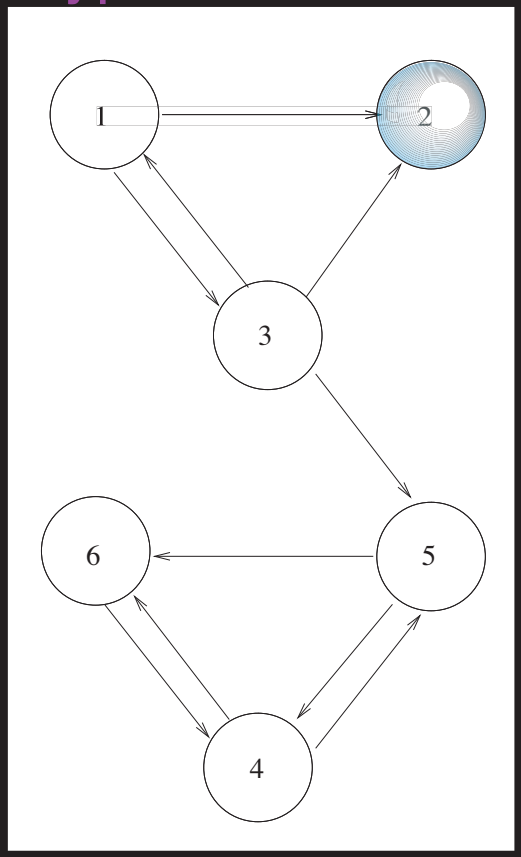
$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_i = amount of time spent at P_i

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote

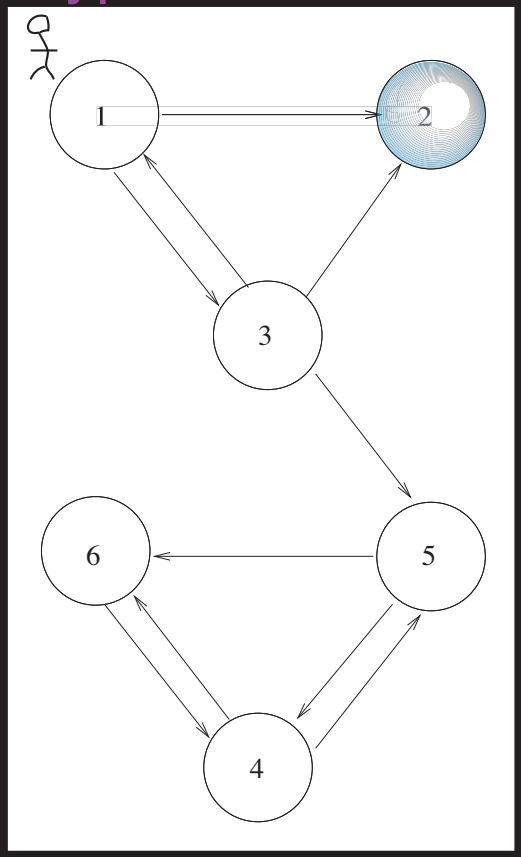


Markov chain

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

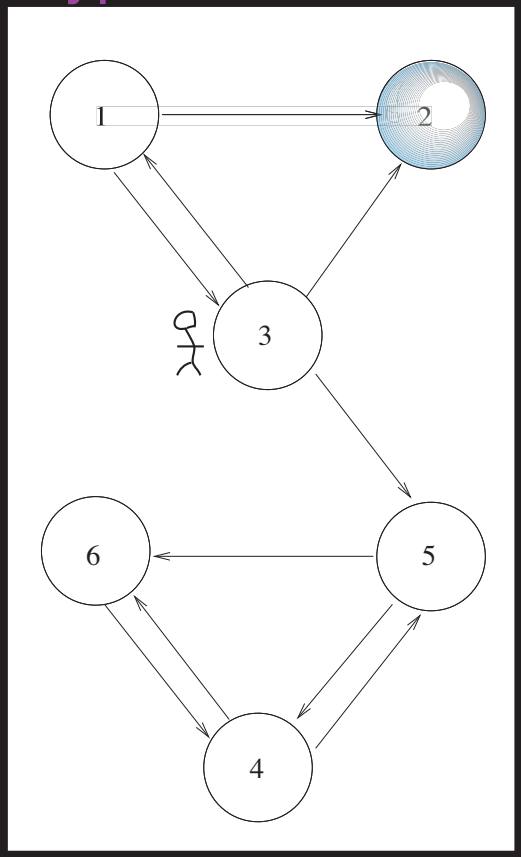
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

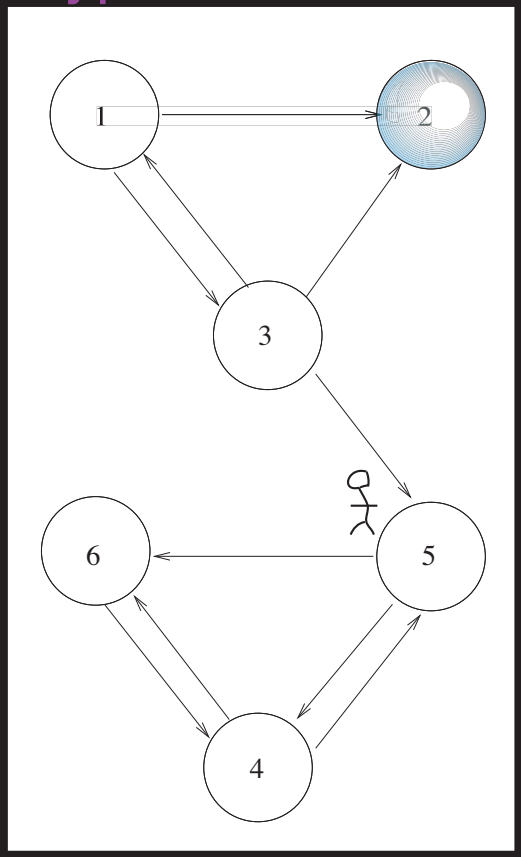
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

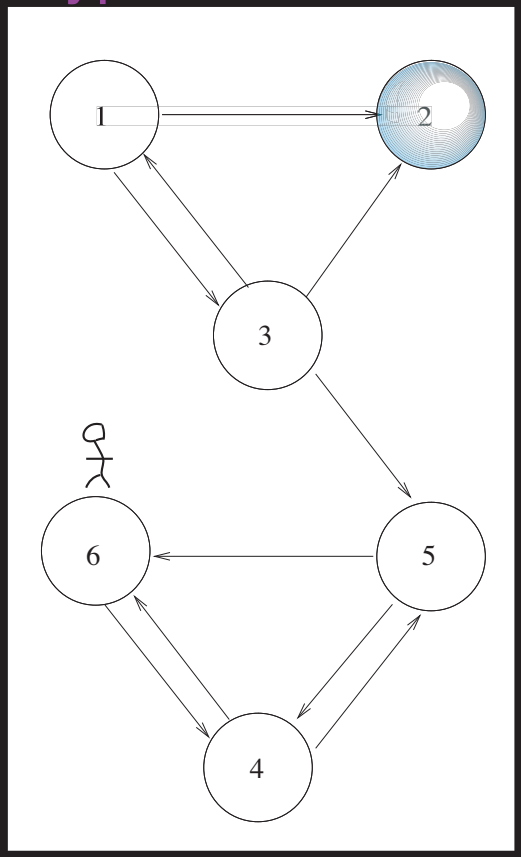
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

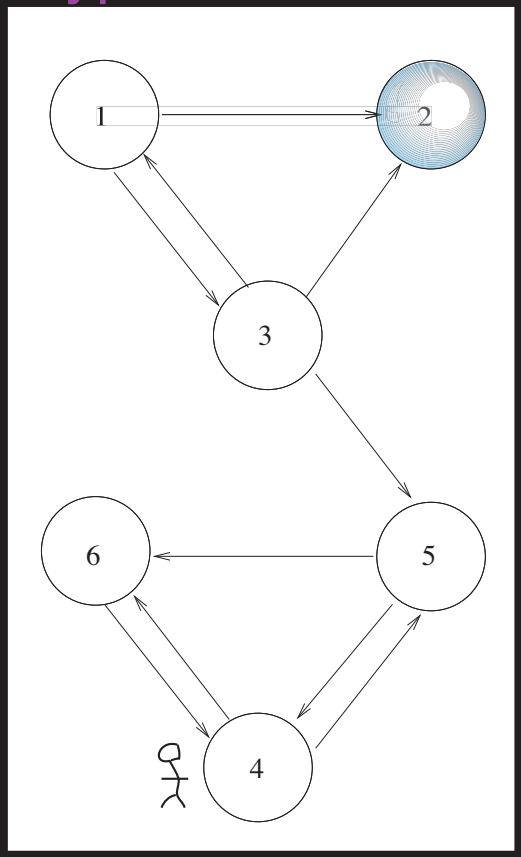
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

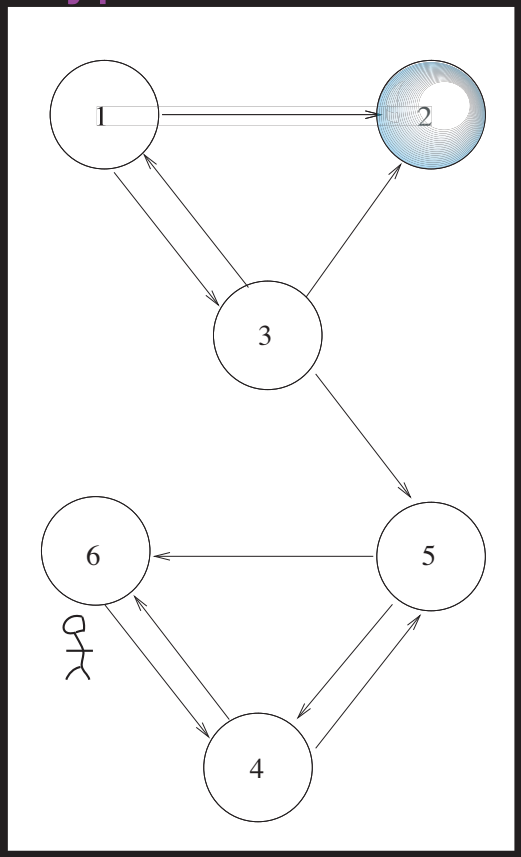
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

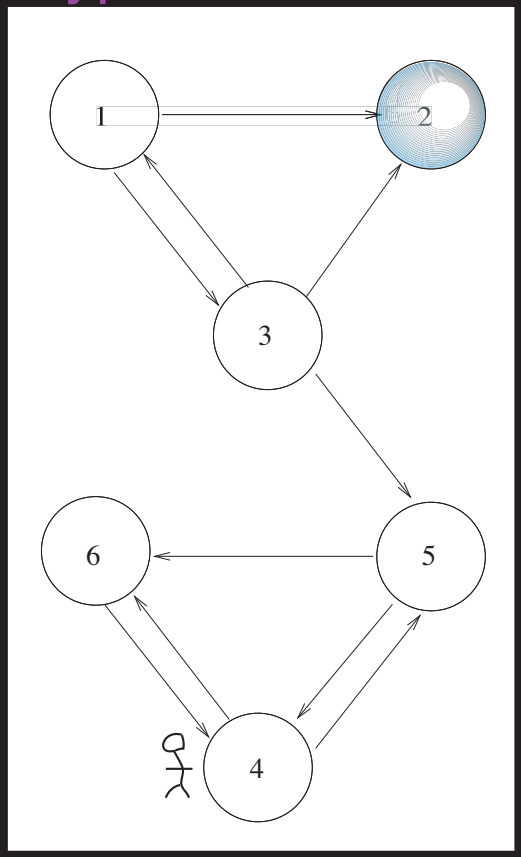
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

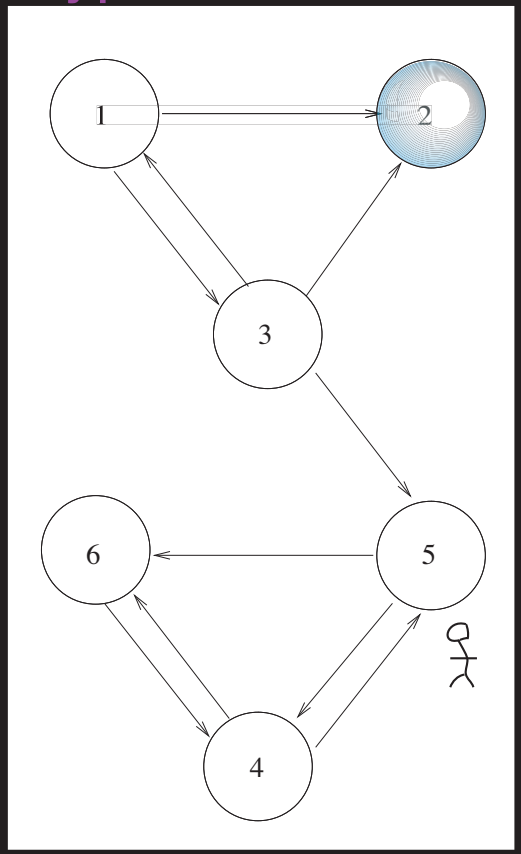
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

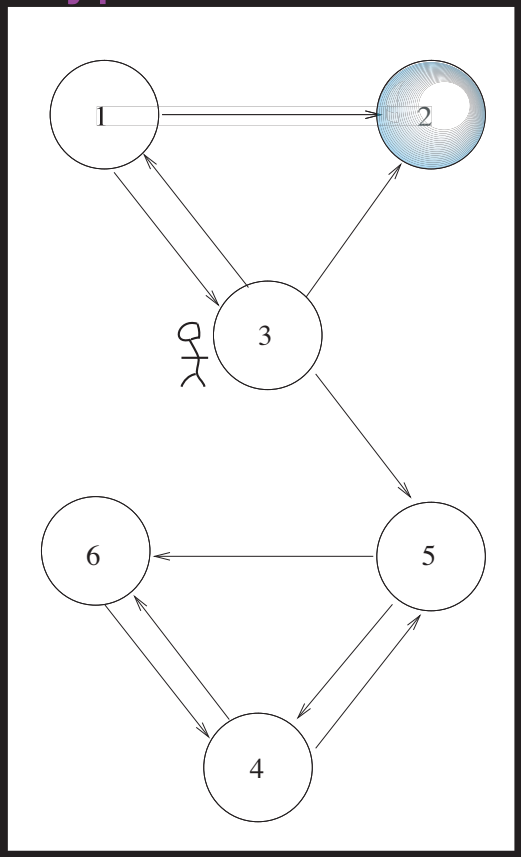
Hyperlink as vote



Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

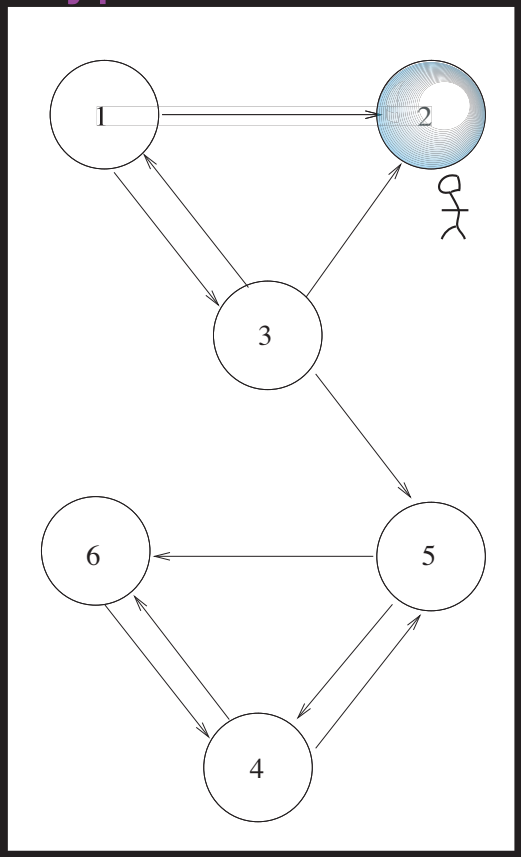
Hyperlink as vote



Ranking with a Random Surfer

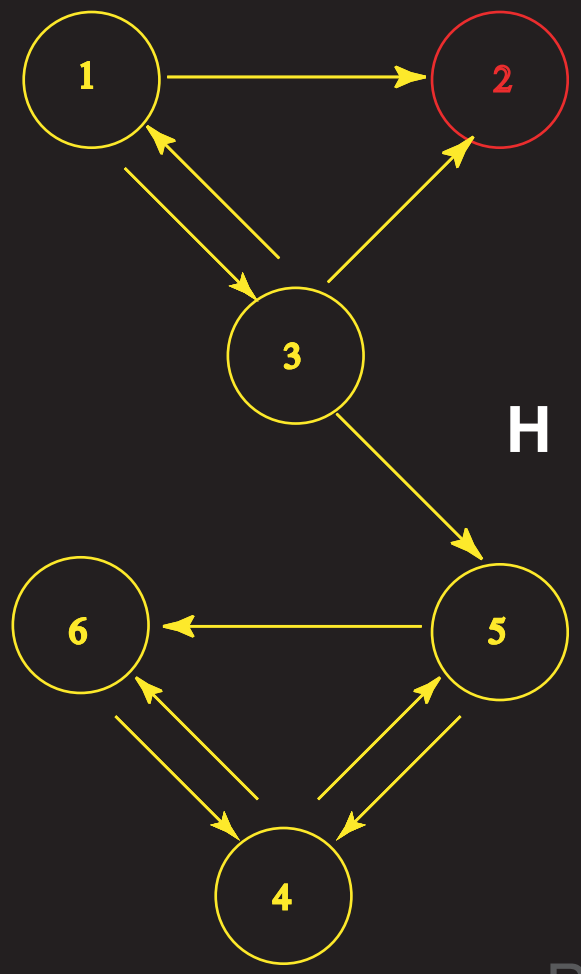
- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote



page 2 is a dangling node

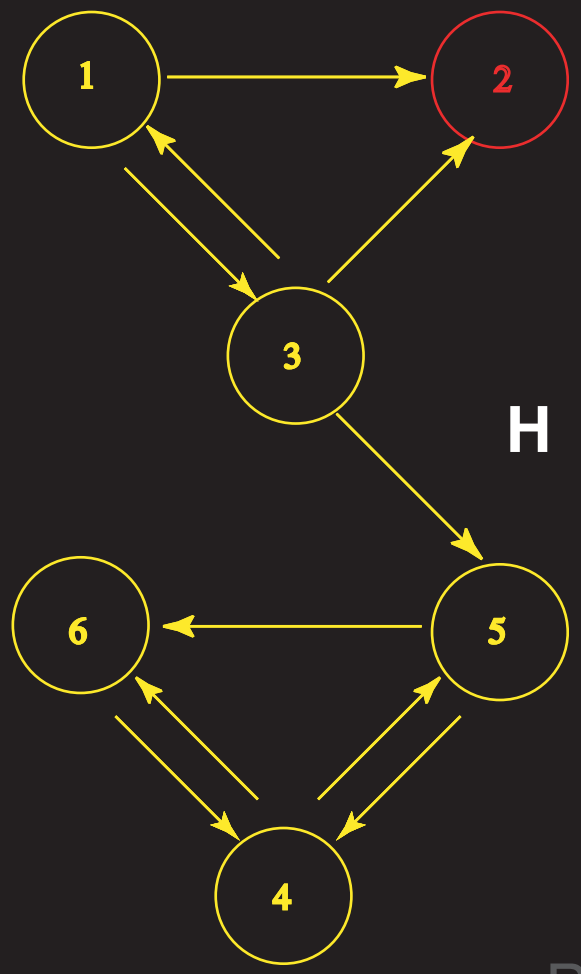
Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- A random walk on the Web Graph
- PageRank = π_i = amount of time spent at P_i
- Dead end page (nothing to click on) — a “dangling node”

