Mathematics Everywhere:
from Google to Sudoku to NCAA to Netflix

Amy Langville
Mathematics Department
College of Charleston
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
- 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 College
- 2005, Postdoc Applied Mathematics, N. C. State University
  - information retrieval, text mining, web search
  - Carl Meyer
  - *Google’s PageRank and Beyond: The Science of Search Engine Rankings*
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, web page ranking
  — 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
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
Tiny Web

$$H = \begin{pmatrix} P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\ P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \end{pmatrix}$$
$H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0
\end{pmatrix}$
Tiny Web

$$H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$
Tiny Web

\[ H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{3} & \frac{1}{3} & 0 & 0 & 1/3 & 0
\end{pmatrix} \]
Tiny Web

\[
\begin{align*}
H &= \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
P_5 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
P_6 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
\end{pmatrix}
\end{align*}
\]
Tiny Web

\[ H = \begin{pmatrix}
    P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
    P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
    P_2 & 0 & 0 & 0 & 0 & 0 & 0 \\
    P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
    P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
    P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
    P_6 & & & & & & 
\end{pmatrix} \]
$$H = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0 
\end{pmatrix}$$
A random walk on the Web Graph
A random walk on the Web Graph

PageRank = $\pi_i = \text{amount of time spent at } P_i$
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 \textit{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

![Diagram](image-url)
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
A random walk on the Web Graph

PageRank = $\pi_i = \text{amount of time spent at } P_i$

Dead end page (nothing to click on) — a “dangling node”
A random walk on the Web Graph

PageRank = $\pi_i$ = amount of time spent at $P_i$

Dead end page (nothing to click on) — a “dangling node”

$\pi^T = (0, 1, 0, 0, 0, 0) = \text{e-vector} \quad \Rightarrow \quad \text{Page } P_2 \text{ 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 $\frac{e^T}{n} = \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right)$

$$s = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
P_1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\
P_2 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
P_3 & 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
P_4 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\
P_5 & 0 & 0 & 0 & 1/2 & 0 & 1/2 \\
P_6 & 0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}$$
The Fix

Allow Web Surfers To Make Random Jumps

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

\[
S = \begin{pmatrix}
P_1 & P_2 & P_3 & P_4 & P_5 & P_6 \\
0 & 1/2 & 1/2 & 0 & 0 & 0 \\
1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\
1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 1/2 \\
0 & 0 & 0 & 1/2 & 0 & 1/2 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]

— \( S = H + \frac{ae^T}{6} \) is now row stochastic \( \implies \rho(S) = 1 \)
The Fix

Allow Web Surfers To Make Random Jumps

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

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

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

— Perron says $\exists \pi^T \geq 0$ s.t. $\pi^T = \pi^T S$ 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.
emerging 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 puzzle
Sudoku

Sudoku puzzle

Sudoku matrix

\[
\begin{array}{cccc}
7 & 8 & 5 & 2 \\
8 & 6 & 4 & 5 \\
1 & 9 & 8 & \\
4 & 2 & 8 & 9 \\
5 & 7 & 6 & 1 \\
7 & 3 & 6 & \\
3 & 1 & 6 & 4 \\
2 & 5 & 8 & 1 \\
\end{array}
\]

\[
\begin{array}{cccccccc}
6 & 4 & 7 & 8 & 1 & 5 & 2 & 3 & 9 \\
8 & 9 & 3 & 6 & 2 & 4 & 7 & 1 & 5 \\
2 & 1 & 5 & 3 & 9 & 7 & 4 & 8 & 6 \\
4 & 3 & 1 & 2 & 8 & 9 & 6 & 5 & 7 \\
7 & 2 & 6 & 4 & 5 & 3 & 8 & 9 & 1 \\
5 & 8 & 9 & 7 & 6 & 1 & 3 & 4 & 2 \\
1 & 7 & 4 & 9 & 3 & 2 & 5 & 6 & 8 \\
3 & 5 & 8 & 1 & 7 & 6 & 9 & 2 & 4 \\
9 & 6 & 2 & 5 & 4 & 8 & 1 & 7 & 3 \\
\end{array}
\]
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:

