Centrality Measures In The Real World

David Kravitz
dkravitz78@gmail.com

Graph Questions For Today

- Given a record of all phone calls made in one town over a period of one month, who is the most influential person?
- 11 games into the college football season, if 13 teams are all 10-1, how do we (fairly*) decide which two teams play for the national championship?
- If I want to advertise through Twitter by spending money to get 3 famous people to tweet about my product, how do I decide which people to pursue?
- If I'm inventing a search engine, how do I decide which pages to rank the highest?

A real-world problem

- Suppose we are told that drug lords are using their phones to facilitate runs in a certain town, and we have all their phone records.
- How do we find the head people in this operation just by phone records?

Ideas?

- Find who makes the most calls.
  High volume=high importance.

Ideas?

- Find who makes the most calls.
  High volume=high importance.
- Most of the high-volume people are middle-men, taking orders from bosses and finding customers.
- The other high-volume callers are nicknamed “pizza shops” (incoming) and “fundraisers” (outgoing). They are almost always unimportant for networking.

Other Ideas?

- Find who only has incoming calls?
- Find who only has outgoing calls?
- Find large ‘cliques’ who all call each other but very few others?

Remember this problem...

Measuring Influence via Twitter

- How do we find out which tweeters are most influential?

Ideas?

- Whoever has the most tweets @ them is the most influential.
- Whoever has the most followers is the most influential.
- Whoever gets the most replies is the most influential.

Ideas?

- Whoever has the most tweets @ them is the most influential.
- Whoever has the most followers is the most influential.
- Whoever gets the most replies is the most influential.
- In 2012, Justin Beiber was top in all 3 categories, more than double anyone else, every single day of the year.

Even Worse

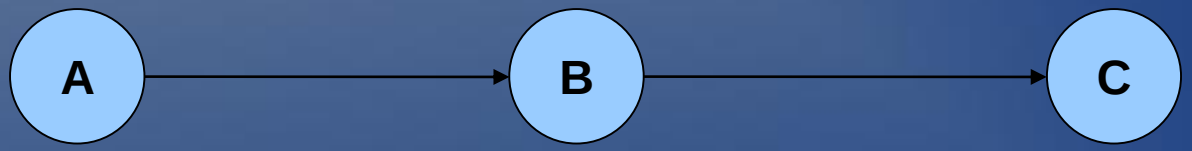
- In 2012, the closest that Justin Beiber ever came to not being number one was on April 18, 2012.
- On that day, Dick Clark was number 2. Dick Clark received 47% of the directed tweets that Justin Beiber did on April 18th.

Even Worse

- In 2012, the closest that Justin Beiber ever came to not being number one was on April 18, 2012.
- On that day, Dick Clark was number 2. Dick Clark received 47% of the directed tweets that Justin Beiber did on April 18th.
- April 18, 2012 was the day Dick Clark passed away.

