Leigh L. Noble, PhD
noble@leighnoble.org

Spuyten Duyvil Undergraduate Mathematics Conference
Some Questions

- What’s a Google bomb?
- How does Google order the results of your search?
- Does “mutual linking” improve a site’s position?

MATH: Use linear algebra to calculate importance scores

- Nice mini-web example
- Dangling node
- Disconnected web
- The Google matrix
Search for **french military victories**

Did you mean: **french military defeats**

No standard web pages containing all your search terms were found.

Your search - **french military victories** - did not match any documents.

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.
- Try fewer keywords.

Also, you can try [Google Answers](http://www.google.com/answers) for expert help with your search.

---

Parody transcripted ©2003 Albino Blacksheep
Search for miserable failure

Biography of George W. Bush
WhiteHouse.gov is the official website for the White House and President Barack ... 6. John Quincy Adams. 7. Andrew Jackson. 8. Martin Van Buren. 9. William ... www.whitehouse.gov/about/presidents/georgewbush - 51k - Cached

Political Google bombs in the 2004 U.S. Presidential election ...
First political... First political... Impact Google's response
During the 2004 U.S. presidential election, Google bombs were used to further various political agendas. Two of the first were the "miserable failure" Google bomb linked to George W. Bush's White House biography and the "waffles" Google bomb linked to John Kerry's...
en.wikipedia.org/wiki/Political_Google_bombs_in_the_2004_U.S._Presi... - Cached

snopes.com: Miserable Failure
Why is the phrase 'miserable failure' tied to President Bush's biography in Google? ... Google Halts 'Miserable Failure' Link to President Bush." The New York Times. ...
www.snopes.com/politics/bush/google.asp
Search for out of touch executives

Corporate Information - Google Management
Susan Wojcicki, Vice President, Product Management. Key executives by function: Engineering ... In Ann Arbor, Larry built an inkjet printer out of Lego™ bricks. ...
www.google.com/corporate/execs.html - 104k - Cached

Reach (C)
Job-hunter says Bank ...
www.cfo.com

AIG exec
On the eve of firing 150,000 workers, 2,000 multi-ethnic employees harbored a TV in Oppenheimer's office and ...
out of touch with reality. March 28, 2009. March 28, 2009 ...
statesmanjournal.com/article/20090328/OPINION.../1050/COMMUNITIES - 74k - Cached

Gizmodo - Letter from a Moto Insider: How Stupid Execs Ran Moto Into ...
... Motorola CEO Greg Brown about how a cabal of inept, out-of-touch ... cabal of inept, out-of-touch executives more worried about their golf score than ...
gizmodo.com/372565/letter-from-a-moto-insider-how-stupid-execs-ran-... - 53k
“Link Bomb” is more descriptive

Works off anchor text:
Click here to read about a <a href = "http://www.whitehouse.gov/about/presidents/georgewbush">miserable failure</a>.

Creates this text on a web page:
Blah blah blah, and more blah. Click here to read about a miserable failure. And now for something completely different...

The phrase miserable failure is now associated with the bomb recipient's website.
Crawl / Webspider

Index the data the spider gathers

Search and Presentation of User Interface
Historical Perspective

→ Exact IP address and file name/location

1990 → Archie: only searched titles on per-server basis
→ Veronica: fancier version of Archie
→ WebCrawler: full text search

1994 → Lycos: used anchor text to rate relevance
→ Yahoo: directory, no search

1995 → AltaVista: indexed 10 million docs in 1995
→ Excite: grouped pages by keyword

1996 → Google: BackRub+PageRank, later added full-text search

* Factoid: From 1993 through 1996 the WWW grew from 130 sites to > 600,000
Importance Scores

Idea behind PageRank algorithm:
Each page lends part of its importance to the pages it links to.

\[
R(p_k) = \sum_{p_j \in B_k} \frac{R(p_j)}{n_j}
\]

- \(B_k\) = all pages linking to \(p_k\) (aka backlinks)
- \(n_j\) = the number of links going OUT of page \(p_j\)
Example 1: Best Case

\[ R(p_k) = \sum_{p_j \in B_k} \frac{R(p_j)}{n_j} \]

- \( R(p_1) = 0 \cdot R(p_1) + 1 \cdot R(p_2) + 1/2 \cdot R(p_3) + 0 \cdot R(p_4) \)
- \( R(p_2) = 1/3 \cdot R(p_1) + 0 \cdot R(p_2) + 1/2 \cdot R(p_3) + 0 \cdot R(p_4) \)
- \( R(p_3) = 1/3 \cdot R(p_1) + 0 \cdot R(p_2) + 0 \cdot R(p_3) + 1 \cdot R(p_4) \)
- \( R(p_4) = 1/3 \cdot R(p_1) + 0 \cdot R(p_2) + 0 \cdot R(p_3) + 0 \cdot R(p_4) \)
\[
\begin{bmatrix}
0 & 1 & \frac{1}{2} & 0 \\
\frac{1}{3} & 0 & \frac{1}{2} & 0 \\
\frac{1}{3} & 0 & 0 & 1 \\
\frac{1}{3} & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4)
\end{bmatrix}
= 
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4)
\end{bmatrix}
\]

