Thinking linearly about data

Tim Chartier
Department of Mathematics

tichartier@davidson.edu

Picture credit: http://cyberkshema.com/2012/10/10/big-data/
Binary bunch

Our technological world is filled with data.

Picture credit: http://skinbeautifulblog.files.wordpress.com/2008/04/weight-loss.jpg
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.

EVERY MINUTE OF THE DAY

571 NEW WEBSITES ARE CREATED.
2,083 FOURSQUARE USERS PERFORM CHECK-INS.
3,125 INSTAGRAM USERS ADD NEW PHOTOS.
3,600 TWITTER USERS SEND OVER 100,000 TWEETS.
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
TWITTER USERS SEND OVER 100,000 TWEETS.

Picture credit: http://www.visualnews.com/2012/06/19/how-much-data-created-every-minute/
Friending data

As of Sept. 2012, Facebook has over one billion active users, over half use mobile devices. (wikipedia.com)
Friending data

• Each minute $\geq 290,000$ status updates are posted and nearly 140,000 photos are uploaded.

• So, 15 min. of Facebook photos uploads $\geq$ number of photographs stored in the NY public photo archives. (digitaltrends.com)

Picture credit: http://digitalgallery.nypl.org/
Data

Organize

Picture credit: http://www.alegoaday.com/category/general-lego/page/3/
Organize 4 meaning

Life is linear

Linear algebra will be the tool we take from the data mining tool belt.

Picture credit: http://www.clker.com/cliparts/9/6/7/c/1280887403729185486tool%20belt.jpg
2 mine

We will discuss two ways to mine through data:

• rank items - who's the best?
• cluster items - who's like each other?

Chip off the tip

This is a huge field.

Focus and foreshadow

This talk will

• highlight my teaching and research both with undergraduates and

• foreshadow topics in tomorrow’s 8:30 AM invited session “Thinking Linearly about Data in Research and Teaching”.

Ideas?

• I’ll give you a sense of the breadth of work as it may give you or your students ideas.
• Details are available in expository and research articles.

Picture credit: http://www.inventionmachine.com/Portals/56687/images/idea_bulb.jpg
Who’s Number 1

It’s a question in pop culture, government, and sports.

WAIT!

Why?
Daily answers

We use rankings with every Google search.
Pages of rankings

We need more than who is ranked first.
Billion dollar math

- Google’s PageRank algorithm is an attractive classroom topic but relies on Markov Chains.
- What are other options?
New Year's math

- I teach ranking methods that help choose which teams played in the Rose Bowl.
- Recent research with Amy Langville allows students to create personalized methods.

Getting graphic

• Suppose we only know who won and who lost each game.
• We denote this with an arrow (directed edge) from the winning to the losing team.
Ready to Rank

How would you rank the following teams?

B  A  C  D
Get Real!

Would this change your mind?
Two conferences

Suppose we have two conferences.
Rank again...

How would you rank now?
and again...

How would you rank now?
Dependable math

Our rankings will have a sense of “quality” of a win or loss.
When life is linear

• We will use a linear model used by the Bowl Championship Series for college football.
• These models take “strength of schedule” into account.

Picture credit: http://www.bcsfootball.org
A real honor

The Massey method started as an honors math project by Ken Massey while he was undergraduate at Bluefield College.
Transitivity?

- Suppose Duke beat UNC-Chapel Hill by 10 points and UNC-Chapel Hill beat Wake Forest by 5 points.
- Could we predict Duke would beat Wake Forest by 15 points?
Crystal ball not included

Can it be true approximately?

Picture credit: http://www.photo-dictionary.com/phrase/5923/crystal-ball.html
Rated $r$

- Let $r_1$, $r_2$, and $r_3$ be the ratings for Duke, UNC-Chapel Hill and Wake Forest.
- The Massey method assumes:
  
  \[
  r_1 - r_2 = 10, \\
  r_2 - r_3 = 5, \\
  r_3 - r_1 = 1, 
  \]

  which produces a singular system.
Being normal

• If we rank Division I NCAA Men’s basketball teams, our matrix system is approximately:

\[ M_{5000 \times 350} r_{350 \times 1} = b_{5000 \times 1} \]

• The normal equations:

\[ M^T M r = M^T b, \]

have zero-row sum and are singular.

• So, we replace a row in \( M^T M \) with 1’s and the RHS with a 0.
Finite nuts and bolts

• For a class like Finite Math, I teach the nuts and bolts of such methods.
• I also supply software to create personalized rankings.

Picture credit: http://www.itracing.co.uk/suspension-nuts-and-bolts
Fictional NFL

Let’s rank this fictional series of games within the South Division of the NFL.