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

\[
\begin{align*}
\min & \quad 0^T x \\
\text{s.t.} & \quad \sum_{i=1}^{n} x_{ijk} = 1, \quad j=1:n, \quad k=1:n \quad \text{(only one } k \text{ in each column)} \\
& \quad \sum_{j=1}^{n} x_{ijk} = 1, \quad i=1:n, \quad k=1:n \quad \text{(only one } k \text{ in each row)} \\
& \quad \sum_{j=mq-mq}^{mq} \sum_{i=mp-mp}^{mp} x_{ijk} = 1, \quad k=1:n, \quad p=1:m, \quad q=1:m \quad \text{(only one } k \text{ in each submatrix)} \\
& \quad \sum_{k=1}^{n} x_{ijk} = 1 \quad i=1:n, \quad j=1:n \quad \text{(every position in matrix must be filled)} \\
& \quad x_{ijk} = 1 \quad \forall (i,j,k) \in G \quad \text{(given elements } G \text{ in matrix are set “on”)} \\
& \quad x_{ijk} \in \{0,1\} 
\end{align*}
\]
Mathematical Model of Sudoku

Define:

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

\[
\begin{align*}
\min & \quad 0^T x \\
\text{s.t.} & \quad \sum_{i=1}^{n} x_{ijk} = 1, \quad j=1:n, k=1:n \quad (\text{only one } k \text{ in each column}) \\
& \quad \sum_{j=1}^{n} x_{ijk} = 1, \quad i=1:n, k=1:n \quad (\text{only one } k \text{ in each row}) \\
& \quad \sum_{j=mp+1}^{mq} \sum_{i=mp+1}^{mp} x_{ijk} = 1, \quad k=1:n, p=1:m, q=1:m \quad (\text{only one } k \text{ in each submatrix}) \\
& \quad \sum_{k=1}^{n} x_{ijk} = 1 \quad i=1:n, j=1:n \quad (\text{every position in matrix must be filled}) \\
& \quad x_{ijk} = 1 \quad \forall (i, j, k) \in G \quad (\text{given elements } G \text{ in matrix are set “on”}) \\
& \quad x_{ijk} \in \{0, 1\} 
\end{align*}
\]

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 \times 9 puzzle takes 16.7 seconds to solve on desktop machine.
Most puzzle creators do not check whether their puzzle has one unique solution.
Most puzzle creators do not check whether their puzzle has one unique solution.

<table>
<thead>
<tr>
<th>Puzzle</th>
<th>Solution 1</th>
<th>Solution 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1 2</td>
<td>4 2</td>
</tr>
<tr>
<td>2</td>
<td>4 3</td>
<td>2 1</td>
</tr>
<tr>
<td>3</td>
<td>3 4</td>
<td>3 1</td>
</tr>
<tr>
<td>4</td>
<td>2 1</td>
<td>2 4</td>
</tr>
<tr>
<td></td>
<td>4 3</td>
<td>1 2</td>
</tr>
</tbody>
</table>
Some Interesting $9 \times 9$ Sudoku Facts

- How many $9 \times 9$ 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 $9 \times 9$ 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 $\leq 16$ givens has been found yet.

- Given one Sudoku matrix, could I make my own Daily Sudoku Calendar?
  