Tiny Web



$$\mathbf{H} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- ▷ A random walk on the Web Graph
- ▷ PageRank = π_i = amount of time spent at P_i
- ▷ Dead end page (nothing to click on) — a "dangling node"
- ▷ $\boldsymbol{\pi}^T = (0, 1, 0, 0, 0, 0)$ = e-vector $\Longrightarrow$ Page P_2 is a "rank sink"

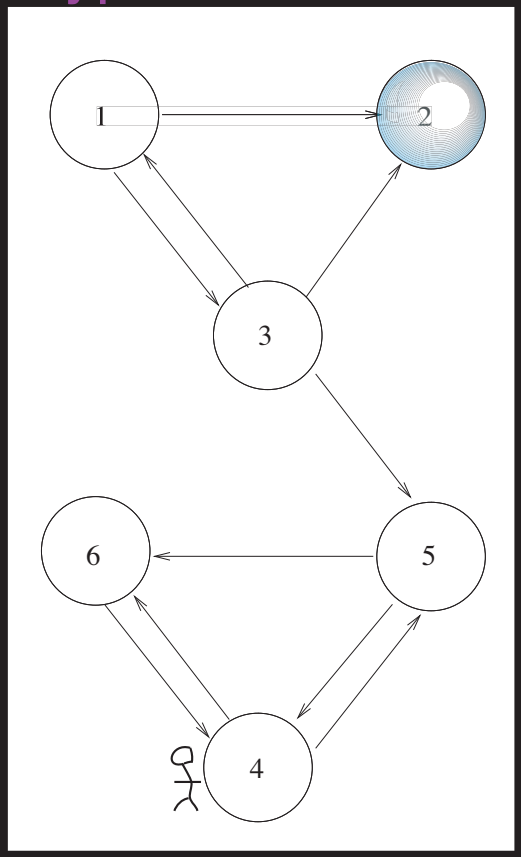
The Fix

Allow Web Surfers To Make Random Jumps

Ranking with a Random Surfer

- Rank each page corresponding to a search term by number and *quality* of votes cast for that page.

Hyperlink as vote



surfer “teleports”

The Fix

Allow Web Surfers To Make Random Jumps

- Replace zero rows with $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

$$\mathbf{S} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

The Fix

Allow Web Surfers To Make Random Jumps

- Replace zero rows with $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

$$\mathbf{S} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- $\mathbf{S} = \mathbf{H} + \frac{\mathbf{a}\,\mathbf{e}^T}{6}$ is now row stochastic $\implies \rho(\mathbf{S}) = 1$

The Fix

Allow Web Surfers To Make Random Jumps

- Replace zero rows with $\frac{\mathbf{e}^T}{n} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$

$$\mathbf{S} = \begin{array}{c} \\ P_1 \\ P_2 \\ P_3 \\ P_4 \\ P_5 \\ P_6 \end{array} \begin{array}{c} \begin{array}{cccccc} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

- $\mathbf{S} = \mathbf{H} + \frac{\mathbf{a}\,\mathbf{e}^T}{6}$ is now row stochastic $\Longrightarrow$ $\rho(\mathbf{S}) = 1$
- Perron says $\exists$ $\boldsymbol{\pi}^T \geq 0$ s.t. $\boldsymbol{\pi}^T = \boldsymbol{\pi}^T\mathbf{S}$ with $\sum_i \pi_i = 1$

Ranking with a Random Surfer

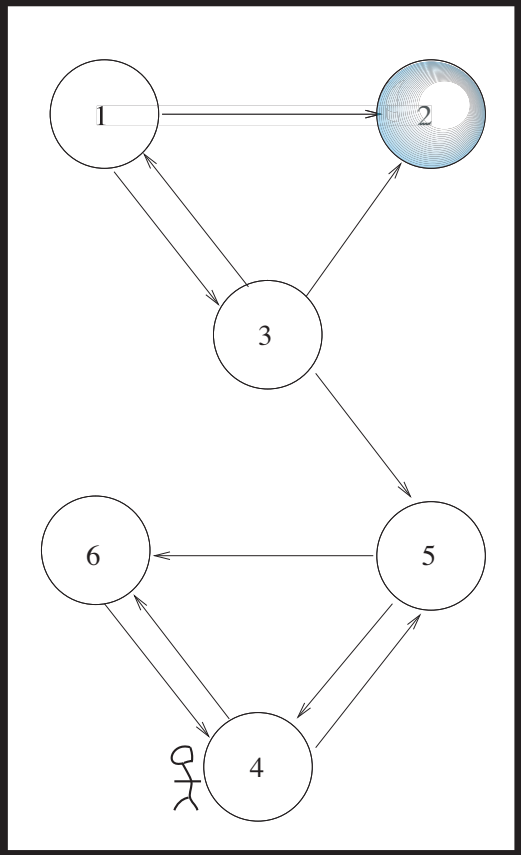
- If a page is “important,” it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or $proportion\ of\ time$, the surfer spends on each page to create ranking of webpages.

Ranking with a Random Surfer

- If a page is "important," it gets lots of votes from other important pages, which means the random surfer visits it often.
- Simply count the number of times, or *proportion of time*, the surfer spends on each page to create ranking of webpages.

Proportion of Time

Page 1 = .04
Page 2 = .05
Page 3 = .04
Page 4 = .38
Page 5 = .20
Page 6 = .29



Ranked List of Pages

Page 4
Page 6
Page 5
Page 2
Page 1
Page 3

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do? I'm a modeler.

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do? I'm a modeler.

What? A model?

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	Applied applied mathematician.

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	Applied applied mathematician.
So your degree's in math.	

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	Applied applied mathematician.
So your degree's in math.	No, it's in OR.

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	Applied applied mathematician.
So your degree's in math.	No, it's in OR.
What the hell?!?	

Mathematical Models are everywhere.

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	Applied applied mathematician.
So your degree's in math.	No, it's in OR.
What the hell?!?	You see, back in the 1950s . . .

Mathematical Models are everywhere.

embracing my OR nature

- Some problems I see myself, some are brought to me.
- I love the challenge of modeling the problem and solving it.

Less enjoyable challenge:

What do you do?	I'm a modeler.
What? A model?	No, a modeler, a math. modeler
A what?	A computational mathematician.
Huh?	Applied applied mathematician.
So your degree's in math.	No, it's in OR.
What the hell?!?	You see, back in the 1950s . . .
Uhh, nevermind.	

Sudoku

Sudoku

brought to me by Lewis Driskell, Spring 2006

Sudoku

- MATH 452/552 Class Assignment
- JOMA article with java applet

Andy Bartlett

Tim Chartier

Tim Rankin

- MATH 105 Business Calculus

Sudoku

Sudoku puzzle

		7	8		5	2		
8			6		4			5
	1			9			8	
4			2	8	9			7
5			7	6	1			2
	7			3			6	
3			1		6			4
		2	5		8	1		

Sudoku

Sudoku puzzle

		7	8		5	2		
8			6		4			5
	1			9			8	
4			2	8	9			7
5			7	6	1			2
	7			3			6	
3			1		6			4
		2	5		8	1		

Sudoku matrix

6	4	7	8	1	5	2	3	9
8	9	3	6	2	4	7	1	5
2	1	5	3	9	7	4	8	6
4	3	1	2	8	9	6	5	7
7	2	6	4	5	3	8	9	1
5	8	9	7	6	1	3	4	2
1	7	4	9	3	2	5	6	8
3	5	8	1	7	6	9	2	4
9	6	2	5	4	8	1	7	3

Sudoku

Sudoku puzzle

		7	8		5	2		
8			6		4			5
	1			9			8	
4			2	8	9			7
5			7	6	1			2
	7			3			6	
3			1		6			4
		2	5		8	1		

Sudoku matrix

6	4	7	8	1	5	2	3	9
8	9	3	6	2	4	7	1	5
2	1	5	3	9	7	4	8	6
4	3	1	2	8	9	6	5	7
7	2	6	4	5	3	8	9	1
5	8	9	7	6	1	3	4	2
1	7	4	9	3	2	5	6	8
3	5	8	1	7	6	9	2	4
9	6	2	5	4	8	1	7	3

Definition A $n \times n$ matrix is called a *Sudoku matrix* if:

1. n is a perfect square (e.g., 4, 9, 16, 25),
2. every row uses the integers 1 through n exactly once,
3. every column uses the integers 1 through n exactly once,
4. every submatrix uses the integers 1 through n exactly once.

Mathematical Model of Sudoku

Define:

$$x_{ijk} = \begin{cases} 1, & \text{if element } (i,j) \text{ of the } n \times n \text{ Sudoku matrix contains the integer } k \\ 0, & \text{otherwise.} \end{cases}$$

$$\min \quad \mathbf{0}^T \mathbf{x}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} x_{ijk} = 1, \quad j{=}1{:}n,\ k{=}1{:}n \quad \text{(only one } k \text{ in each column)} \tag{1}$$

$$\sum_{j=1}^{n} x_{ijk} = 1, \quad i{=}1{:}n,\ k{=}1{:}n \quad \text{(only one } k \text{ in each row)} \tag{2}$$

$$\sum_{j=mq-m+1}^{mq} \sum_{i=mp-m+1}^{mp} x_{ijk} = 1, \quad k{=}1{:}n,\ p{=}1{:}m,\ q{=}1{:}m \quad \text{(only one } k \text{ in each submatrix)} \tag{3}$$

$$\sum_{k=1}^{n} x_{ijk} = 1 \quad i{=}1{:}n,\ j{=}1{:}n \quad \text{(every position in matrix must be filled)} \tag{4}$$

$$x_{ijk} = 1 \quad \forall (i,j,k) \in G \quad \text{(given elements } G \text{ in matrix are set "on")} \tag{5}$$

$$x_{ijk} \in \{0,1\} \tag{6}$$

Mathematical Model of Sudoku

Define:

$$x_{ijk} = \begin{cases} 1, & \text{if element } (i,j) \text{ of the } n \times n \text{ Sudoku matrix contains the integer } k \\ 0, & \text{otherwise.} \end{cases}$$

$$\begin{aligned}
\min \quad & \mathbf{0}^T \mathbf{x} \\
s.t. \quad & \sum_{i=1}^{n} x_{ijk} = 1, \quad j{=}1{:}n,\ k{=}1{:}n \quad \text{(only one } k \text{ in each column)} && (1) \\
& \sum_{j=1}^{n} x_{ijk} = 1, \quad i{=}1{:}n,\ k{=}1{:}n \quad \text{(only one } k \text{ in each row)} && (2) \\
& \sum_{j=mq-m+1}^{mq} \sum_{i=mp-m+1}^{mp} x_{ijk} = 1, \quad k{=}1{:}n,\ p{=}1{:}m,\ q{=}1{:}m \quad \text{(only one } k \text{ in each submatrix)} && (3) \\
& \sum_{k=1}^{n} x_{ijk} = 1 \quad i{=}1{:}n,\ j{=}1{:}n \quad \text{(every position in matrix must be filled)} && (4) \\
& x_{ijk} = 1 \quad \forall (i,j,k) \in G \quad \text{(given elements } G \text{ in matrix are set "on")} && (5) \\
& x_{ijk} \in \{0,1\} && (6)
\end{aligned}$$

Value of the Model

With a computer algorithm, we can solve any Sudoku puzzle, regardless of:

size n
number of givens
level of difficulty

9×9 puzzle takes 16.7 seconds to solve on desktop machine.

Unique Solution?

- Most puzzle creators do not check whether their puzzle has one unique solution.

Puzzle

2	2
3	4

Unique Solution?

- Most puzzle creators do not check whether their puzzle has one unique solution.

Puzzle

	2		
			2
3			
			4

Solution 1

1	**2**	3	4
4	3	**2**	1
3	4	1	2
2	1	**4**	3

Solution 2

4	**2**	3	1
1	3	**2**	4
3	4	1	2
2	1	**4**	3

Some Interesting 9×9 Sudoku Facts

- How many 9×9 matrices deserve the title of Sudoku matrices?

  $6{,}670{,}903{,}752{,}021{,}072{,}936{,}960 \approx 6.67 \times 10^{21}$

- What is the fewest number of givens that must be provided to create a 9×9 puzzle with a unique solution?

  17; 35,396 distinct puzzles with 17 givens and a unique solution have been found. No unique solution puzzle with ≤ 16 givens has been found yet.

- Given one Sudoku matrix, could I make my own Daily Sudoku Calendar?

Puzzle

	2		
			1
3			
		4	

Unique Solution

1	**2**	3	4
4	3	2	**1**
3	4	1	2
2	1	**4**	3

$1 \Leftrightarrow 4$

Puzzle

	2		
			4
3			
		1	

Unique Solution

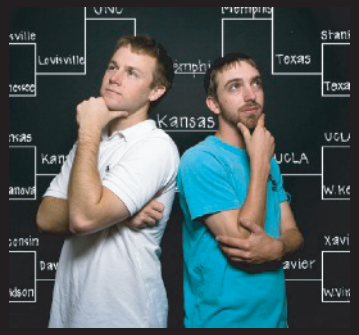
4	**2**	3	1
1	3	2	**4**
3	1	4	2
2	4	**1**	3

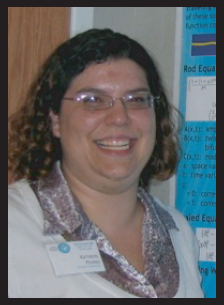
By using mathematical operations 362,879 (≈ 991 years worth of) Sudoku matrices can be created from one 9×9 Sudoku matrix.