A Smarter Algorithm

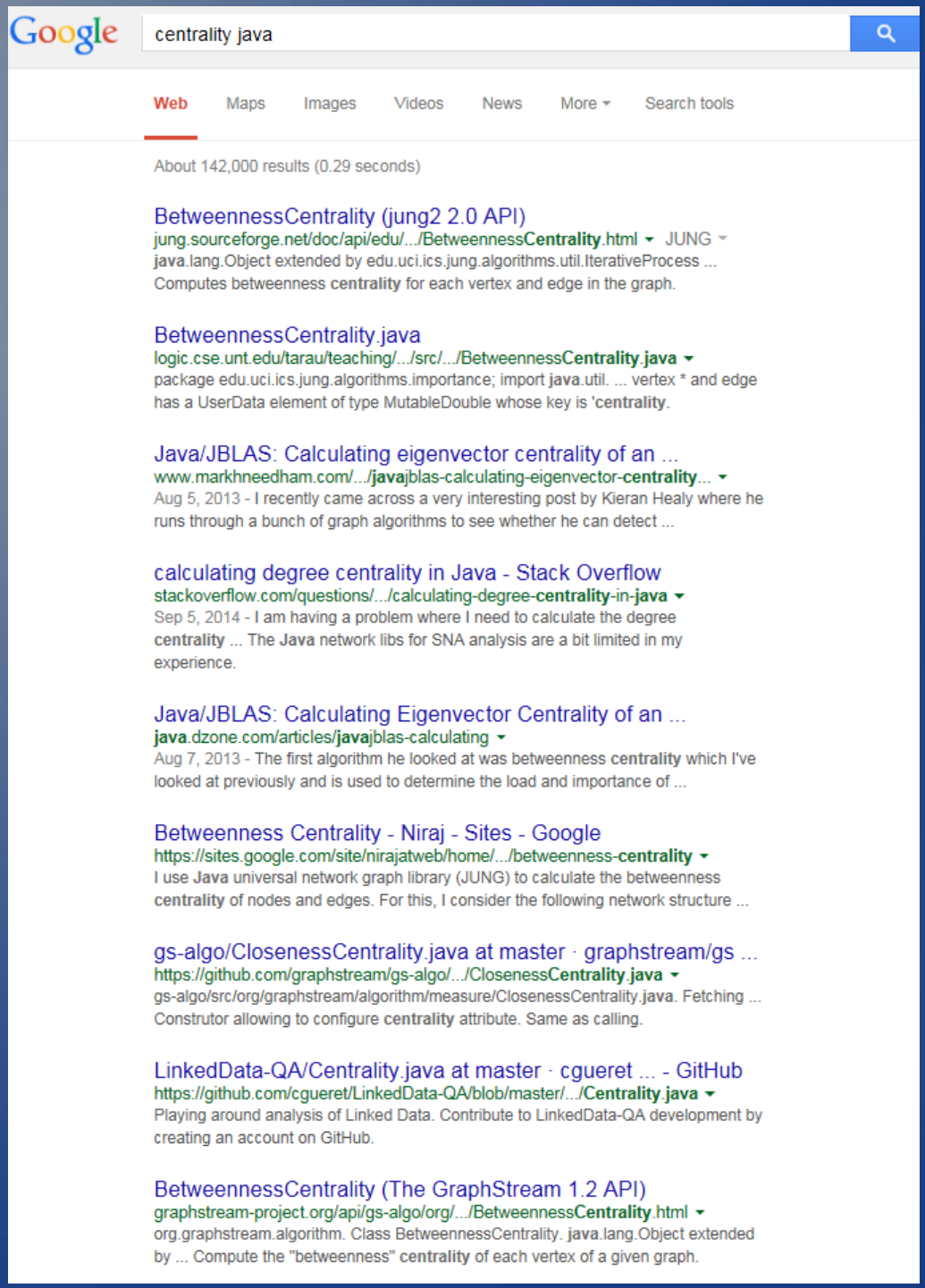
- If A makes a call to B, and B makes a call to C, but A never called C, then B is on the shortest path between A and C.
- Idea: More shorter paths = more importance.