We’ll represent the records of the teams by a weighted graph.
Row, row, row

- There are 4 teams so we need a 4 x 4 matrix.
- Each row of the matrix corresponds to a team.
Naming your columns

The columns have the same ordering as the rows.
Corner to Corner

Each diagonal element equals the number of games played by the corresponding team.
Team to team

Each off-diagonal element equals \(-g\) where \(g\) is the number of times the corresponding teams played.
‘Tis the season

Let’s construct the matrix for our season.

\[
\begin{pmatrix}
2 & -1 & 0 & -1 \\
-1 & 3 & -1 & -1 \\
0 & -1 & 2 & -1 \\
-1 & -1 & -1 & 3
\end{pmatrix}
\]
Go to the right

• We need to form the vector for the right-hand side of the linear equation.

• Each row in that vector corresponds to the same team that it did in the matrix.
A Saint

- The RHS equals the sum of the point differentials where a win gives a positive differential and a loss, a negative one.
- The Saints won by 7 and lost by 3. So, the RHS associated with this team would be 7-3 = 4.
I’m right

For the season, the right-hand side vector is,

$$
\begin{pmatrix}
4 \\
7 \\
13 \\
-24
\end{pmatrix}
$$
Playing the system

The entire linear system is then:

\[
\begin{pmatrix}
2 & -1 & 0 & -1 \\
-1 & 3 & -1 & -1 \\
0 & -1 & 2 & -1 \\
-1 & -1 & -1 & 3
\end{pmatrix}
\begin{pmatrix}
S \\
B \\
P \\
F
\end{pmatrix}
=
\begin{pmatrix}
4 \\
7 \\
13 \\
-24
\end{pmatrix}
\]

**Note:** There is one more step.
Playing the system

Replace any row by 1’s in the matrix and 0 on the RHS.

\[
\begin{pmatrix}
2 & -1 & 0 & -1 \\
-1 & 3 & -1 & -1 \\
0 & -1 & 2 & -1 \\
1 & 1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
S \\
B \\
P \\
F
\end{pmatrix}
= 
\begin{pmatrix}
4 \\
7 \\
13 \\
0
\end{pmatrix}
\]
Who’s #1?

For this season, the teams are ranked (in decreasing order of their ratings) Panthers, Buccaneers, Saints and Falcons.
Prediction

If the Panthers played the Saints, this method predicts the Panthers will win by 4.375 + 0.125 = 4.5 points.
Bigger and bigger

- Note such a method can reward blowouts.
- It can be adapted to look only at wins.
- Another option is the Colley method.
Get a Colley

• The Colley matrix

\[ C = 2I + M^TM. \]

• The entries in the RHS vector equal

\[ 1 + \frac{1}{2}(W - L) \]

for each corresponding team.

*Picture credit: http://cooperativeintelligence.files.wordpress.com/2011/12/win-loss-image.jpg*
Colley’s return

Let’s construct the matrix for our season.
Colley’s return

The linear system for the Colley method is:

\[
\begin{pmatrix}
4 & -1 & 0 & -1 \\
-1 & 5 & -1 & -1 \\
0 & -1 & 4 & -1 \\
-1 & -1 & -1 & 5 \\
\end{pmatrix}
\begin{pmatrix}
S \\
B \\
P \\
F \\
\end{pmatrix}
=
\begin{pmatrix}
1 \\
1/2 \\
2 \\
1/2 \\
\end{pmatrix}
\]
Who’s #1?

For Colley, the teams are ranked (from best to worst) Panthers, **Saints** and for a tie for third between the Buccaneers and Falcons.
Further investigation

What further ideas can we explore in this area?

An alternate norm

- The Massey method used linear regression with the 2-norm, that is, least squares.
- What if we change norms?
- In his talk (9:30 tomorrow), Ken Massey will consider such methods, their computational algorithms, solution properties and goodness to fit using college football.
It’s one big game

- We can also apply ranking to other contexts.
- We’ve ranked movies from user ratings, social network graphs from the stance of influence, and social game users.
- **Question**: What’s a game?
Ranking movies

What’s a game?

- A game occurs over a prediction.
- Note, this necessitates ties.

Picture credit: http://www.herdaily.com/health/400/popcorn-a-healthy-snack.html
Ranking predictors

What’s a game?

• A game occurs over a prediction.
• Note, this necessitates ties.

Picture credit: http://www.prediculous.com
Ranking tweeters

What’s a game?
• One user follows another user
• One user retweets another
• ...

Games within the game

- We can also apply ranking sports beyond ranking teams.
- John Harris, Kevin Hutson and their students (Furman U.) ranked pitchers and batters.
- What’s a game? Each plate appearance.
Research

These research directions create new ideas, often out of need.

Picture credit: http://99u.com/tips/5661/Generate-Ideas-In-Moderation
All tied up

• Integrating ties can be important in many applications.
• While natural in Massey, Colley isn’t necessarily obvious.

Picture credit: http://tabithasbewitched.files.wordpress.com/2010/07/4306knot.jpg
Anyone’s game

Possibly more interesting than integrating ties was our treating close games as ties.