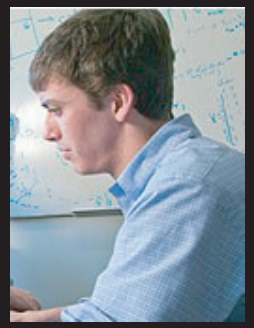
Ranking Sports Teams

Ranking Teams

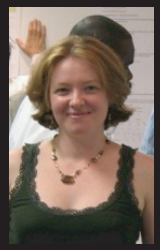
- MATH 452/552 Class Project = March Madness
  - 2006: Luke Ingram and John McConnell
  - 2008: Neil Goodson and Colin Stephenson
  - 2009: Kathryn Pedings, Ryan Parker, Erich Kreutzer, Max Win, Tim Chartier, Nick Dovidio, Yoshi Yamamoto







- M.S. Thesis 2007: Luke Ingram
- Ph.D. Thesis 2008: Anjela Govan

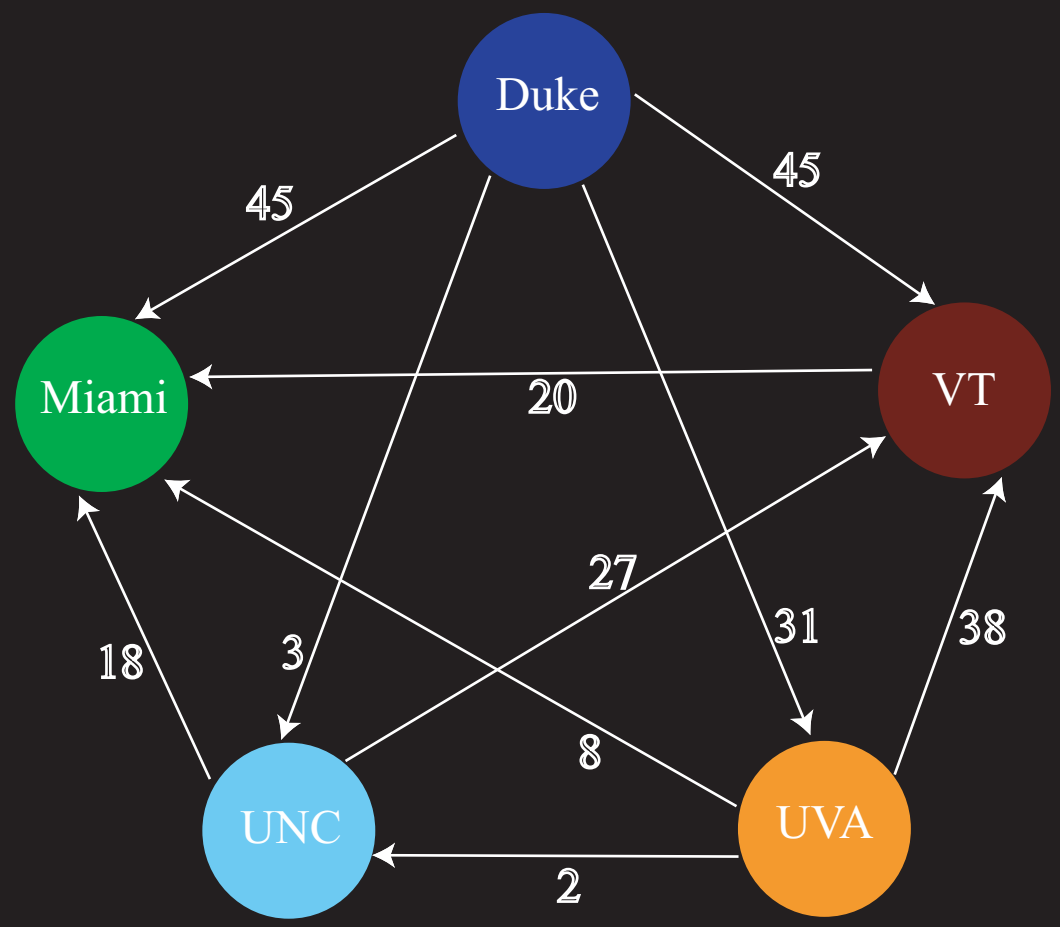


Ranking Teams

PageRank applied to Sports

(joint work with Carl Meyer)

- webpages vote with hyperlinks
- losing teams vote with point differentials

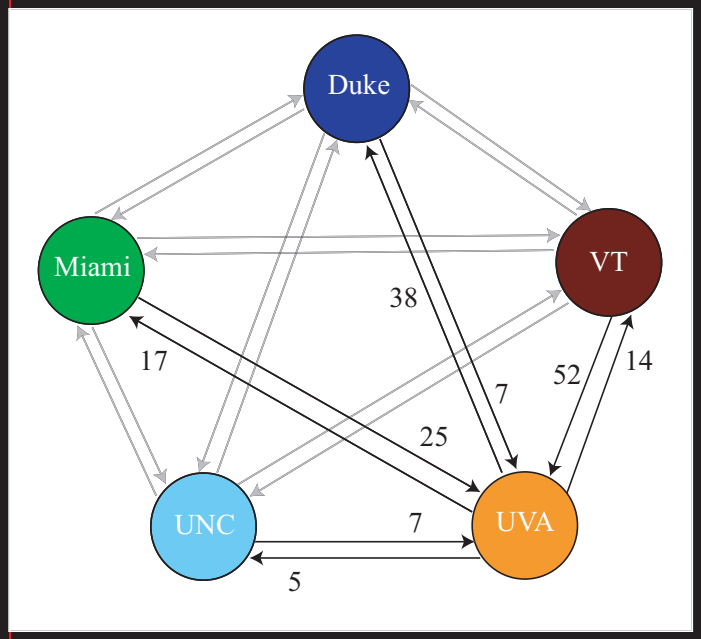


- not as successful at ranking and predicting winners as mHITS

mHITS

each team i gets both offensive rating $\mathbf{o}_i$ and defensive rating $\mathbf{d}_i$

- **mHITS Thesis:** A team is a good defensive team (i.e., deserves a high defensive rating d_j) when it holds its opponents (particularly strong offensive teams) to low scores. A team is a good offensive team (i.e., deserves a high offensive rating o_i) when it scores many points against its opponents (particularly opponents with high defensive ratings).



Graph

$$\mathbf{P} = \begin{array}{c} \\ \text{Duke} \\ \text{Miami} \\ \text{UNC} \\ \text{UVA} \\ \text{VT} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{Duke} & \text{Miami} & \text{UNC} & \text{UVA} & \text{VT} \end{array} \\ \begin{pmatrix} 0 & 7 & 21 & 7 & 0 \\ 52 & 0 & 34 & 25 & 27 \\ 24 & 16 & 0 & 7 & 3 \\ 38 & 17 & 5 & 0 & 14 \\ 45 & 7 & 30 & 52 & 0 \end{pmatrix} \end{array}$$

Point Matrix $\mathbf{P} \geq 0$

mHITS Equations

Summation Notation

$$d_j = \sum_{i \in I_j} p_{ij} \frac{1}{o_i} \qquad \text{and} \qquad o_i = \sum_{j \in L_i} p_{ij} \frac{1}{d_j}$$

Matrix Notation: iterative procedure

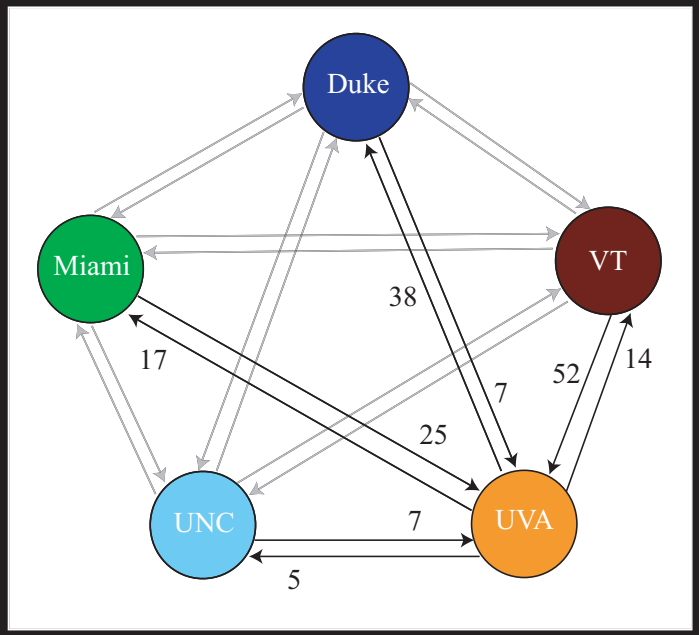
$$\mathbf{d}^{(k)} = \mathbf{P}^T \frac{1}{\mathbf{o}^{(k)}} \qquad \text{and} \qquad \mathbf{o}^{(k)} = \mathbf{P} \frac{1}{\mathbf{d}^{(k-1)}}$$

- related to the Sinkhorn-Knopp algorithm for matrix balancing
  (uses successive row and column scaling to transform $\mathbf{P} \geq 0$ into doubly stochastic matrix $\mathbf{S}$)
- P. Knight uses Sinkhorn-Knopp algorithm to rank webpages

mHITS Results: tiny NCAA

(data from Luke Ingram)

Team	Off. Rating o	Off. Rank	Def. Rating d	Def. Rank
Duke	.4	5th	140	5th
Miami	1.8	1st	67	2nd
UNC	.6	4th	97	4th
UVA	1.0	3rd	80	3rd
VT	1.4	2nd	34	1st



	r = o/d Rating	r = o/d Rank
Duke	0.003	5th
Miami	0.027	2nd
UNC	0.006	4th
UVA	0.012	3rd
VT	0.041	1st

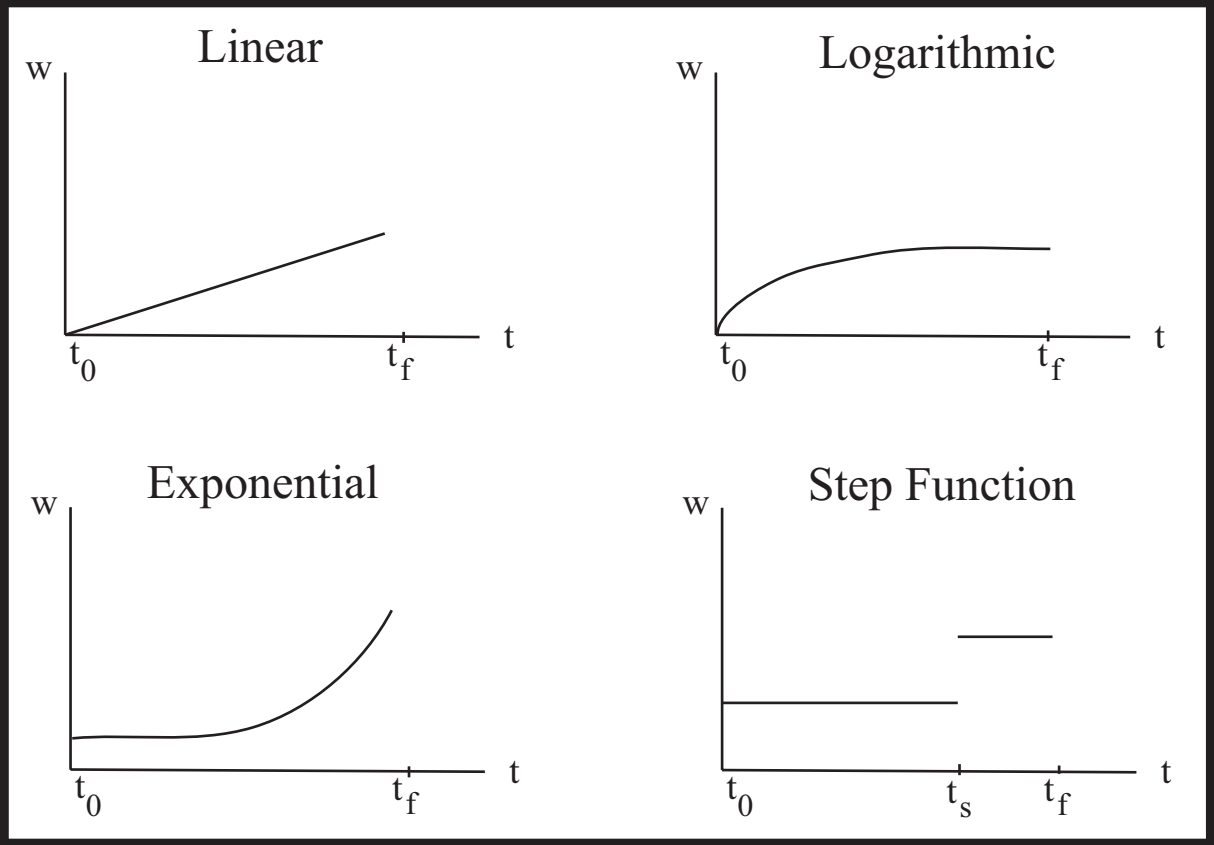
Weighted mHITS

- Weighted Point Matrix $\bar{\mathbf{P}} \geq 0$

$$\bar{\mathbf{p}}_{ij} = w_{ij} \ \mathbf{p}_{ij}$$

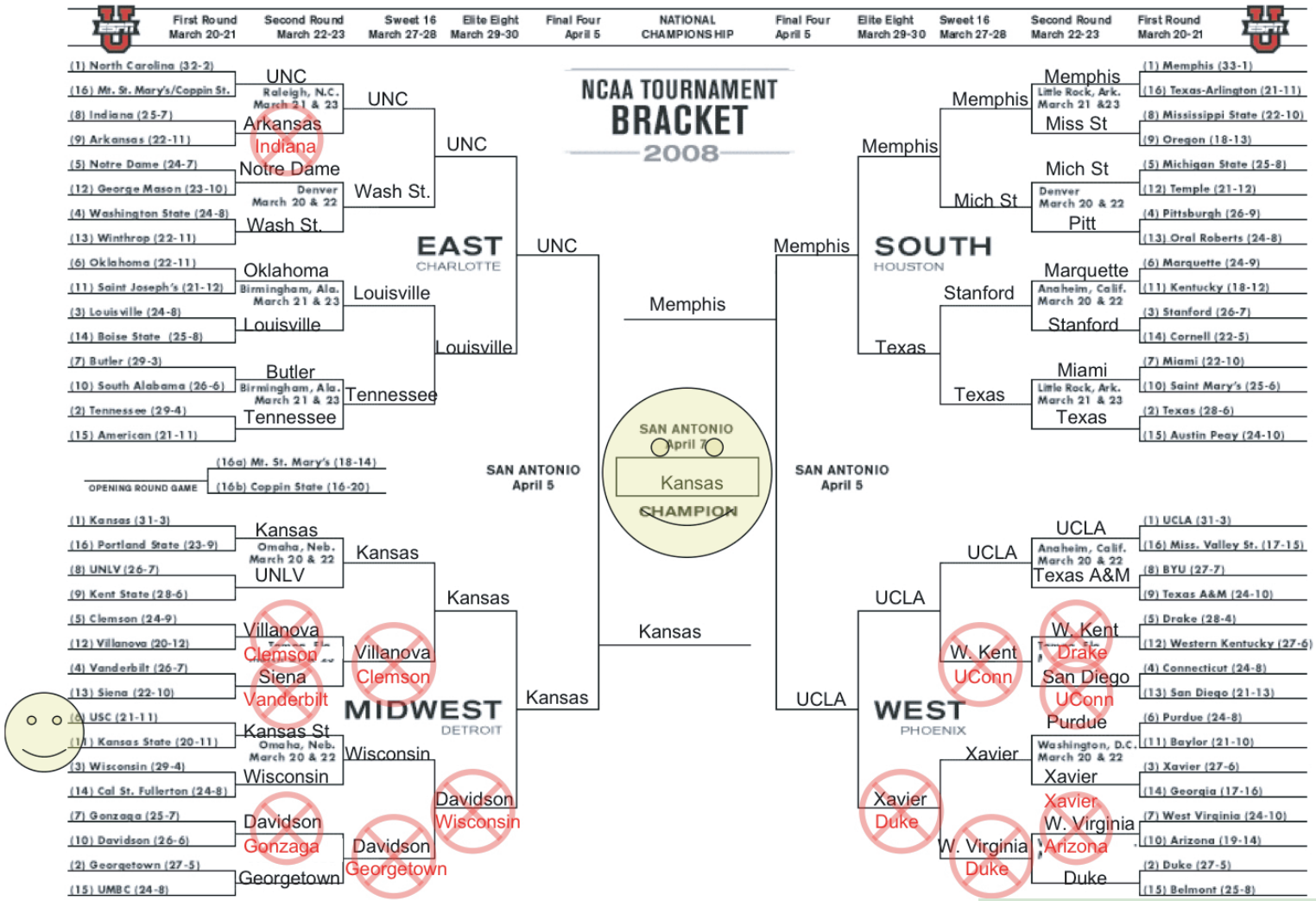
(w_{ij} = weight of matchup between teams i and j)

- possible weightings w_{ij}



mHITS Results: full NCAA

(image from Neil Goodson and Colin Stephenson)



mHITS - Log Weighted

ESPN Pts: 1450 out of 1680
National Rank: 834 out 3.6 mil.
Percentile: 100%`

mHITS Results: full NCAA

Method	ESPN score
Massey Linear	1450
Massey Log	1450
mHITS Log	1450
Massey Step	1420
Massey Exponential	1400
mHITS Step	1320
Massey Uniform	1310
mHITS Linear	1310
mHITS Uniform	1310
Colley Linear	1100
Colley Log	1010

- weightings can easily be applied to all ranking methods
  - $\Rightarrow$ interesting possibilities for weighted webpage ranking

mHITS Point Matrix $\mathbf{P} \geq 0$

Perron-Frobenius guarantees (for irreducible $\mathbf{P}$ with total support)

- existence of $\mathbf{o}$ and $\mathbf{d}$
- uniqueness of $\mathbf{o}$ and $\mathbf{d}$
- convergence of mHITS algorithm
- rate of convergence of mHITS algorithm

  $\sigma_2^2(\mathbf{S})$, where $\mathbf{S} = \mathbf{D}\,(1/\mathbf{o})\;\mathbf{P}\;\mathbf{D}\,(1/\mathbf{d})$

A New Algorithm for Ranking Sports Teams Using Evolutionary Optimization

Kathryn Pedings (kathryn.pedings@gmail.com) and Dr. Amy Langville
College of Charleston Department of Mathematics

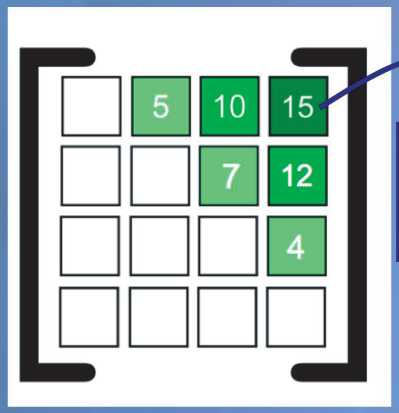
COLLEGE of CHARLESTON
THE GRADUATE SCHOOL

Abstract

Evolutionary optimization is an algorithm using Darwinian ideas of mating, mutating, fitness, and survival of the fittest. Its use has been limited to intractable problems, but it was our belief that the algorithm could be modified to be successful with tractable problems such as sports ranking.

1. Never before used on sports ranking.
2. We have shown that with some datasets our evolutionary algorithm is better than other well-known algorithms.
3. Sophisticated changes made to speed up the algorithm.

Hillside Form and Point-Differential Matrices

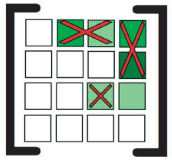


Team 1 beat Team 4 by 15 points.

Initial Population

$$\begin{bmatrix} 2 & 3 & 1 & 1 \\ 4 & 2 & 2 & 3 \\ 1 & 1 & 4 & 4 \\ 3 & 4 & 3 & 2 \end{bmatrix}$$

Calculate the Fitness



Get an Offspring

Mating

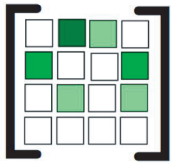
$$\begin{bmatrix} 2 \\ 4 \\ 1 \\ 3 \end{bmatrix} \text{ and } \begin{bmatrix} 3 \\ 2 \\ 1 \\ 4 \end{bmatrix} \rightarrow \begin{bmatrix} 2 \\ 3 \\ 1 \\ 4 \end{bmatrix}$$

Mutating

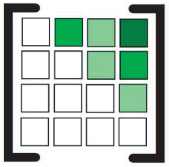
$$\begin{bmatrix} 1 \\ 2 \\ 4 \\ 3 \end{bmatrix} \rightarrow \begin{bmatrix} 4 \\ 2 \\ 1 \\ 3 \end{bmatrix}$$

Then calculate the fitness one step at a time.