  By using mathematical operations $362,879 \approx 991$ years worth of Sudoku matrices can be created from one $9 \times 9$ 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 $o_i$ and defensive rating $d_i$

- **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](image)

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

Point Matrix $P \geq 0$
mHITS Equations

Summation Notation

\[ d_j = \sum_{i \in I_j} p_{ij} \frac{1}{o_i} \quad \text{and} \quad o_i = \sum_{j \in L_i} p_{ij} \frac{1}{d_j} \]

Matrix Notation: iterative procedure

\[ d^{(k)} = P^T \frac{1}{o^{(k)}} \quad \text{and} \quad o^{(k)} = P \frac{1}{d^{(k-1)}} \]

- related to the Sinkhorn-Knopp algorithm for matrix balancing
  (uses successive row and column scaling to transform \( P \geq 0 \) into doubly stochastic matrix \( S \))
- P. Knight uses Sinkhorn-Knopp algorithm to rank webpages
mHITS Results: tiny NCAA
(data from Luke Ingram)

<table>
<thead>
<tr>
<th>Team</th>
<th>Off. Rating $o$</th>
<th>Off. Rank</th>
<th>Def. Rating $d$</th>
<th>Def. Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duke</td>
<td>0.4</td>
<td>5th</td>
<td>140</td>
<td>5th</td>
</tr>
<tr>
<td>Miami</td>
<td>1.8</td>
<td>1st</td>
<td>67</td>
<td>2nd</td>
</tr>
<tr>
<td>UNC</td>
<td>0.6</td>
<td>4th</td>
<td>97</td>
<td>4th</td>
</tr>
<tr>
<td>UVA</td>
<td>1.0</td>
<td>3rd</td>
<td>80</td>
<td>3rd</td>
</tr>
<tr>
<td>VT</td>
<td>1.4</td>
<td>2nd</td>
<td>34</td>
<td>1st</td>
</tr>
</tbody>
</table>

\[ r = \frac{o}{d} \text{ Rating} \quad r = \frac{o}{d} \text{ Rank} \]

<table>
<thead>
<tr>
<th>Team</th>
<th>$r$ = $\frac{o}{d}$ Rating</th>
<th>$r$ = $\frac{o}{d}$ Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duke</td>
<td>0.003</td>
<td>5th</td>
</tr>
<tr>
<td>Miami</td>
<td>0.027</td>
<td>2nd</td>
</tr>
<tr>
<td>UNC</td>
<td>0.006</td>
<td>4th</td>
</tr>
<tr>
<td>UVA</td>
<td>0.012</td>
<td>3rd</td>
</tr>
<tr>
<td>VT</td>
<td>0.041</td>
<td>1st</td>
</tr>
</tbody>
</table>
Weighed mHITS

- Weighted Point Matrix $\tilde{P} \geq 0$

$$\tilde{p}_{ij} = w_{ij} \cdot p_{ij} \quad (w_{ij} = \text{weight of matchup between teams } i \text{ and } j)$$

- possible weightings $w_{ij}$

![Graphs showing Linear, Logarithmic, Exponential, and Step Function weightings](image)
mHITS Results: full NCAA
(image from Neil Goodson and Colin Stephenson)
mHITS Results: full NCAA

- weightings can easily be applied to all ranking methods
  ⇒ interesting possibilities for weighted webpage ranking

<table>
<thead>
<tr>
<th>Method</th>
<th>ESPN score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Massey Linear</td>
<td>1450</td>
</tr>
<tr>
<td>Massey Log</td>
<td>1450</td>
</tr>
<tr>
<td>mHITS Log</td>
<td>1450</td>
</tr>
<tr>
<td>Massey Step</td>
<td>1420</td>
</tr>
<tr>
<td>Massey Exponential</td>
<td>1400</td>
</tr>
<tr>
<td>mHITS Step</td>
<td>1320</td>
</tr>
<tr>
<td>Massey Uniform</td>
<td>1310</td>
</tr>
<tr>
<td>mHITS Linear</td>
<td>1310</td>
</tr>
<tr>
<td>mHITS Uniform</td>
<td>1310</td>
</tr>
<tr>
<td>Colley Linear</td>
<td>1100</td>
</tr>
<tr>
<td>Colley Log</td>
<td>1010</td>
</tr>
</tbody>
</table>
mHITS Point Matrix $P \geq 0$

Perron-Frobenius guarantees (for irreducible $P$ with total support)

- existence of $o$ and $d$
- uniqueness of $o$ and $d$
- convergence of mHITS algorithm
- rate of convergence of mHITS algorithm

\[ \sigma_2^2(S), \text{ where } S = D(1/o) \cdot P \cdot D(1/d) \]
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

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.