Colley Predictions for the NFL

- No Ties
- Induced Ties

Correct Predictions

- 63%
- 61%
- 59%
- 57%
- 55%
- 53%
- 51%
- 49%
- 47%

Season

- 1999
- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006
- 2007
- 2008
Evening the field

**Question:** How is ranking college teams different from the other contexts we just outlined?
Evening the field

**Question:** How is ranking college teams different from the other contexts we just outlined?

**Answer:** College teams play approximately the same number of games.
Evening the field

• Note, in the Colley method, if a team never plays, it’s rating is $\frac{1}{2}$ which will put it as an average team.
• In his talk (9:00 tomorrow), Chuck Wessell will discuss such issues and modifications that improve the rankings for tennis and golf.
Research how to rank

Two purposes of ranking are:

• to rank the teams for their play over an entire season, and

• to predict future play.

got purpose?
Apply in March

One time to test predictive methods is creating brackets for March Madness.

Measuring mojo

For March Madness, it may be helpful to differentiate between a team that

- lost its first 5 games but won its last 5 and
- won its first 5 games but lost its last 5
Adding weight

- Rather than counting each game as 1, you decide what to weight a game.
- A game may count as 1/5 or 5 games!

*Picture credit: http://howtogainweight123.com/gain-weight-wrong*
Colley’s diagonal

Each diagonal element equals $2 + t$, where $t$ equals the number of weighted games the corresponding team played.
Team to team

Each off-diagonal element equals \(-g\) where \(g\) is the number of weighted games the corresponding teams played.
Go to the right

Each element in the right-hand side vector equals $1 + 1/2(W - L)$, where again $W$ and $L$ are the number of weighted wins and losses for the corresponding team.
How do we do?

• We’ve entered our brackets into the ESPN Tournament Challenge and competed against more than 4 million brackets!

• You get 10 points for each correct choice in the first round and each successive round doubles in the points allotted to a correct prediction.
## 2009

<table>
<thead>
<tr>
<th>Method</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colley – no weight</td>
<td>62</td>
</tr>
<tr>
<td>Massey – no weight</td>
<td>88</td>
</tr>
<tr>
<td>Colley – bi-weekly</td>
<td>97</td>
</tr>
<tr>
<td>Massey – bi-weekly</td>
<td>79</td>
</tr>
<tr>
<td>Obama</td>
<td>80</td>
</tr>
<tr>
<td>Mike Greenberg (sports analyst)</td>
<td>70</td>
</tr>
<tr>
<td>Mike Golic (sports analyst)</td>
<td>43</td>
</tr>
</tbody>
</table>
10 and 11

- In 2010, a student in my math modeling class beat 99% of the close to 5 million brackets.
- In 2011, the best bracket beat most of the sports experts in the ESPN Challenge.
Bracketology

• I teach bracketology using sports ranking methods to majors and nonmajors.
• I will be replacing Amy Langville’s talk (10:00 tomorrow) to discuss bracketology in the classroom.
2 mine

Again, we can mine through data by:
- ranking items - who's the best?
- clustering items - who's like each other?

Rated Film

Rate the following films between -5 and 5

- 5 – really liked or want to see the film
- -5 – did not like or want to see the film

1. Inception
2. The King's Speech
3. Toy Story 3
# Taste test

Which pair has the most similar tastes?

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>3</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friend 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friend 2</td>
<td>-1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>You</td>
<td>4</td>
<td>1</td>
<td>-3</td>
</tr>
</tbody>
</table>

![Movie Posters](image1.png) ![Movie Posters](image2.png) ![Movie Posters](image3.png)
Creating distance

• Create ratings vectors like \([5, 3, -1], [-1, 2, 3]\) and \([4, 1, -3]\).

• **Method 1:** Find the Euclidean distance between the vectors.
Angling

**Method 2:** Find angle between the vectors using cosine similarity.

\[ \theta = \cos^{-1} \left( \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \right) \]
100-D dating

Comment in classes:
• Create a 100 question survey and find a date.
• What distance measure would you use?

I always knew my perfect mate was out there somewhere in 100 dimensional space.

Social fingerprint

• Tomorrow (8:30 AM), Michael Berry will discuss data forensics as applied to electronic socialization such as Facebook activity, mobile phone connections, email exchanges, tweets....

• Similarity measures applied to known social identities aid in this work.
Block buster

Here we have a movie by movie matrix were we kept the nearest neighbors of movies as measure by cosine similarity.

*Pictures thanks to David Gleich.*
Eigenvalue genres

Rearranging the order of the movies uncovers well-connected components of the graph.

**Tool:** An eigenvector!
Community detection

• Tomorrow, Van Emden Henson (10:30) will discuss the importance of bipartite or near-bipartite components in a community network.
• Spectral-based methods aid in identifying strong near-bipartite communities.
Adding data

At this meeting alone, many of us will add data to the mass of digital information.

When life is linear

• Linear algebra is a frequent tool used to analyze our often nonlinear world.
• Accessible to students in teaching and research.
• Where to start? What are your ideas?

Photo credit: http://commons.wikimedia.org/wiki/File:Triple-Spiral-labyrinth-variant.png