Fitness is Bad



Fitness is Good



Population Changes Drastically

$$\begin{bmatrix} 1 & 3 & 2 & 4 \\ 3 & 2 & 1 & 3 \\ 2 & 4 & 3 & 1 \\ 4 & 1 & 4 & 2 \end{bmatrix}$$

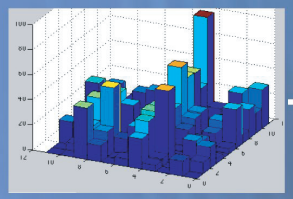
Population Changes Very Little

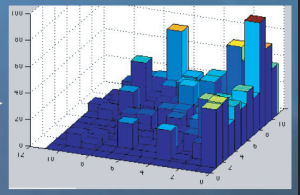
$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 3 & 3 & 2 & 4 \\ 2 & 4 & 3 & 3 \\ 4 & 2 & 4 & 2 \end{bmatrix}$$

STOP! Print out the best parent solution; this is your team ranking!

Results

The dataset used is the basketball data for the Southern Conference for the 2007-2008 season.





Massey Ranking	Evolutionary Optimization	Colley Ranking
# of Viol = 269	# of Viol = 265	# of Viol = 278
Davidson	Davidson	Davidson
UNC - G	UNC - G	App St.
GA So.	GA So.	Chatt.
App St.	App St.	GA So.
Chatt.	Chatt.	UNC - G
CofC	Elon	Elon
Elon	CofC	CofC
Wofford	W. Car.	Wofford
W. Car.	Wofford	Furman
Furman	Furman	W. Car.
Citadel	Citadel	Citadel

Future Work

- Use to rank other items such as books, movies, graduate school student applicants, etc.
- Enter a bracket to ESPN for March Madness.

Acknowledgements

- Emmie Douglas for her input and scholarly advice.
- Dr. Amy Langville for her original Evolutionary Optimization code, datasets, and continued guidance throughout the research process.

Work Cited

Michalewicz, Zbigniew, and David B. Fogel. How to Solve It. New York: Springer, 2004.

Goodson, Neil, and Colin Stephenson. "March "Mathness": Sports Ranking Using Linear Algebra." Math 481 Final Presentation. College of Charleston, Charleston, SC. 17 Apr. 2008.

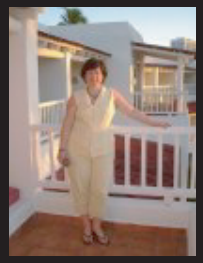
Recommendation Systems

Recommendation Systems

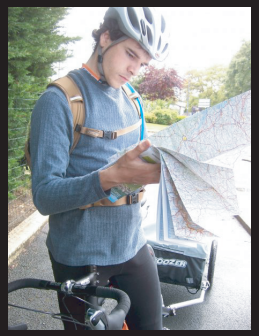
Can you cluster on a directed graph?—Leonid Zhukov, Yahoo! Research, 2005

Clustering

- MATH 552 Project: Barbara Ball and Clare Rodgers, 2006



- B.S. Thesis: Ibai Basabe, 2006

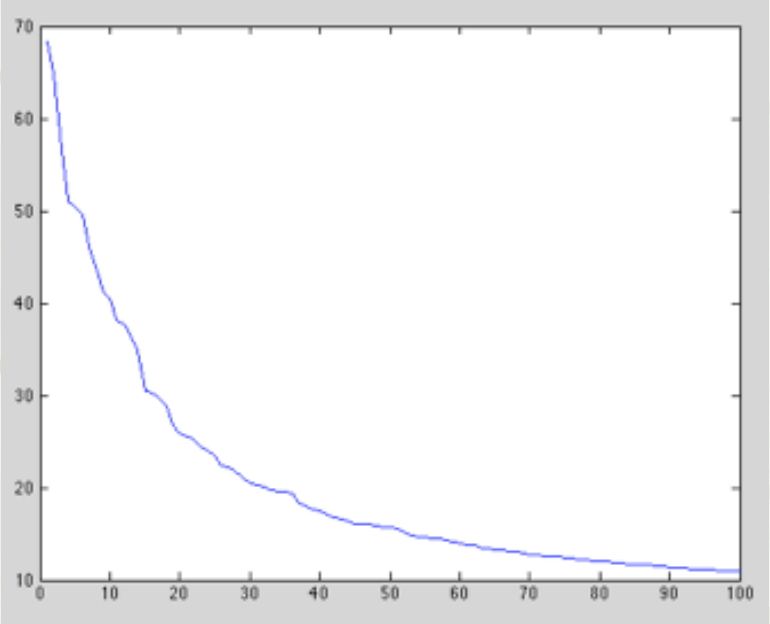


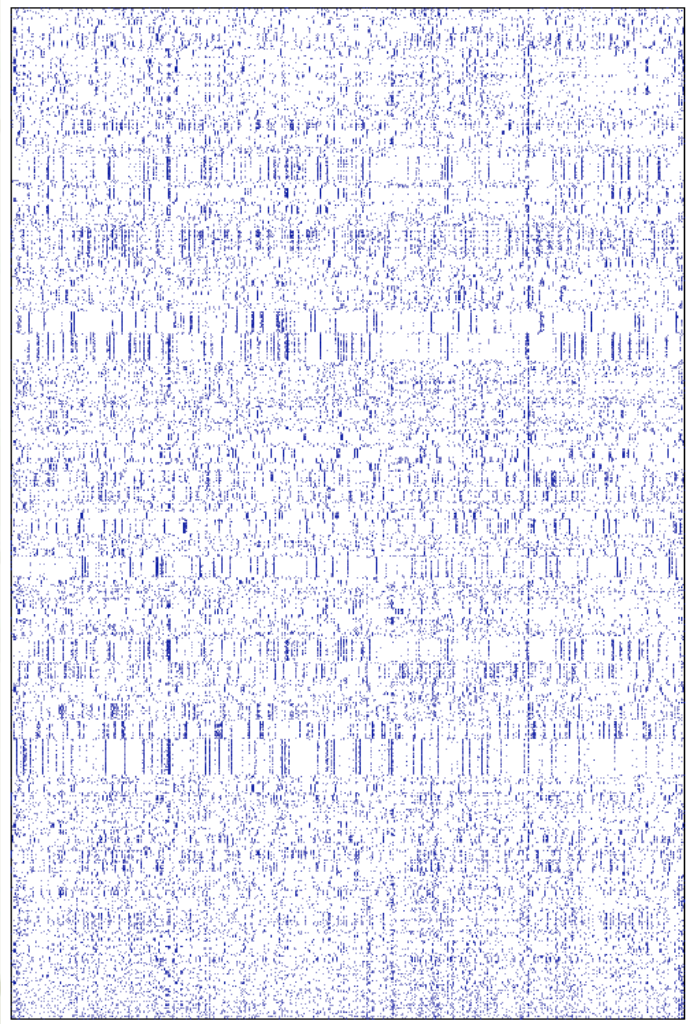
- M.S. Thesis: Barbara Ball, 2007
- M.S. Thesis: Ibai Basabe, 2007
- M.S. Thesis: Emmie Douglas, 2008

- Independent Studies: Emmie Douglas, Kathryn Pedings, Patrick Moran

Raw Yahoo! Dataset

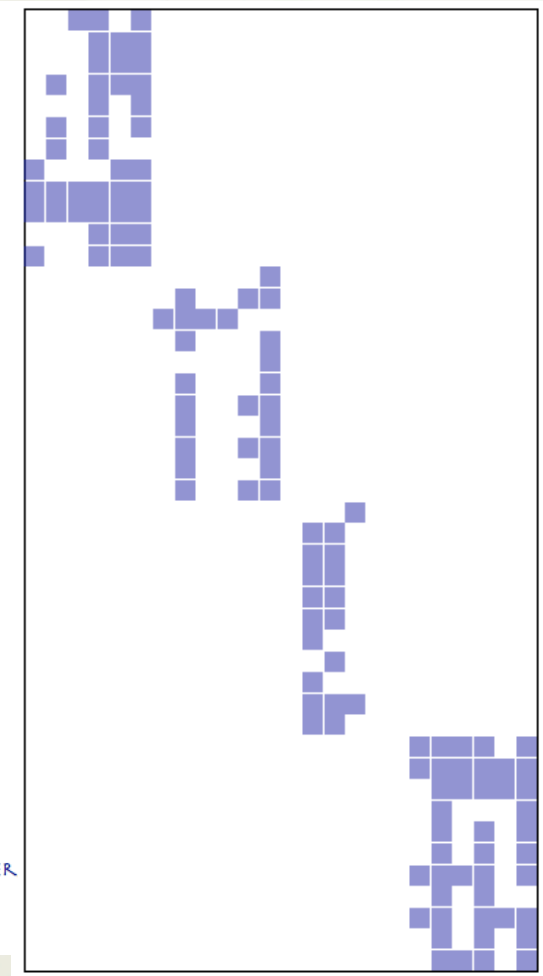
- 3,000 phrases by 2,000 advertisers

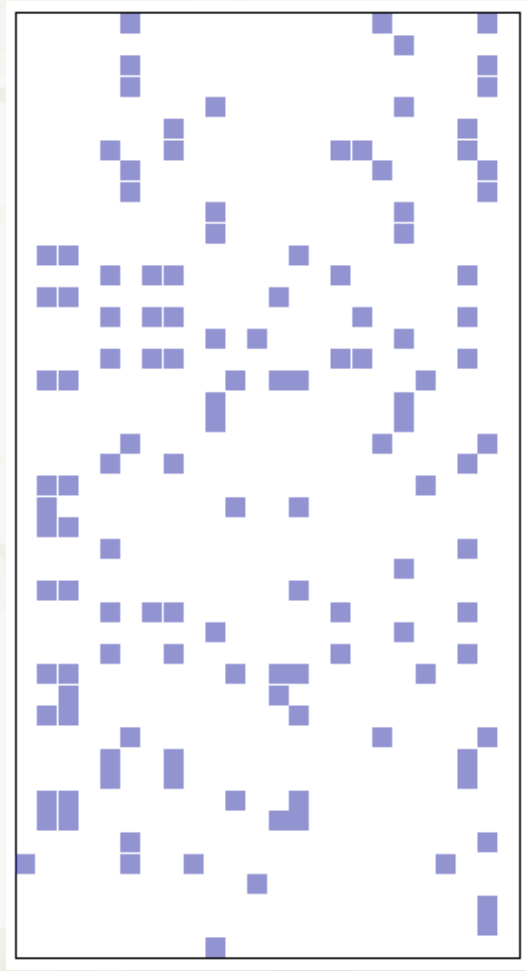




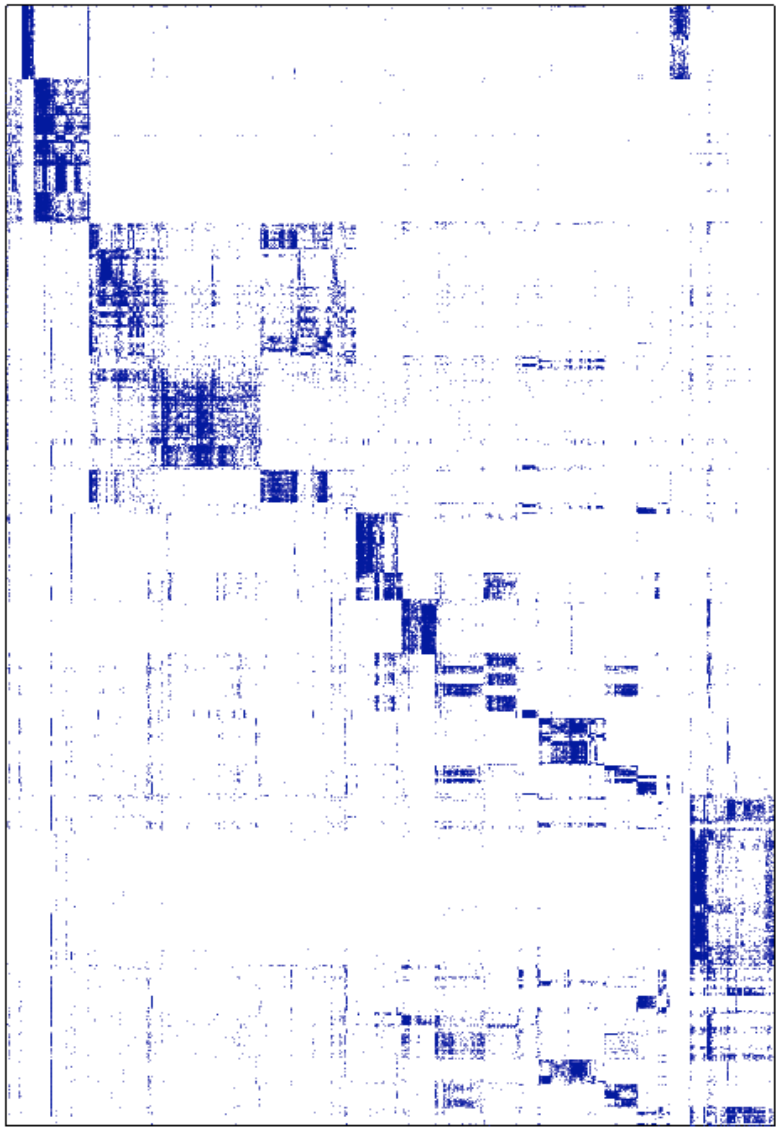
Note about Small Yahoo!

FL hotel Miami;
FL hotel Orlando;
Flagstaff hotel;
Florida fort hotel Lauderdale;
Florida hotel in Jacksonville;
Florida hotel Jacksonville;
Florida hotel key west;
Florida hotel naples;
Florida hotel Orlando;
Florida hotel Sarasota;
Florida hotel Tampa;
Florida Orlando ;
Best gambling site;
Betting black jack;
Betting casino jack;
Betting casino jack poker;
Black free game jack;
Black free jack;
Casino gambling;
Casino gambling internet;
Casino gambling online;
Casino game;
Casino game online;
Casino internet;
Cash earn extra;
Company mlm;
Computer make money;
Earn extra income money;
Earn extra income;
Earn extra money;
Earn from home money;
Earn home money;
Earn internet money;
Earn money;
CARTRIDGE EPSON;
CARTRIDGE EPSON INK;
CARTRIDGE EPSON INKJET;
CARTRIDGE HEWLETT INK PACKARD;
CARTRIDGE HEWLETT INKJET PACKARD;
CARTRIDGE HEWLETT PACKARD;
CARTRIDGE HEWLETT PACKARD PRINTER
CARTRIDGE HP;
CARTRIDGE HP INK;
CARTRIDGE HP INKJET;
CARTRIDGE HP PRINTER;



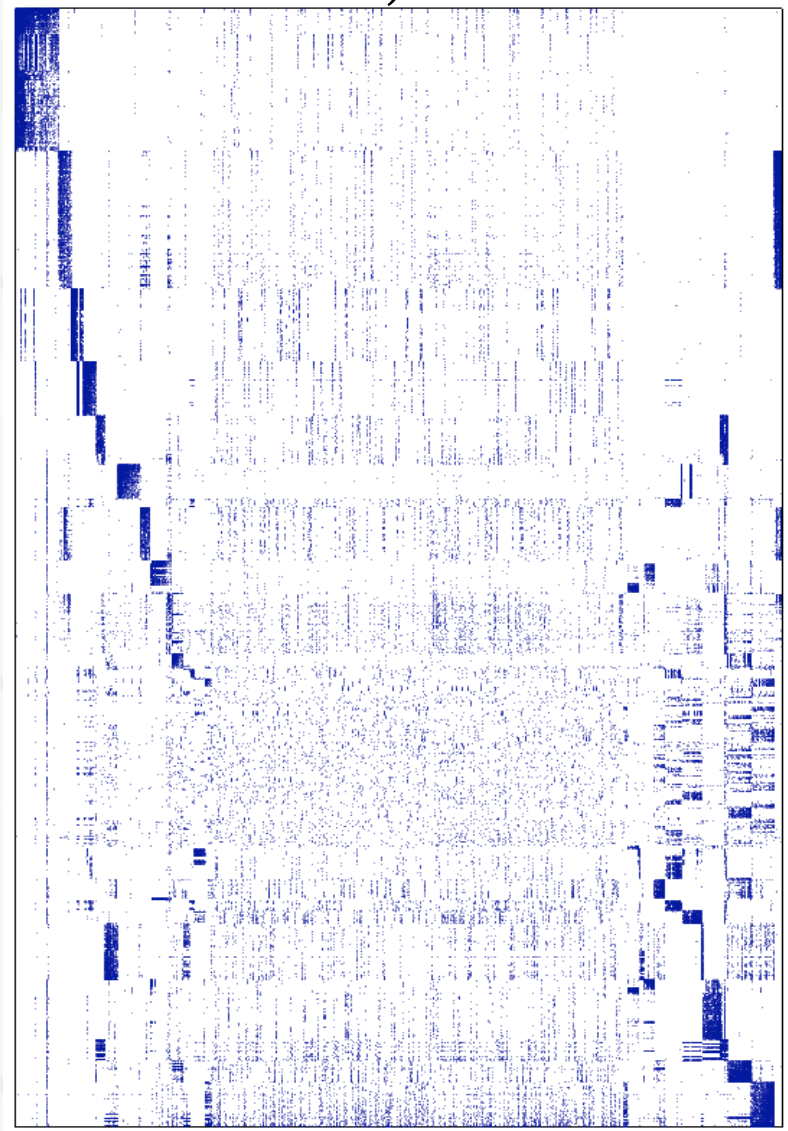


SVD Signs
k=25



Row entropy: 0.0555
Column entropy: 0.0482

SVD Gaps
k=25; tol = 2.3

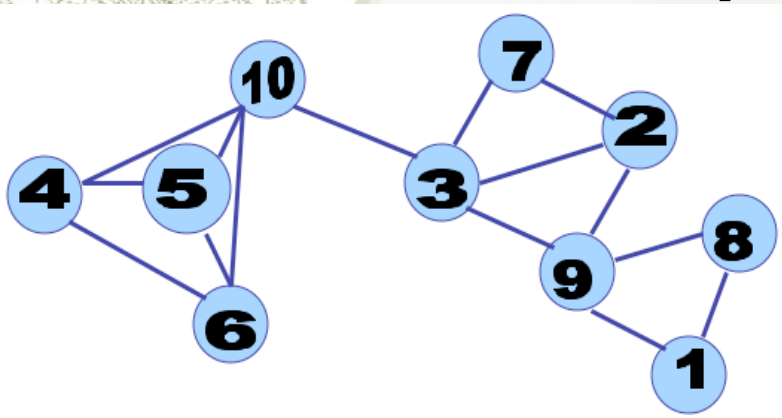


Row entropy:0.1023
Column entropy: 0.1421

To be Discussed:

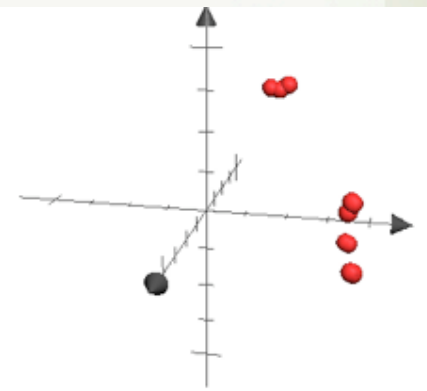
- Brief review of Fiedler Clustering Methods
- SVD based clustering methods
  - *SVD Signs*
  - *SVD Gaps*
- Measuring cluster “goodness”
- Cluster Aggregation
- Experiments with Large Datasets

Example Datasets Used

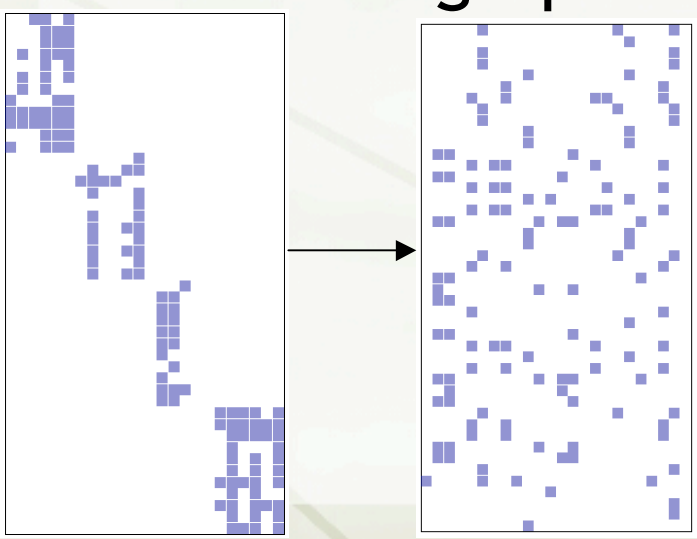


10 node graph

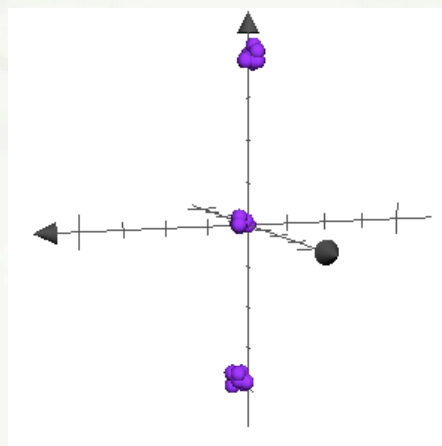
0.9218	0.9355	0.0579
0.9257	0.9301	0.0611
0.9187	0.9399	0.0545
0.7382	0.9169	0.3529
0.7367	0.9098	0.3597
0.7275	0.8936	0.3006
0.1763	0.4103	0.8132
0.1764	0.5087	0.8326
0.1758	0.4607	0.8046
0.4057	0.8936	0.0099
0.4026	0.8763	0.0167



12 points in space



Small Yahoo!

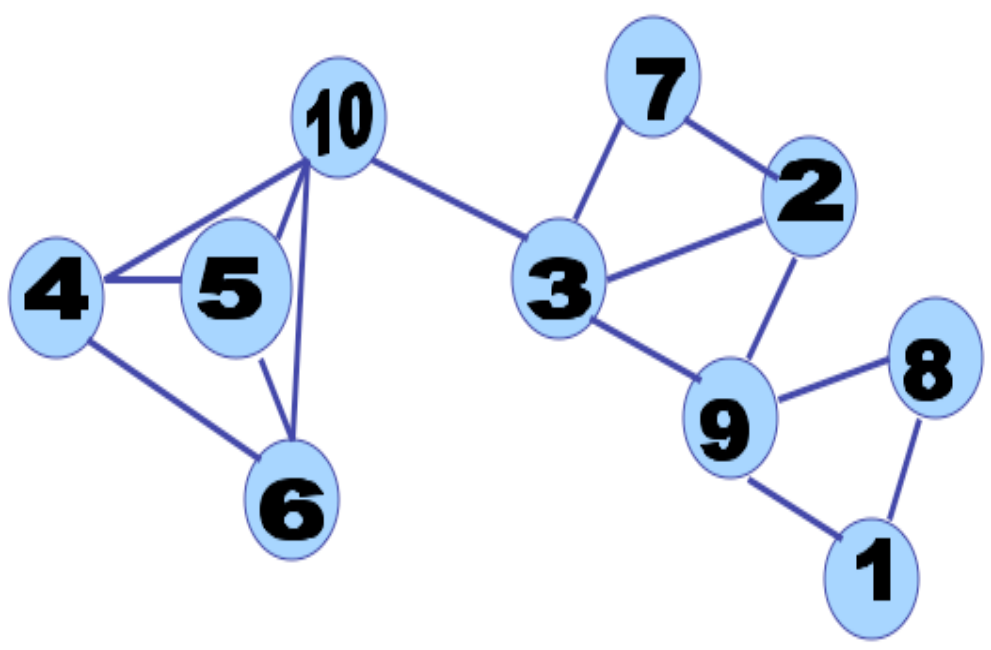


Tri-modal dataset

Origins

- SVD based methods rooted in two graph theoretic clustering methods
  - *Fiedler Method*
  - *Extended Fiedler Method*
- Both use sign patterns of eigenvectors to cluster

Connection to Graph Theory



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Finding the Laplacian

- L = D - A

$$\underbrace{\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 \\ 0 & 3 & -1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\ 0 & -1 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & -1 \\ 0 & 0 & 0 & 3 & -1 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 3 & -1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & -1 & 3 & 0 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & -1 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & 4 & 0 \\ 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \end{bmatrix}}_{L} = \underbrace{\begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix}}_{D} - \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}}_{A}$$

What is a Fiedler Vector?

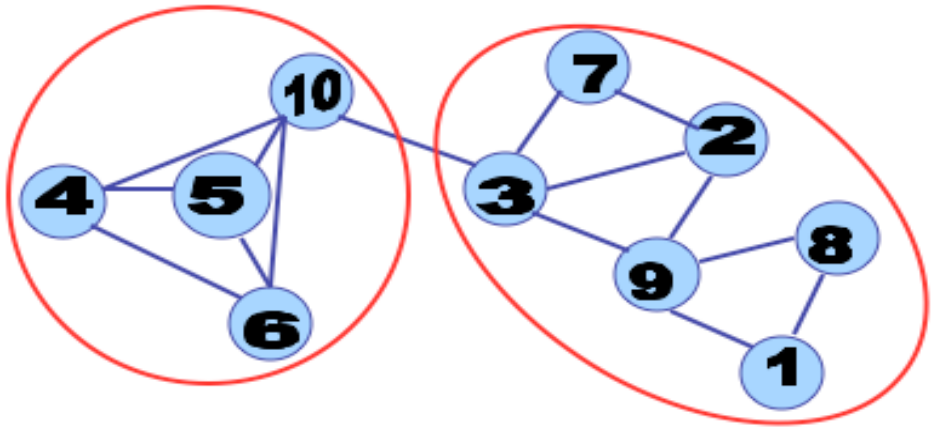
- Eigenvector corresponding to the second smallest eigenvalue of the Laplacian
- Named for Miroslav Fiedler

$$P = \begin{bmatrix} 0.32 & 0.38 & 0.43 & 0.71 & -0.19 & -0.00 & 0.00 & 0.00 & 0.17 & -0.09 \\ 0.32 & 0.19 & -0.39 & -0.00 & 0.45 & 0.06 & -0.51 & 0.18 & 0.45 & 0.01 \\ 0.32 & 0.09 & -0.29 & -0.00 & 0.28 & -0.06 & 0.51 & -0.18 & -0.11 & -0.64 \\ 0.32 & -0.40 & 0.11 & -0.00 & -0.05 & 0.38 & 0.11 & 0.73 & -0.11 & -0.11 \\ 0.32 & -0.40 & 0.11 & 0.00 & -0.05 & 0.47 & -0.35 & -0.60 & -0.11 & -0.11 \\ 0.32 & -0.40 & 0.11 & -0.00 & -0.05 & -0.79 & -0.27 & 0.05 & -0.11 & -0.11 \\ 0.32 & 0.17 & -0.59 & 0.00 & -0.69 & -0.00 & 0.00 & 0.00 & -0.11 & 0.17 \\ 0.32 & 0.38 & 0.43 & -0.71 & -0.19 & 0.00 & 0.00 & -0.00 & 0.17 & -0.09 \\ 0.32 & 0.29 & 0.05 & -0.00 & 0.38 & 0.00 & -0.00 & -0.00 & -0.68 & 0.45 \\ 0.32 & -0.29 & 0.01 & -0.00 & 0.11 & -0.06 & 0.51 & -0.18 & 0.45 & 0.54 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.26 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.86 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.06 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.81 \end{bmatrix}$$

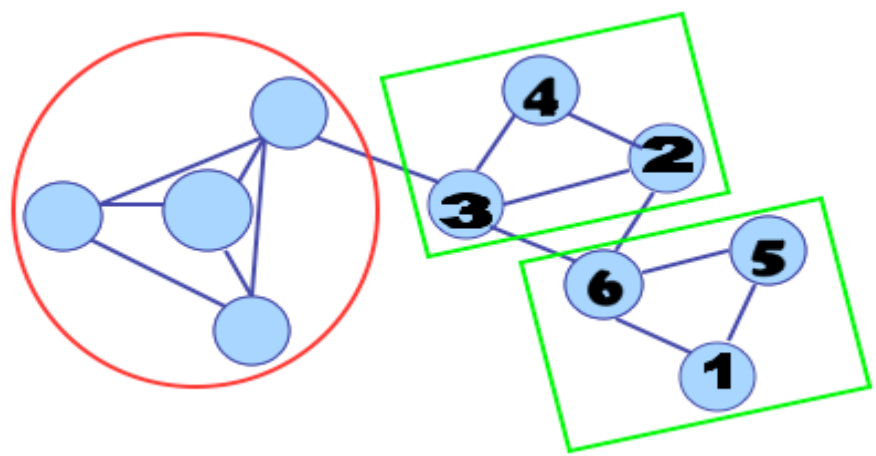
- Signs can be used to bisect the graph

Applying the Fiedler Method

$$\begin{bmatrix} + \\ + \\ + \\ - \\ - \\ - \\ + \\ + \\ + \\ - \end{bmatrix}$$



First Iteration



$$\begin{bmatrix} - \\ + \\ + \\ + \\ - \\ - \end{bmatrix}$$

Second Iteration

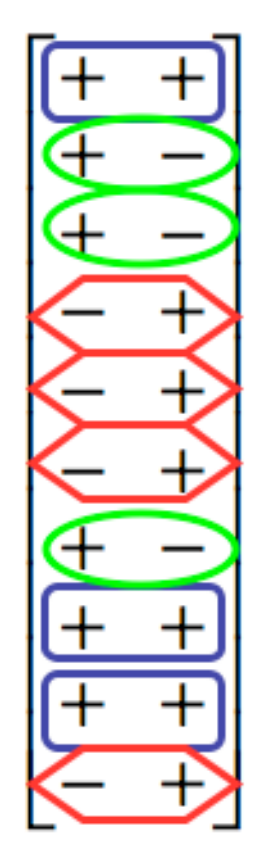
Some Problems with the Fiedler Method

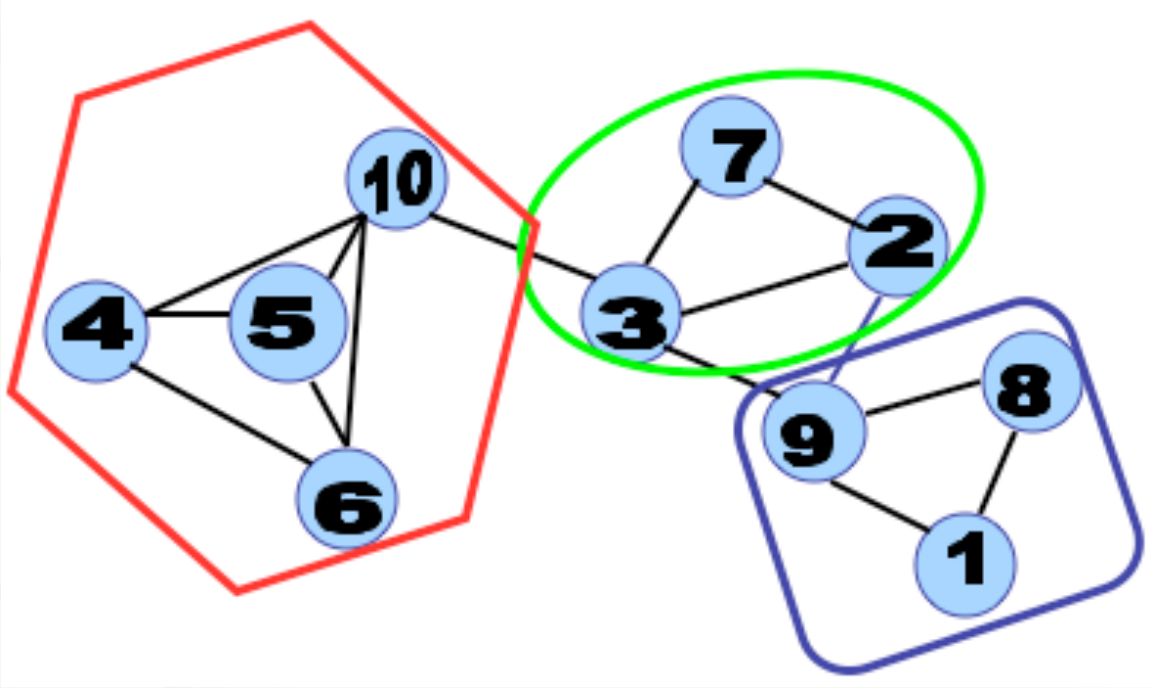
- Only works for square symmetric matrices
- Is an iterative algorithm
  - Mistakes are made worse by further iteration
  - Must compute an eigen decomposition at every iteration

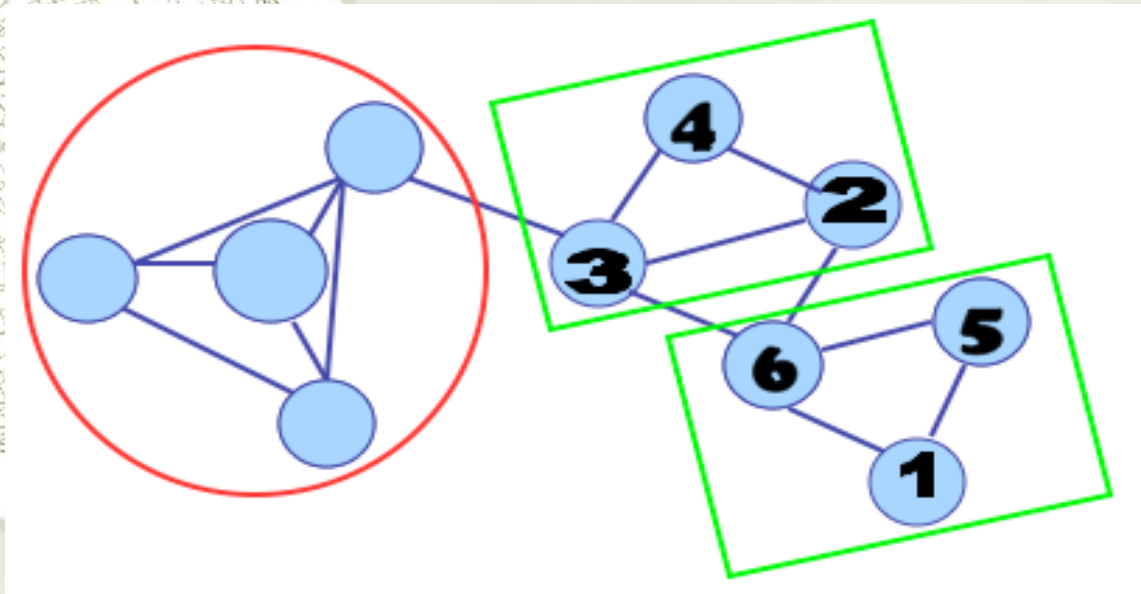
Extended Fiedler

- Using multiple eigenvectors is even more effective
  - No iteration required!
  - Only one eigen decomposition necessary!
- This is the method explored by Ibai Basabe in his thesis