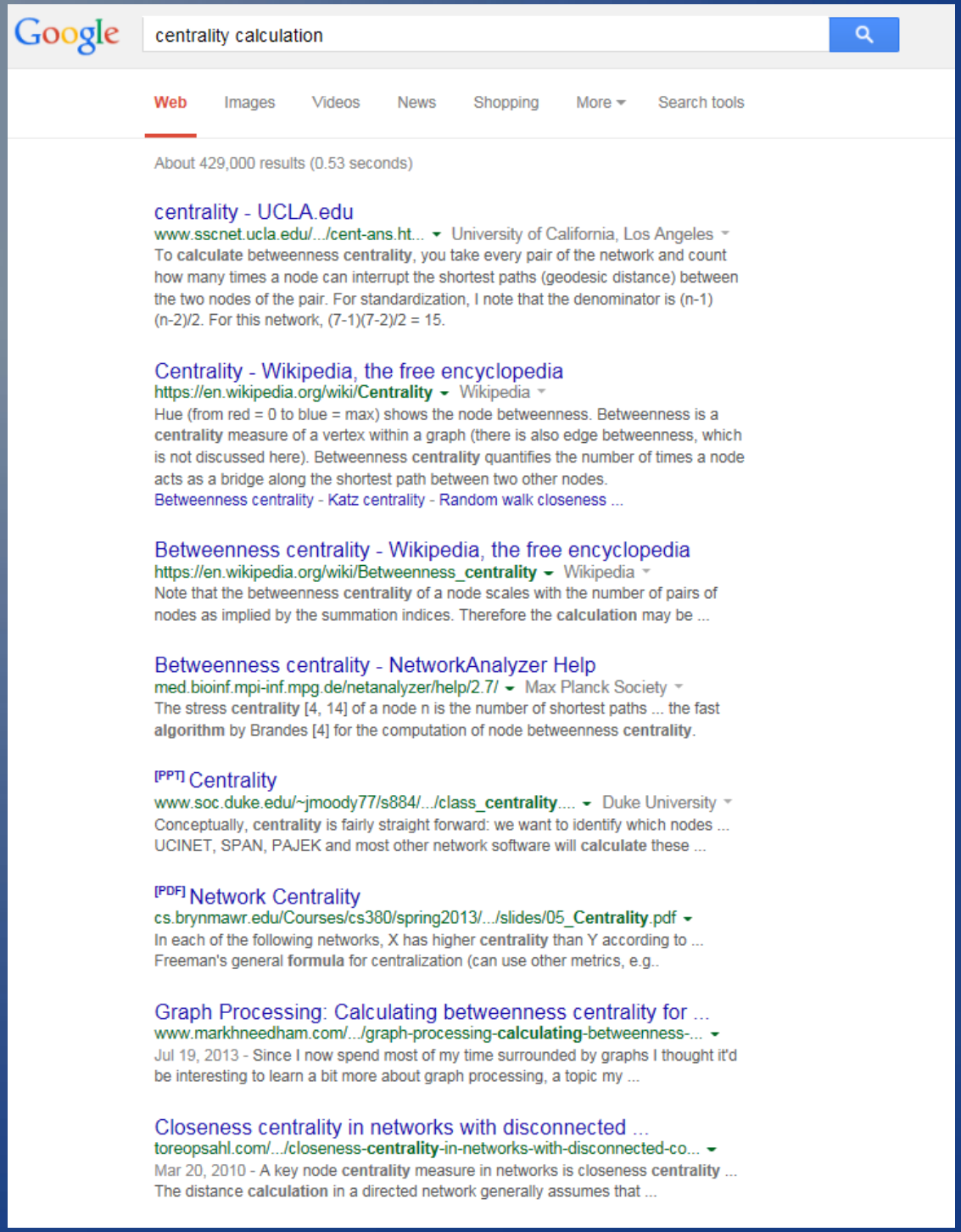


- Kevin Bacon showed us that most shortest paths are length 6 or less.

Betweenness Centrality

- For all pairs of vertices in a graph, find the shortest path between them.
- Each occurrence of a vertex on a shortest path is one point.
- Total points (or a ratio therein) is the betweenness centrality.
- Idea: More points = more important.