Ways to solve:

- Like a traditional eigenvector problem with eigenvalue \( \lambda = 1 \)
- Or find a stationary vector so that \( P^m R = R \).
Normalize column sum to 1:

\[
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4)
\end{bmatrix} = \begin{bmatrix} 3 \\ 2 \\ 2 \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0.375 \\ 0.250 \\ 0.250 \\ 0.125 \end{bmatrix}
\]

Page 1 has the highest importance score!
Example 2: Dangling Node

\[
R(p_k) = \sum_{p_j \in B_k} \frac{R(p_j)}{n_j}
\]

\[
\begin{bmatrix}
0 & 1 & \frac{1}{2} & 0 \\
\frac{1}{3} & 0 & \frac{1}{2} & 0 \\
\frac{1}{3} & 0 & 0 & 0 \\
\frac{1}{3} & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4)
\end{bmatrix}
= 
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4)
\end{bmatrix}
\]

Has no eigenvalue of 1!

\[
\lambda \in \{0, \ 0.836, \ -0.418 \pm 0.156i\}
\]
Random Surfer to the rescue

When a page has no link OUT, treat it as if the page links everywhere. Replace $P$ with

$$
\begin{bmatrix}
0 & 1 & 1/2 & 1/4 \\
1/3 & 0 & 1/2 & 1/4 \\
1/3 & 0 & 0 & 1/4 \\
1/3 & 0 & 0 & 1/4 \\
\end{bmatrix}
$$

Solution:

$$
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4)
\end{bmatrix}
= c
\begin{bmatrix}
9/4 \\
3/2 \\
1 \\
1
\end{bmatrix}
\longrightarrow
\begin{bmatrix}
0.391 \\
0.261 \\
0.174 \\
0.174
\end{bmatrix}
$$
Example 3: Disconnected Web

\[
\begin{bmatrix}
0 & 1 & 1/2 & 0 & 0 \\
1/2 & 0 & 1/2 & 0 & 0 \\
1/2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 \\
\end{bmatrix}
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4) \\
R(p_5) \\
\end{bmatrix}
= 
\begin{bmatrix}
R(p_1) \\
R(p_2) \\
R(p_3) \\
R(p_4) \\
R(p_5) \\
\end{bmatrix}
\]

Eigenvectors for \( \lambda = 1 \):

\[
\begin{bmatrix}
0.444 \\
0.333 \\
0.222 \\
0 \\
0 \\
\end{bmatrix} \quad \text{and} \quad 
\begin{bmatrix}
0 \\
0 \\
0.5 \\
0.5 \\
\end{bmatrix}
\]
A little theory

These had unique solutions
- No dangling nodes!
- Matrices were column stochastic

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

\[
\begin{bmatrix}
0 & 1 & \frac{1}{2} & \frac{1}{4} \\
\frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{4} \\
\frac{1}{3} & 0 & 0 & \frac{1}{4} \\
\frac{1}{3} & 0 & 0 & \frac{1}{4}
\end{bmatrix}
\]
The Google Matrix

Theorem

For a column stochastic matrix $G$ with all positive entries, there exists a unique vector $q$ with all positive entries which is also column stochastic so that $Gq = q$.

Consider an $n \times n$ matrix

$$G = (1 - \alpha)P + \alpha N$$

where $N_{ij} = \frac{1}{n}$ and $P$ has no dangling nodes.

$G$ guarantees we get a page ranking we can use.

* Factoid: In 2007 Google altered their search engine ranking system to counteract link bombs.
You trade links in attempt to improve your rank...

You were tied for 2nd place. What happens to your ranking?

You will move up! But this doesn’t always happen. [Try it.]
Google’s efficient and effective algorithm has changed our generation’s ability to find info (and entertainment!) we seek.

Still, Challenges Ahead –

- The web is still expanding and evolving at an enormous pace.
- The *type of content* is rapidly shifting away from text based content to flash, video, and other content.
- *Recommendation engines* are another wave of innovation based on powerful linear algebra and numerical analysis techniques.
References

- Kurt Bryan & Tanya Leise
  The $25,000,000,000$ Eigenvector: The Lin.Alg.behind Google

- David Austin
  How Google Finds Your Needle in the Web’s Haystack.
  *Feature Column at AMS website, Dec 2006.*

- Michael W. Berry & Murray Browne
  Understanding Search Engines: Mathematical Modeling. . . 2nd Ed.
  *SIAM, 2005.*

- John Battelle
  The Search: How Google and Its Rivals Rewrote the Rules. . .
  *Portfolio, 2005.*

- Anton and Rorres
  Elementary Linear Algebra, Applications Version, 8th Ed.
  *Wiley & Sons, 2000, pp. 581-583.*
Ordering of Search Engine Results: google it!

Thank you for listening!

http://www.LeighNoble.org/