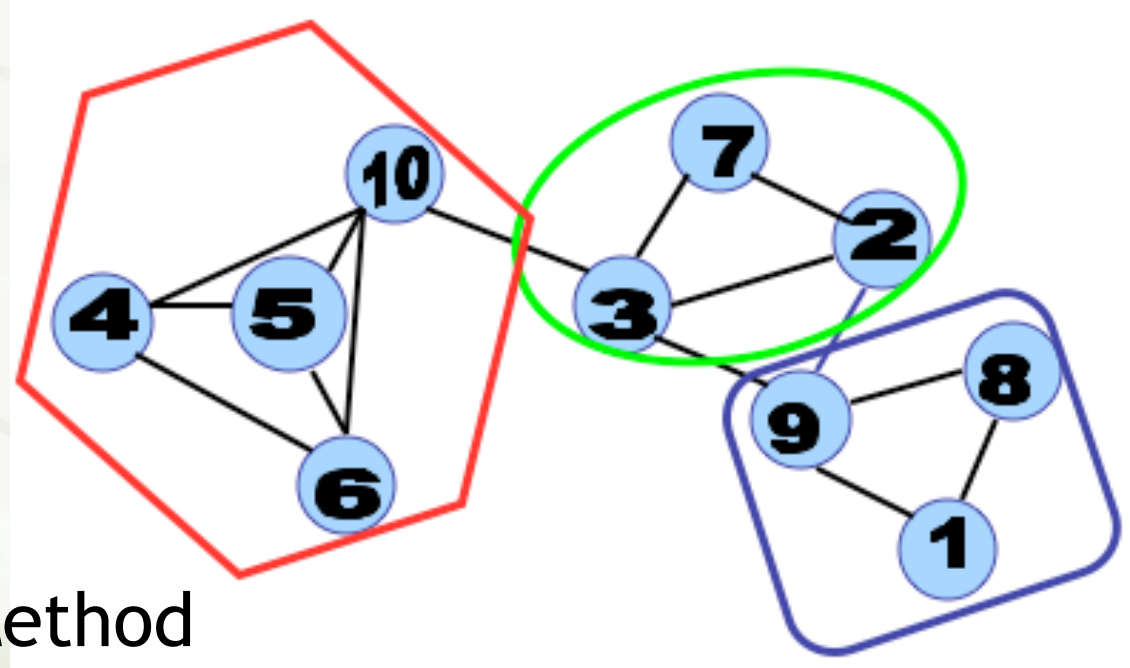
Applying Extended Fiedler







Fiedler Method



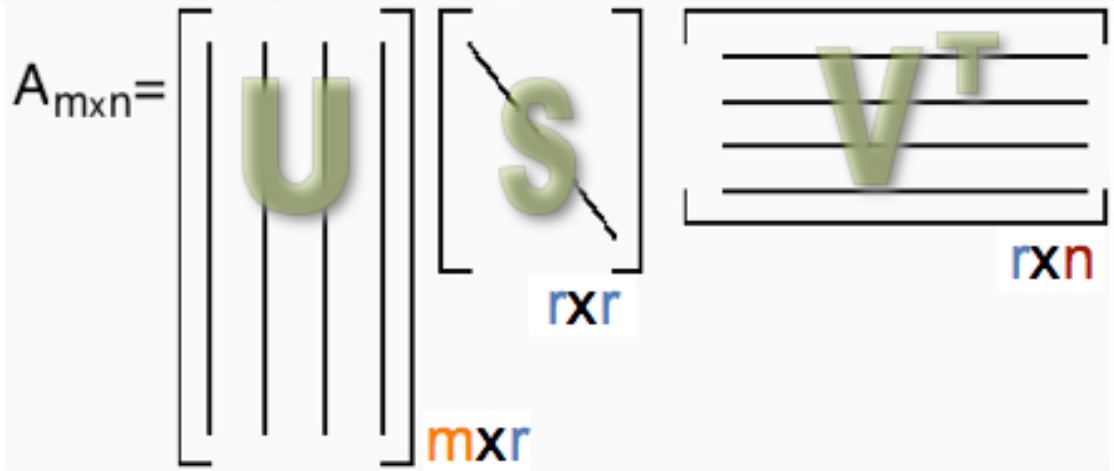
Extended Fiedler Method

Problems with Extended Fiedler

- Limited Flexibility
  - Cannot be applied to non-symmetric or rectangular matrices
- Dependent on Eigen Decomposition

The Singular Value Decomposition

- For *every* m by n matrix *A* there exist three matrices *U*, *S*, and *V* such that



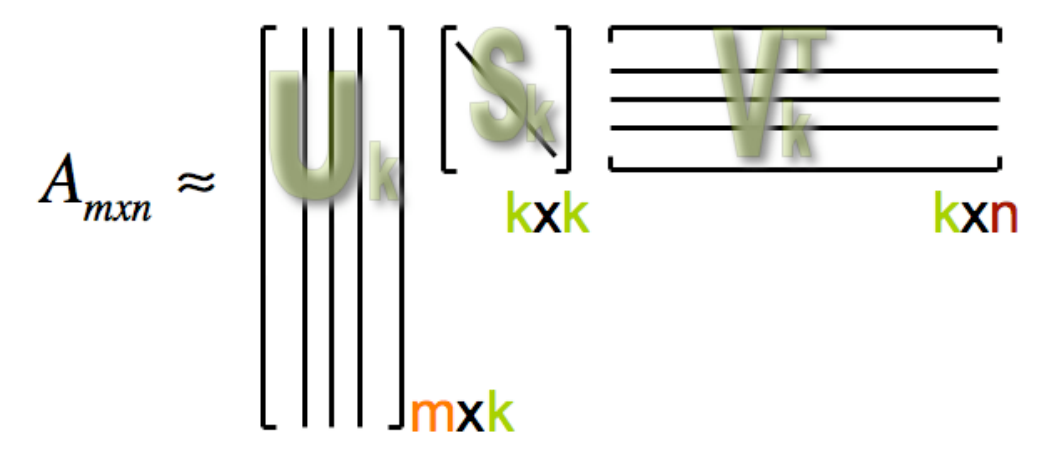
- *U* and V^T contain left and right singular vectors
- *S contains singular values*
  - *Real numbers*
  - *Appear in descending order*

Properties of the SVD

- An orthogonal decomposition
- An exact decomposition
- Closely related to eigen decomposition
  - If A is a square symmetric matrix
    - The singular vectors of A are the eigenvectors of A
    - The singular values of A are the squares of the eigenvalues of A
- Can be truncated

The Truncated SVD

- Can use SVD to find a rank *k* approximation of *A*

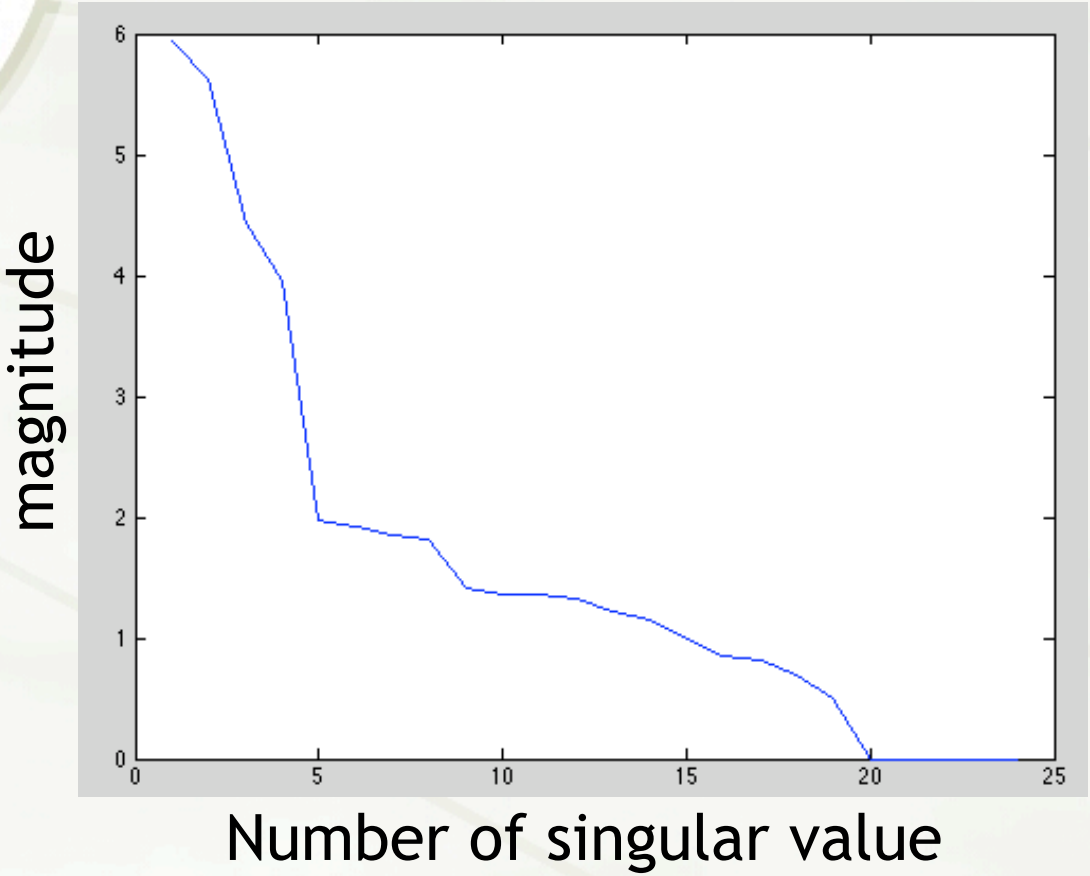


- Proven to be the closest rank *k* approximation by Eckert and Young
- When *k*<<*r*, truncated SVD is less expensive to find

How to choose k

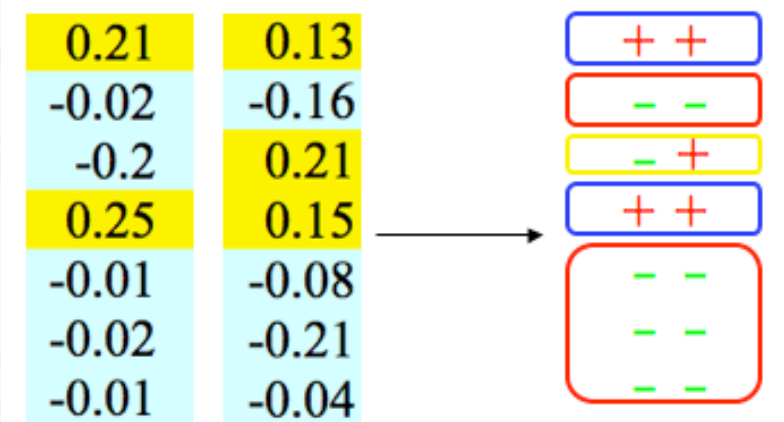
- This is a problem for *Extended Fiedler* as well, since it uses a truncated Eigen Decomposition
- Not a trivial problem
  - Many papers written on just this subject
  - Hundreds of possible solutions

Simple way to choose k



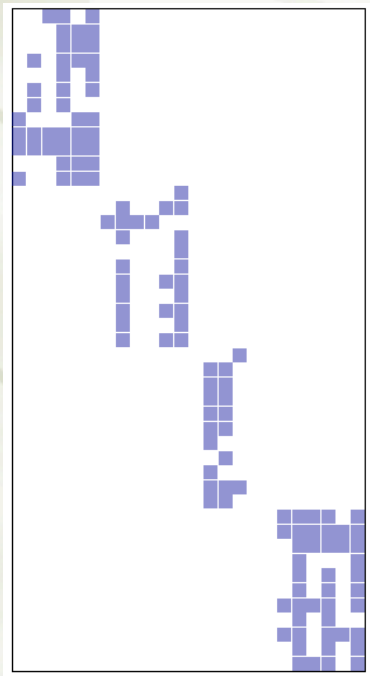
The SVD Signs Method

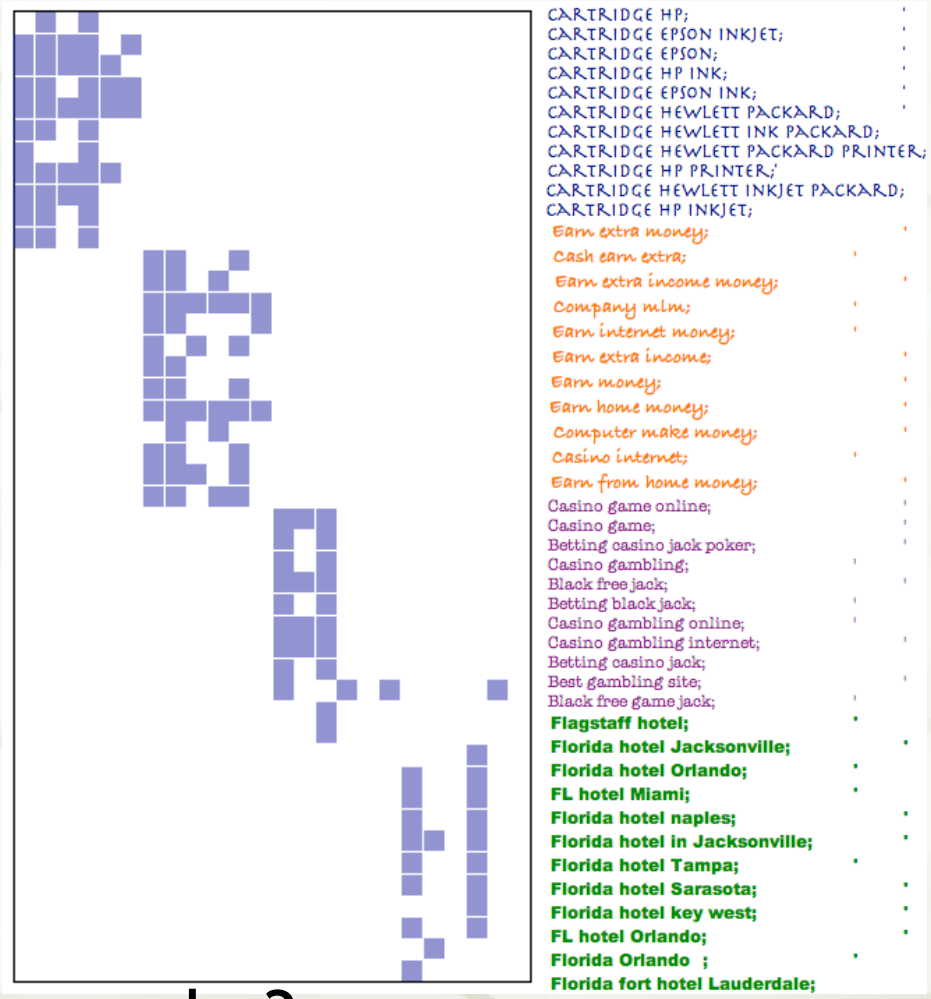
- Use sign patterns of singular vectors to cluster, like *Extended Fiedler*



- Use left singular vectors, U, for rows
- Use right singular vectors, V^T, for columns

Results on Small Yahoo!



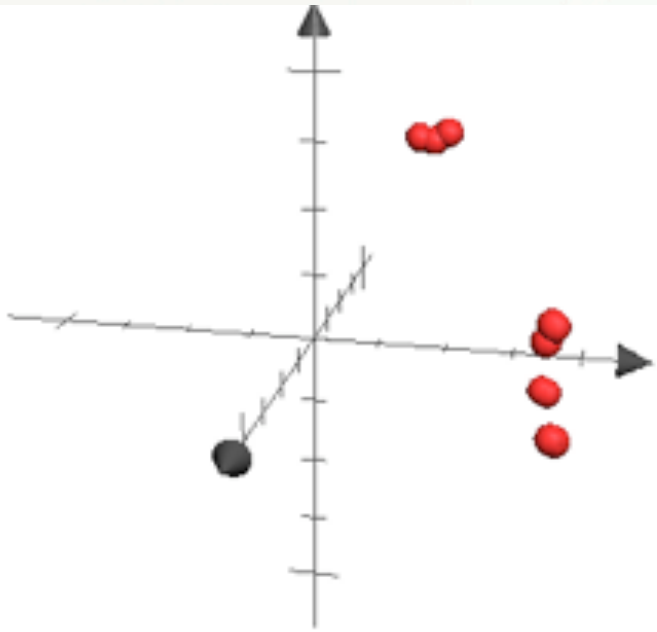


k=3

Why SVD Signs Works

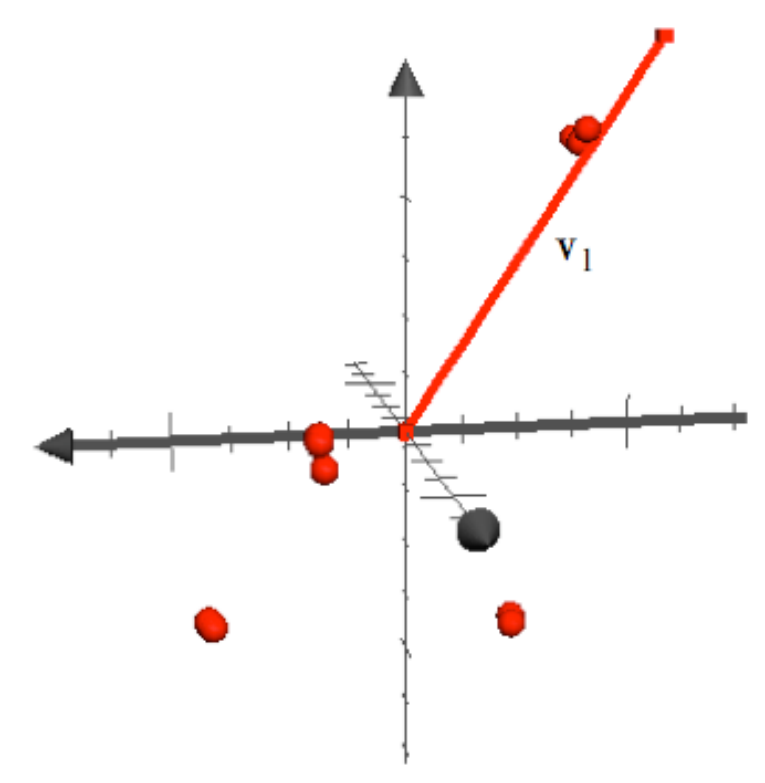
- Note that any m by n matrix can be seen as a set of m points in n dimensional space