Google
centrality java
Web
Maps
Images
Videos
News
More
Search tools
About 142,000 results (0.29 seconds)
BetweennessCentrality (jung2 2.0 API)
jung.sourceforge.net/doc/api/edu/.../BetweennessCentrality.html JUNG
java.lang.Object extended by edu.uci.ics.jung.algorithms.util.IterativeProcess ...
Computes betweenness centrality for each vertex and edge in the graph.
BetweennessCentrality.java
logic.cse.unt.edu/tarau/teaching/.../src/.../BetweennessCentrality.java
package edu.uci.ics.jung.algorithms.importance; import java.util. ... vertex * and edge has a UserData element of type MutableDouble whose key is 'centrality.
Java/JBLAS: Calculating eigenvector centrality of an ...
www.markhneedham.com/.../javajblas-calculating-eigenvector-centrality...
Aug 5, 2013 - I recently came across a very interesting post by Kieran Healy where he runs through a bunch of graph algorithms to see whether he can detect ...
calculating degree centrality in Java - Stack Overflow
stackoverflow.com/questions/.../calculating-degree-centrality-in-java
Sep 5, 2014 - I am having a problem where I need to calculate the degree centrality ... The Java network libs for SNA analysis are a bit limited in my experience.
Java/JBLAS: Calculating Eigenvector Centrality of an ...
java.dzone.com/articles/javajblas-calculating
Aug 7, 2013 - The first algorithm he looked at was betweenness centrality which I've looked at previously and is used to determine the load and importance of ...
Betweenness Centrality - Niraj - Sites - Google
https://sites.google.com/site/nirajatweb/home/.../betweenness-centrality
I use Java universal network graph library (JUNG) to calculate the betweenness centrality of nodes and edges. For this, I consider the following network structure ...
gs-algo/ClosenessCentrality.java at master · graphstream/gs ...
https://github.com/graphstream/gs-algo/.../ClosenessCentrality.java
gs-algo/src/org/graphstream/algorithm/measure/ClosenessCentrality.java. Fetching ... Construtor allowing to configure centrality attribute. Same as calling.
LinkedData-QA/Centrality.java at master · cgueret ... - GitHub
https://github.com/cgueret/LinkedData-QA/blob/master/.../Centrality.java
Playing around analysis of Linked Data. Contribute to LinkedData-QA development by creating an account on GitHub.
BetweennessCentrality (The GraphStream 1.2 API)
graphstream-project.org/api/gs-algo/org/.../BetweennessCentrality.html
org.graphstream.algorithm. Class BetweennessCentrality. java.lang.Object extended by ... Compute the "betweenness" centrality of each vertex of a given graph.


Google
centrality calculation
Web
Images
Videos
News
Shopping
More
Search tools
About 429,000 results (0.53 seconds)
centrality - UCLA.edu
www.sscnet.ucla.edu/.../cent-ans.ht... University of California, Los Angeles
To calculate betweenness centrality, you take every pair of the network and count how many times a node can interrupt the shortest paths (geodesic distance) between the two nodes of the pair. For standardization, I note that the denominator is (n-1)(n-2)/2. For this network, (7-1)(7-2)/2 = 15.
Centrality - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Centrality Wikipedia
Hue (from red = 0 to blue = max) shows the node betweenness. Betweenness is a centrality measure of a vertex within a graph (there is also edge betweenness, which is not discussed here). Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.
Betweenness centrality - Katz centrality - Random walk closeness ...
Betweenness centrality - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/Betweenness_centrality Wikipedia
Note that the betweenness centrality of a node scales with the number of pairs of nodes as implied by the summation indices. Therefore the calculation may be ...
Betweenness centrality - NetworkAnalyzer Help
med.bioinf.mpi-inf.mpg.de/netanalyzer/help/2.7/ Max Planck Society
The stress centrality [4, 14] of a node n is the number of shortest paths ... the fast algorithm by Brandes [4] for the computation of node betweenness centrality.
[PPT] Centrality
www.soc.duke.edu/~jmoody77/s884/.../class_centrality.... Duke University
Conceptually, centrality is fairly straight forward: we want to identify which nodes ... UCINET, SPAN, PAJEK and most other network software will calculate these ...
[PDF] Network Centrality
cs.brynmawr.edu/Courses/cs380/spring2013/.../slides/05_Centrality.pdf
In each of the following networks, X has higher centrality than Y according to ... Freeman's general formula for centralization (can use other metrics, e.g..
Graph Processing: Calculating betweenness centrality for ...
www.markhneedham.com/.../graph-processing-calculating-betweenness-...
Jul 19, 2013 - Since I now spend most of my time surrounded by graphs I thought it'd be interesting to learn a bit more about graph processing, a topic my ...
Closeness centrality in networks with disconnected ...
toreopsahl.com/.../closeness-centrality-in-networks-with-disconnected-co...
Mar 20, 2010 - A key node centrality measure in networks is closeness centrality ... The distance calculation in a directed network generally assumes that ...