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.

Results

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

Massey Ranking | Evolutionary Optimization | Colley Ranking
--- | --- | ---
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 | Elson | Elson
Elon | CofC | CofC
Wofford | W. Car. | Wofford
W. Car. | Furman | Furman
Furman | Citadel | W. Car.
Citadel | Citadel | Citadel

# of Viol = 269 | # of Viol = 265 | # of Viol = 278

Future Work

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

Work Cited


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!
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

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
Finding the Laplacian

\[ L = D - A \]

\[
\begin{bmatrix}
2 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & 0 \\
0 & 3 & -1 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\
0 & -1 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & -1 \\
0 & 0 & 0 & 3 & -1 & -1 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & 3 & -1 & 0 & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & -1 & 3 & 0 & 0 & 0 & -1 \\
0 & -1 & -1 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
-1 & 0 & 0 & 0 & 0 & 0 & 2 & -1 & 0 \\
-1 & -1 & -1 & 0 & 0 & 0 & 0 & -1 & 4 & 0 \\
0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 4 \\
\end{bmatrix}
\]  

\[
\begin{bmatrix}
2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\
\end{bmatrix}
\]  

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
What is a Fiedler Vector?

- Eigenvector corresponding to the second smallest eigenvalue of the Laplacian
- Named for Miroslav Fiedler
- Signs can be used to bisect the graph
Applying the Fiedler Method

First Iteration

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
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

  $A_{m \times n} = \begin{bmatrix} U & \vdots \end{bmatrix} \begin{bmatrix} S \end{bmatrix} \begin{bmatrix} V^T \end{bmatrix}$

- $U$ 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

\[
A_{mxn} \approx \begin{bmatrix} U_k^{mxk} & S_k^{kxk} & V_k^{kxn} \end{bmatrix}
\]
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
The first right singular vector points in the 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.
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

\[ A = \begin{pmatrix}
\text{User 1} & \text{User 2} & \ldots & \text{User n} \\
\text{Item 1} & 1 & 5 & \ldots & 0 \\
\text{Item 2} & 0 & 0 & \ldots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{Item m} & 0 & 1 & \ldots & 2 \\
\end{pmatrix} \]

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

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

17,770 movies, ≈ .5 million users

\[
A = \begin{pmatrix}
\text{User} & 1 & 2 & \ldots & n \\
\text{movie} 1 & 1 & 5 & \ldots & 0 \\
\text{movie} 2 & 0 & 0 & \ldots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\text{movie} m & 0 & 1 & \ldots & 2
\end{pmatrix}
\]

- Find similar users
- Find similar movies
- **Cluster users and movies into classes**
- Rank users
- Rank movies
Clustering Netflix
(movie-movie matrix; 17770 x 17770)
Netflix

17770 movies, \( \approx \) .5 million users

\[
A = \begin{pmatrix}
\text{User 1} & \text{User 2} & \ldots & \text{User n} \\
\text{movie 1} & 1 & 5 & \ldots & 0 \\
\text{movie 2} & 0 & 0 & \ldots & 1 \\
: & : & : & \ddots & : \\
\text{movie m} & 0 & 1 & \ldots & 2
\end{pmatrix}
\]

- 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 $m_i$ and each user gets a rating $u_j$

- 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

\[
\begin{align*}
u &= e; \\
\text{for } i &= 1 : \text{maxiter} \\
m &= A u; \\
m &= \frac{5(m - \min(m))}{\max(m) - \min(m)}; \\
u &= \frac{1}{((R - (R > 0) \times (em^T))^2) e}; \\
\end{align*}
\]
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.
  ⇒ class projects
  ⇒ independent studies
  ⇒ 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.
  - class projects
  - independent studies
  - theses

- Use your resources: collaborate in teams with researchers around the world and in various fields.
  - I’m in my educational utopia.