0.9218	0.9355	0.0579
0.9257	0.9301	0.0611
0.9187	0.9399	0.0545
0.7382	0.9169	0.3529
0.7367	0.9098	0.3597
0.7275	0.8936	0.3006
0.1763	0.4103	0.8132
0.1764	0.5087	0.8326
0.1758	0.4607	0.8046
0.4057	0.8936	0.0099
0.4026	0.8763	0.0167

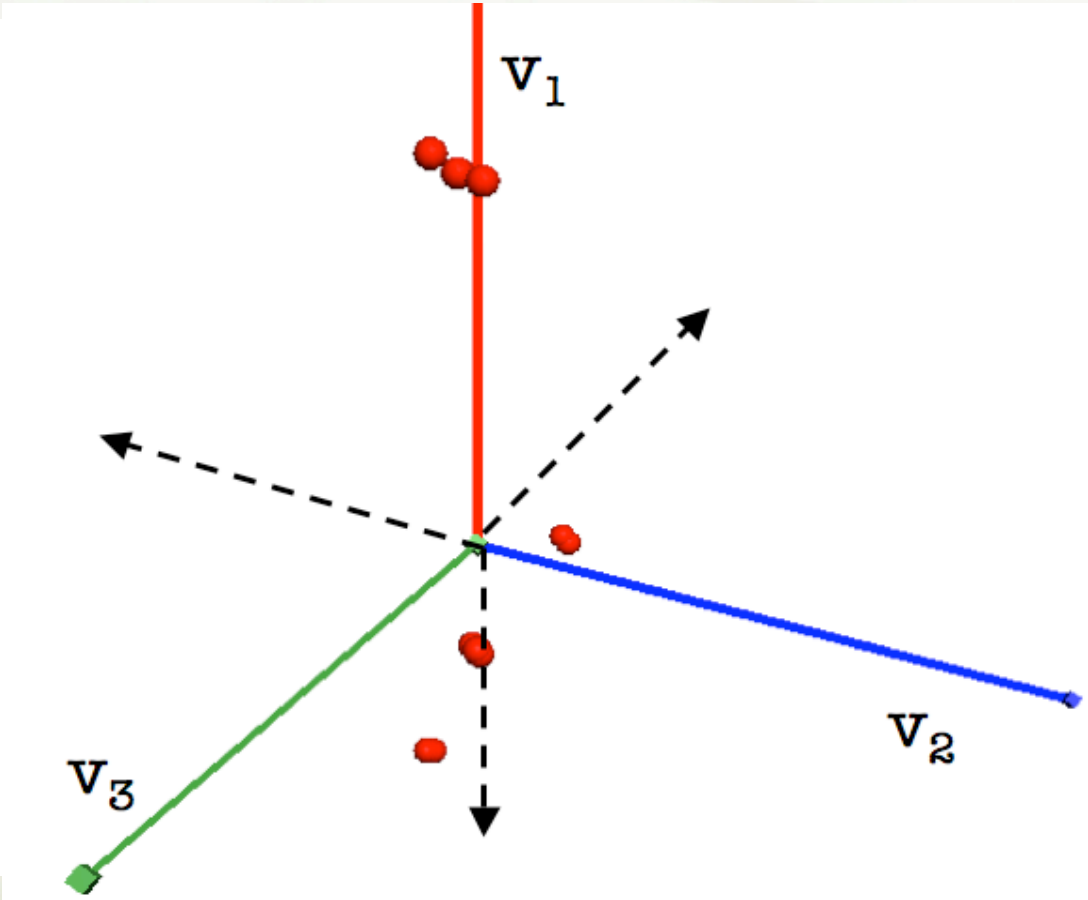


The Geometry of SVD

- The first right singular vector points in direction of principal trend

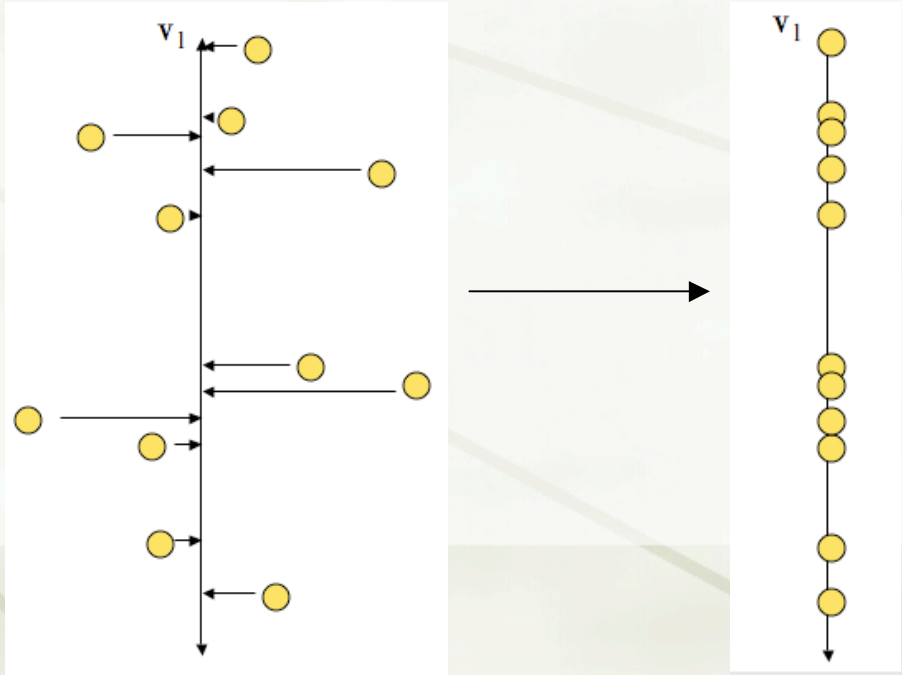


- Second and third right singular vectors point in direction of secondary and tertiary trend

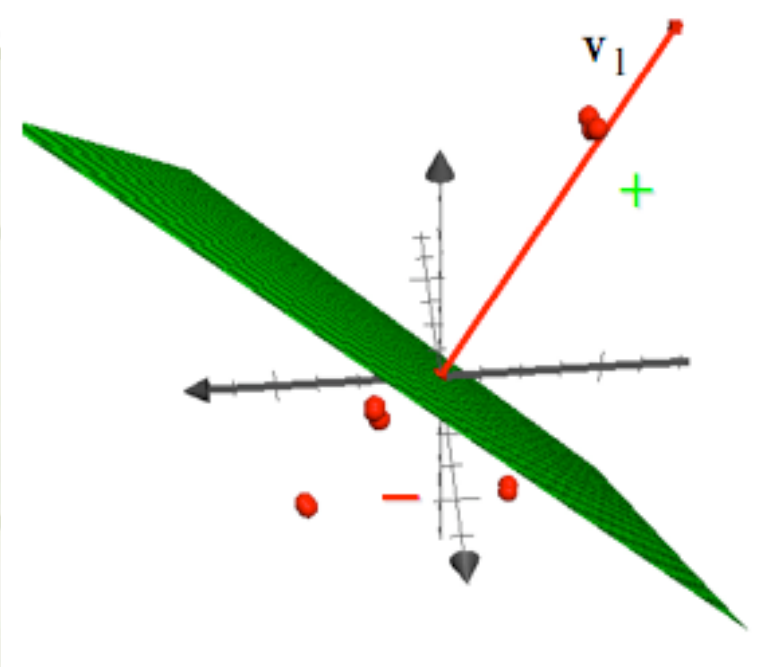


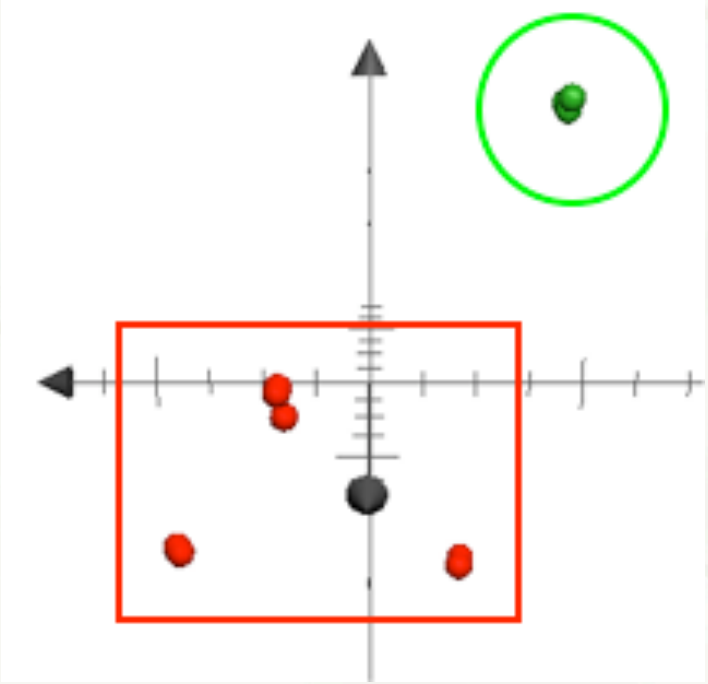
The Left singular vectors

- Contain orthogonal projections of points onto new axes
  - i.e. they give the coordinates of the points w.r.t. the new axes