Great Idea?

- Important people are on more shortest paths, especially celebrities and mavens.
- Pizza shops and fundraisers don’t call both ways, so they are not on shortest paths.
- Computation complexity is not much worse than quadratic, this is acceptable for a lot of problems.
- For several years, this was the preferred method by our government customer.

Airports

- We put every airport and flight in North America into a database and measured betweenness centrality.
- Houston being on the shortest path from Baltimore to San Diego means one point for Houston. Ties split the point.
- It turns out there was a landslide winner for “most important airport in the US.” More than double second place, triple third place.

Betweenness Centrality

1	ANC (Anchorage, AK, USA)	465272
2	FAI (Fairbanks, AK, USA)	215503
3	YYZ (Toronto, Canada, Canada)	131562
4	LAX (Los Angeles, CA, USA)	129246
5	SEA (Seattle/Tacoma, WA, USA)	125151
6	JFK (New York, NY, USA)	124927
7	HPN (White Plains, NY, USA)	121096
8	MIA (Miami, FL, USA)	120643
9	DEN (Denver, CO, USA)	120342
10	MSP (Minneapolis, MN, USA)	111188

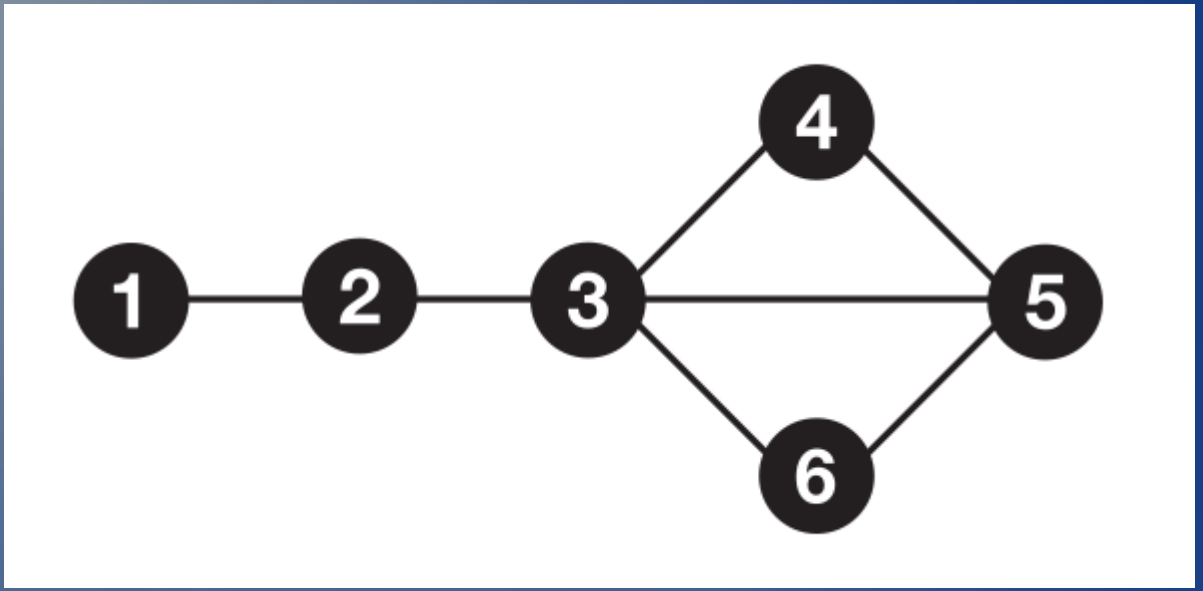
WHAT?

- Alaska has hundreds of airports. All of them fly through Anchorage or Fairbanks, mostly Anchorage.
- Betweenness centrality counted each of them equal to New York City, even with only one flight per month.