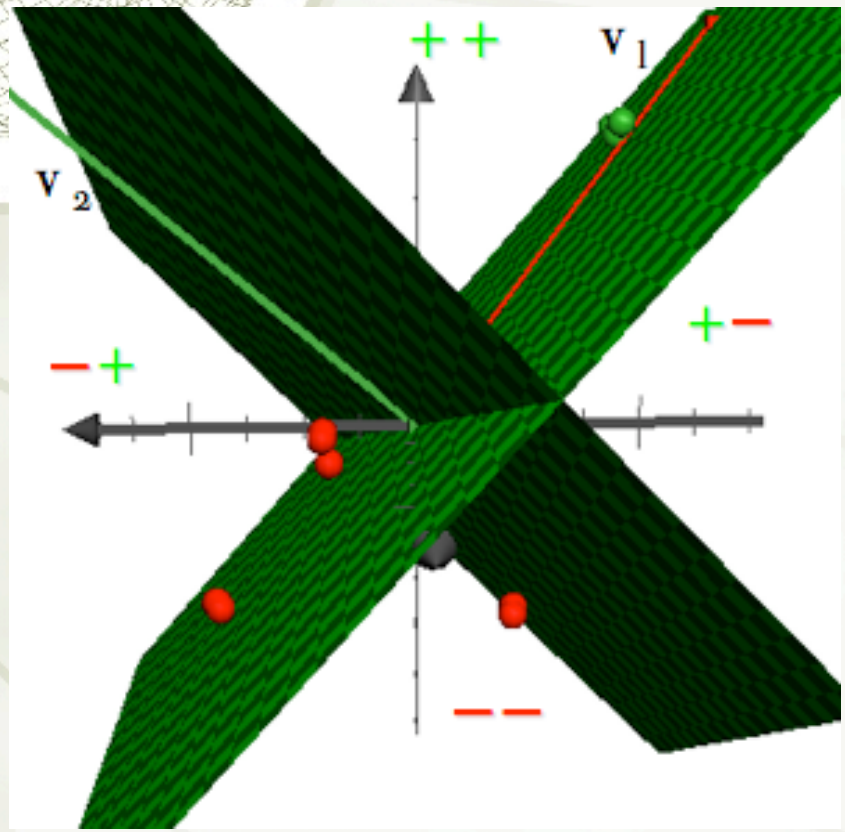


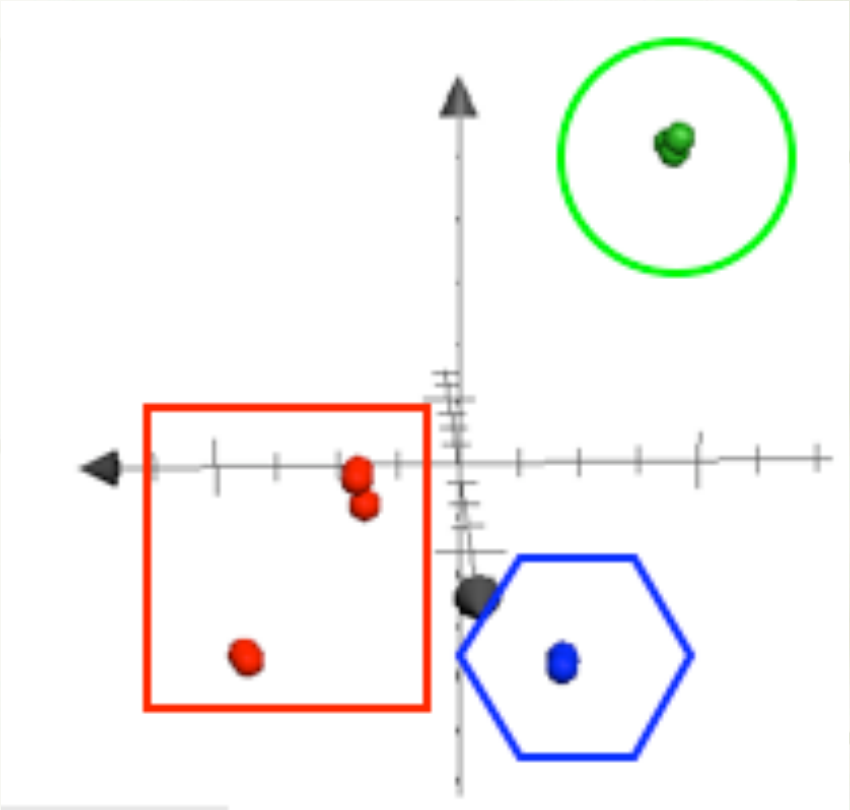
Clustering by signs of singular vectors



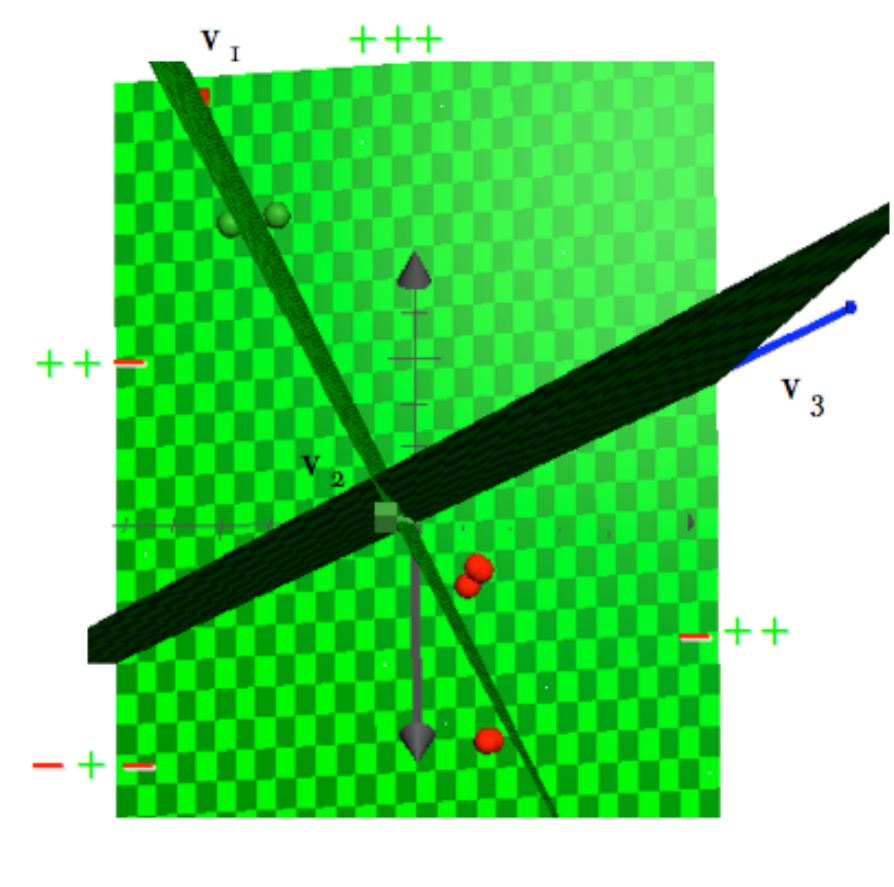


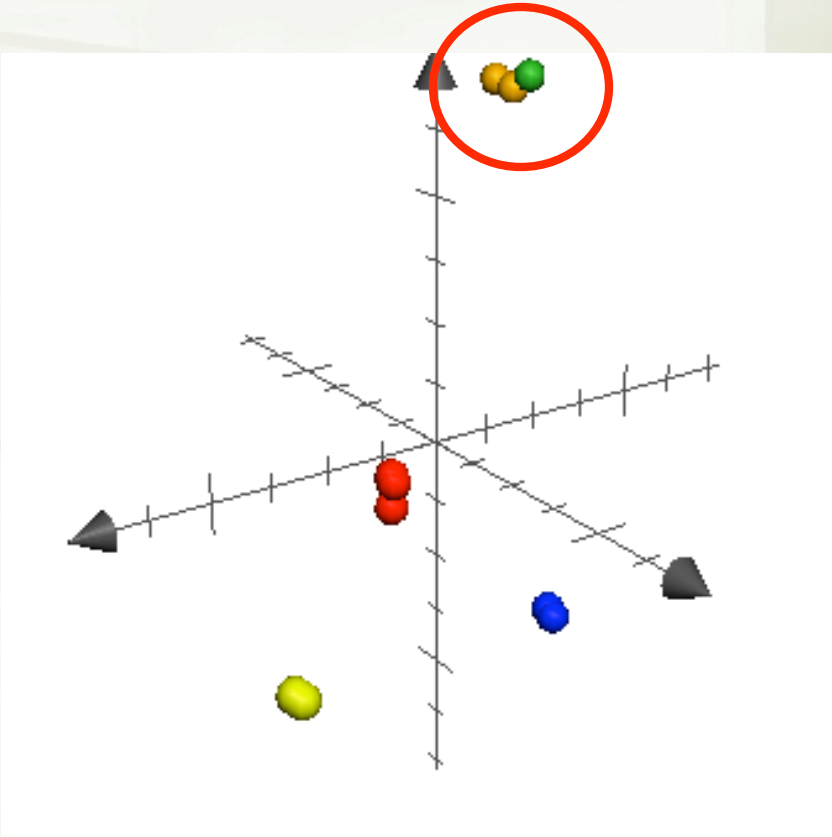
Clustering by signs of singular vectors



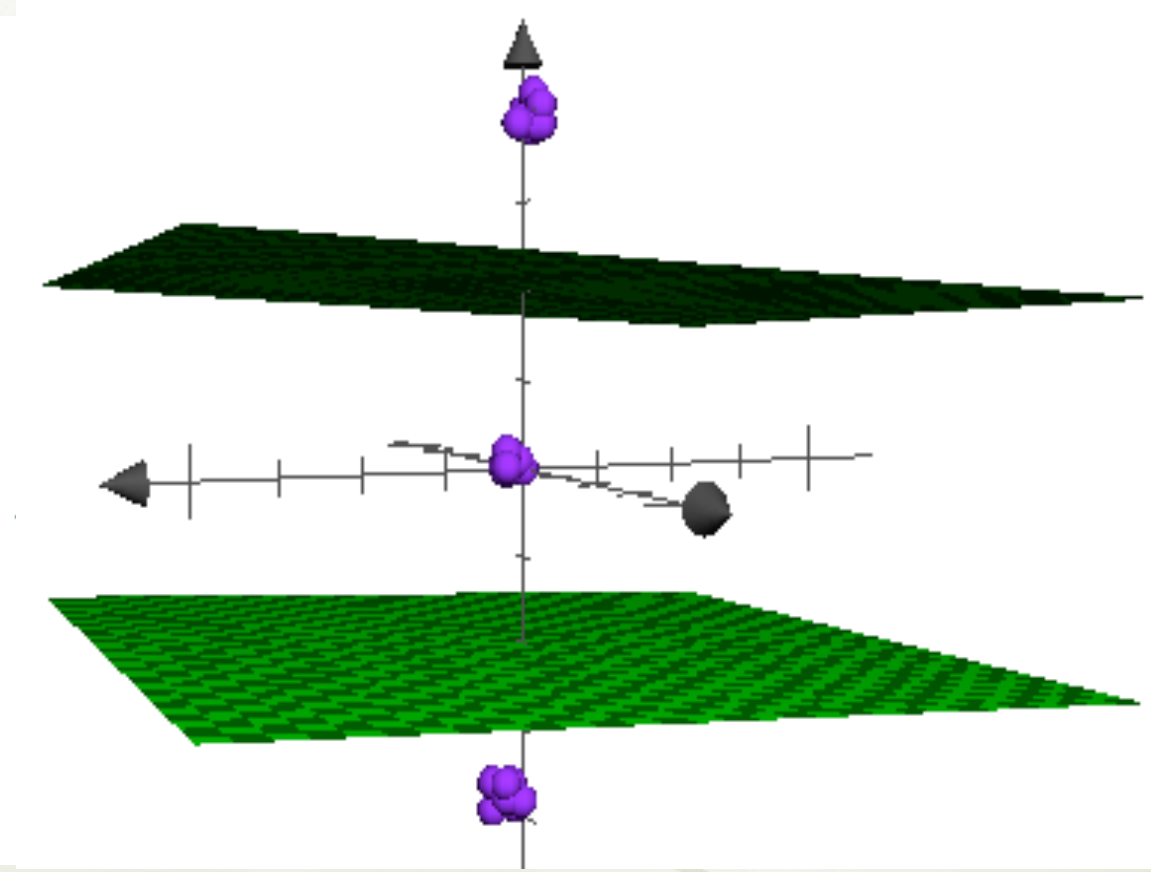


Clustering by signs of singular vectors



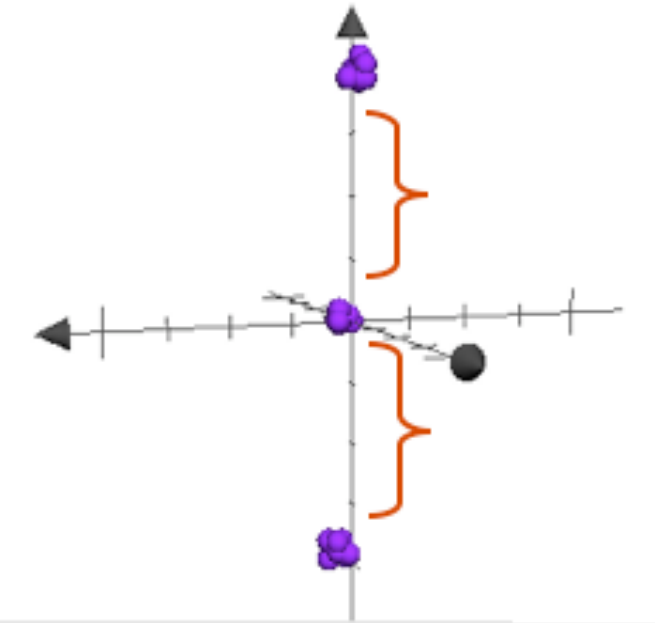


Problem with SVD Signs



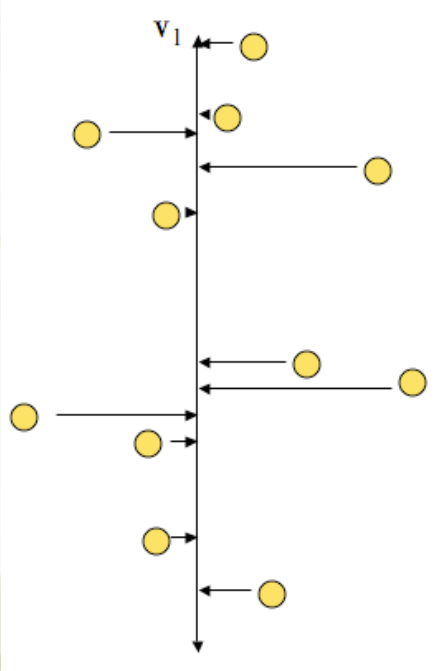
Proposed solution:

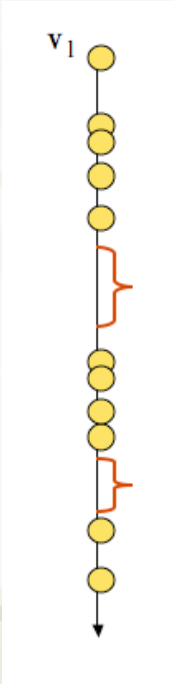
- Should find a way to locate gaps in the dataset
- Use the gaps as dividing points

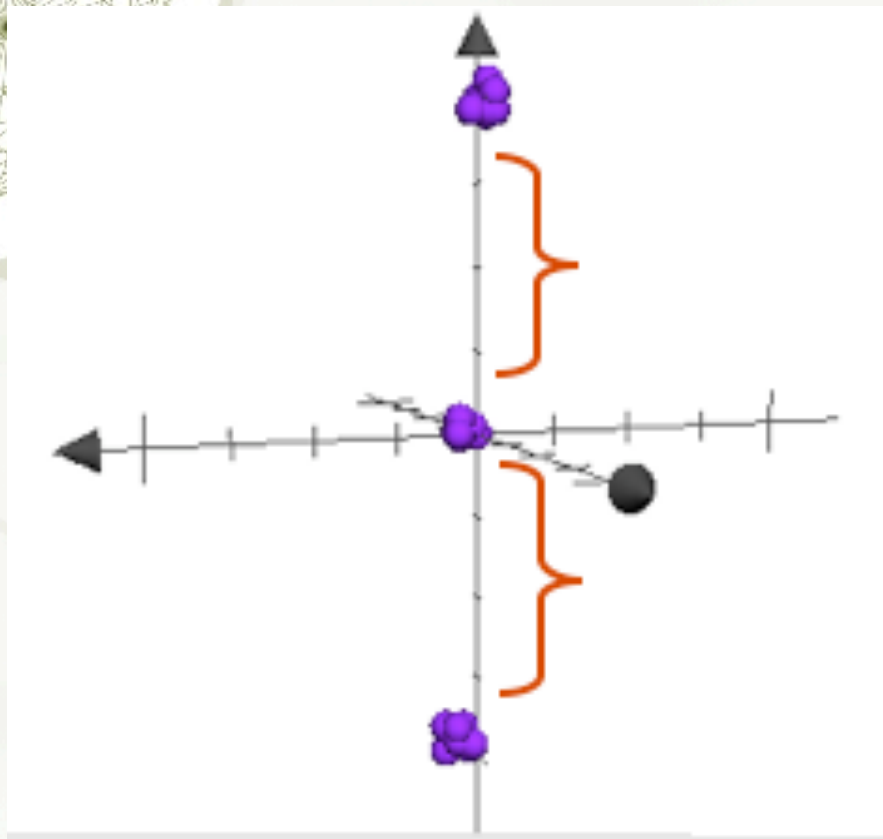


How do we find the gaps?

- Recall that the left singular vectors contain projections of points onto singular vectors







v_1	Gaps in v_1
-0.2537	
-0.252	0.0017
-0.244	0.008
-0.2419	0.002
-0.2347	0.0072
-0.2321	0.0025
-0.2317	0.0003
-0.2305	0.0012
-0.001	0.2294
0	0.001
0.0012	0.0011
0.0017	0.0005
0.0023	0.0005
0.0066	0.0043
0.0093	0.0026
0.013	3.60E-03
0.2519	0.2388
0.252	0.0001
0.2529	0.0008
0.2532	0.0002
0.2581	0.0049
0.2642	0.0061
0.2662	0.0019
0.2753	0.009

When are gaps large enough?

- Should be greater than average gap for each singular vector
- How much greater?
  - Use standard deviations
  - Convert each gap to a “z-score”
- What should be the cut-off “z-score”?
  - If too low then too many cuts
  - If too high then not enough cuts

Recommendation Systems

grocery store data from SAS

purchase history matrix

$$\mathbf{A} = \begin{array}{c} \\ \text{Item 1} \\ \text{Item 2} \\ \vdots \\ \text{Item m} \end{array} \begin{array}{c} \begin{array}{cccc} \text{User 1} & \text{User 2} & \dots & \text{User n} \end{array} \\ \begin{pmatrix} 1 & 5 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 2 \end{pmatrix} \end{array}$$

- Find similar users
- Find similar items
- Cluster users into classes

  EX: use NMF, SDD, SVD Signs, SVD Gaps

Netflix

17770 movies, $\approx$ .5 million users

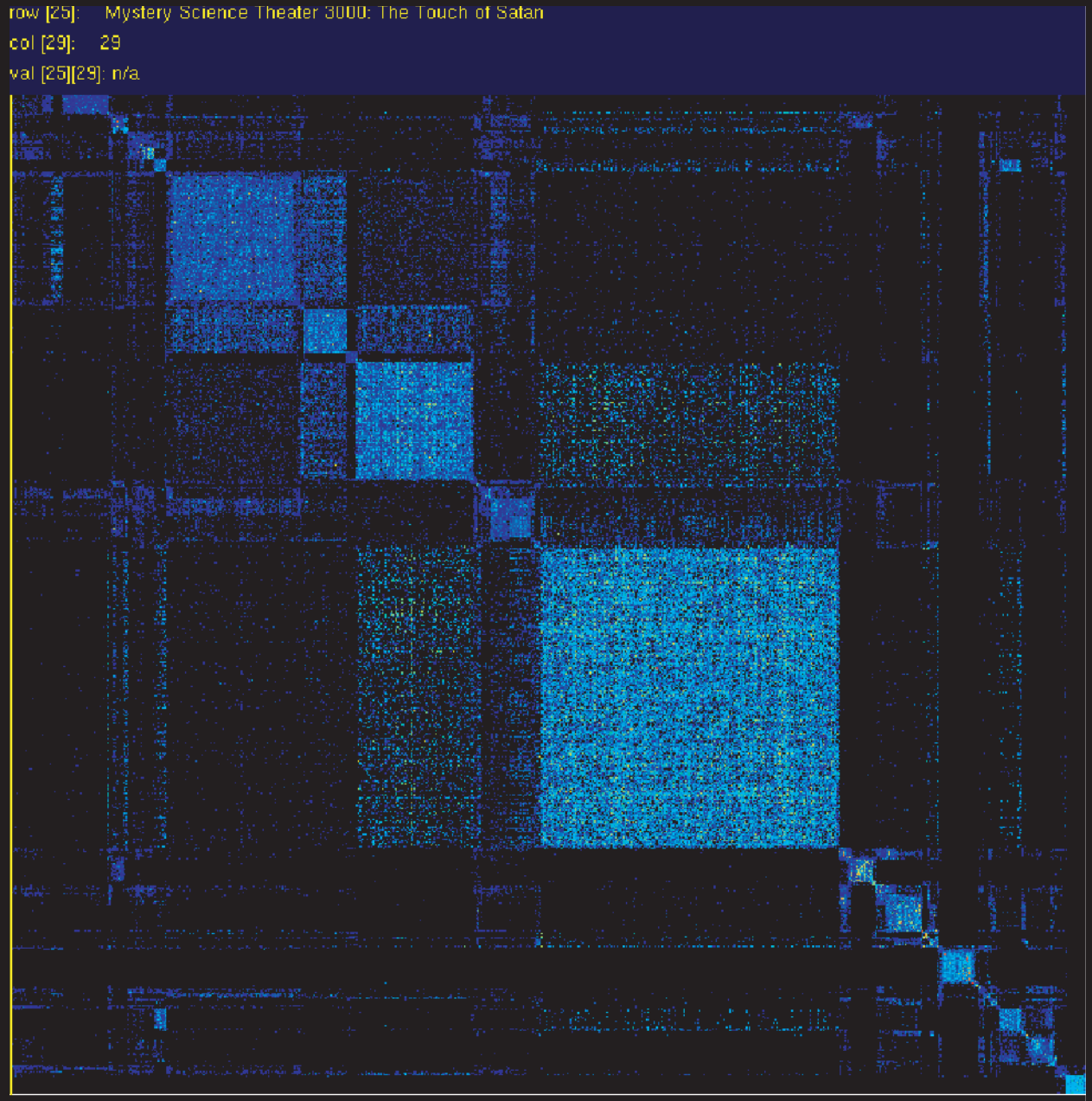
user rating matrix

$$\mathbf{A} = \begin{array}{c} \\ \text{movie 1} \\ \text{movie 2} \\ \vdots \\ \text{movie m} \end{array} \begin{array}{c} \begin{array}{cccc} \text{User 1} & \text{User 2} & \dots & \text{User n} \end{array} \\ \begin{pmatrix} 1 & 5 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 2 \end{pmatrix} \end{array}$$

- Find similar users
- Find similar movies
- Cluster users and movies into classes
- Rank users
- Rank movies

Clustering Netflix

(movie-movie matrix; 17770 x 17770)



(vismatrix tool - David Gleich)

Netflix

17770 movies, $\approx$.5 million users

user rating matrix

$$\mathbf{A} = \begin{array}{c} \\ \text{movie 1} \\ \text{movie 2} \\ \vdots \\ \text{movie m} \end{array} \begin{array}{c} \begin{array}{cccc} \text{User 1} & \text{User 2} & \dots & \text{User n} \end{array} \\ \begin{pmatrix} 1 & 5 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 2 \end{pmatrix} \end{array}$$

- Find similar users
- Find similar movies
- Cluster users and movies into classes
- Rank users
- Rank movies highest ranked, most polarizing, etc.

mHITS on Netflix

each movie i gets a rating $\mathbf{m}_i$ and each user gets a rating $\mathbf{u}_j$

- mHITS Thesis: A movie is a good (i.e., deserves a high rating m_i) if it gets high ratings from good (i.e., discriminating) users. A user is good (i.e., deserves a high rating u_j) when his or her ratings match the true rating of a movie.

mHITS Netflix Algorithm

$\mathbf{u} = \mathbf{e}$;

for i = 1 : maxiter

$\mathbf{m} = \mathbf{A}\,\mathbf{u}$;

$\mathbf{m} = \frac{5(\mathbf{m} - min(\mathbf{m}))}{max(\mathbf{m}) - min(\mathbf{m})}$;

$\mathbf{u} = \frac{1}{((\mathbf{R} - (\mathbf{R} > 0).*(\mathbf{e}\mathbf{m}^T)).^2)\mathbf{e}}$;

end

Netflix mHITS results

subset: 1500 "super users" (rate $\geq$ 1000 movies)

1st	Raiders of the Lost Ark
2nd	Silence of the Lambs
3rd	The Sixth Sense
4th	Shawshank Redemption
5th	LOR: Fellowship of the Ring
6th	The Matrix
7th	LOR: The Two Towers
8th	Pulp Fiction
9th	LOR: The Return of the King
10th	Forrest Gump
11th	The Usual Suspects
12th	American Beauty
13th	Pirates of the Carribbean: Black Pearl
14th	The Godfather
15th	Braveheart

Why Do Math?

from SIAM Marty Golubitsky

Random Surfer Video Project

Team

- Ryan Dumville, filmmaker
- Tanya Chartier, director
- Tim Chartier, mimematician ("Randy, the random surfer")
- Emmie Douglas, main actress
- Kathryn Pedings, props ("Page")
- me, props ("Brin")

will be used by SKV, LLC for court case

Conclusions

- Mathematical models can be found everywhere.
- Tackling problems together with students is mutually beneficial.

  $\Rightarrow$ class projects

  $\Rightarrow$ independent studies

  $\Rightarrow$ theses
- Use your resources: collaborate in teams with researchers around the world and in various fields.

Conclusions

- Mathematical models can be found everywhere.
- Tackling problems together with students is mutually beneficial.

  $\Rightarrow$ class projects

  $\Rightarrow$ independent studies

  $\Rightarrow$ theses
- Use your resources: collaborate in teams with researchers around the world and in various fields.

$\Rightarrow$ I'm in my educational utopia.