Other Problems

- Justin Beiber obviously was first place in betweenness centrality, but almost all of his tweets were from “teenage girls.”
- College Football had one year with a “rock-paper-scissors” situation, 3 teams with similar schedules all beat each other for their only loss. They tie under this method.
- A home page automatically routes from www.david.com to www.david.org, which wrongly makes www.david.com look super-important.

New Method - An Example



$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

$$B = A + I = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

Eigenvalues

- Given a matrix M, an eigenvalue e and eigenvector v are such that $Mv = ev$.
  Further, for any exponent n, $M^n v = e^n v$.

- A “nice” matrix can be completely described by its eigenvalues and eigenvectors.

Eigenvalues

- Given a matrix M, an eigenvalue e and eigenvector v are such that $Mv = ev$.
  Further, for any exponent n, $M^n v = e^n v$.

- A “nice” matrix can be completely described by its eigenvalues and eigenvectors.

- Claim 1: The question relevant to probability is “What happens when the matrix is raised to an infinite exponent?”

Eigenvalues

- Given a matrix M, an eigenvalue e and eigenvector v are such that $Mv = ev$.
  Further, for any exponent n, $M^n v = e^n v$.

- A “nice” matrix can be completely described by its eigenvalues and eigenvectors.

- Claim 1: The question relevant to probability is “What happens when the matrix is raised to an infinite exponent?”

- Claim 2: When looking at large/infinite exponents of a matrix, all that really matters is the principal eigenvalue.

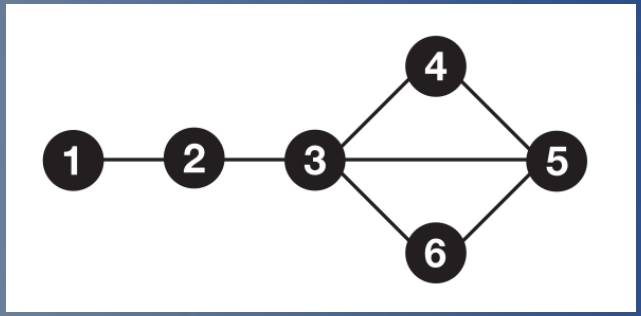
An Example

- In our example, the principal eigenvalue is 2.70559.
- The other eigenvalues are -1.851, -1.350, 1.056, -0.560, and 0.

$$2.70559$$

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

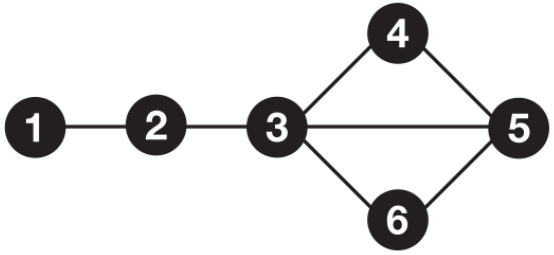
$$\mathbf{v_0} = \begin{bmatrix} 0.092 \\ 0.249 \\ 0.581 \\ 0.405 \\ 0.514 \\ 0.405 \end{bmatrix}$$



$$\mathbf{d} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$B = A + I = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{Bd} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{B^2 d} = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{B^3 d} = \begin{bmatrix} 5 \\ 7 \\ 8 \\ 4 \\ 5 \\ 4 \end{bmatrix}$$

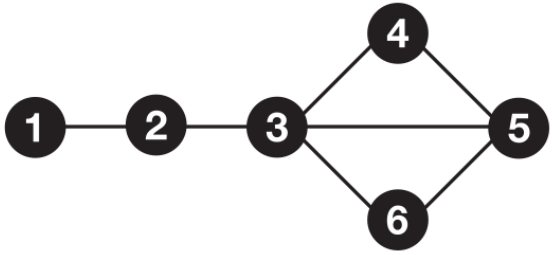


$$\mathbf{d} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$B = A + I = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{d}_k = \frac{\mathbf{B}^k \mathbf{d}}{|\mathbf{B}^k \mathbf{d}|}$$

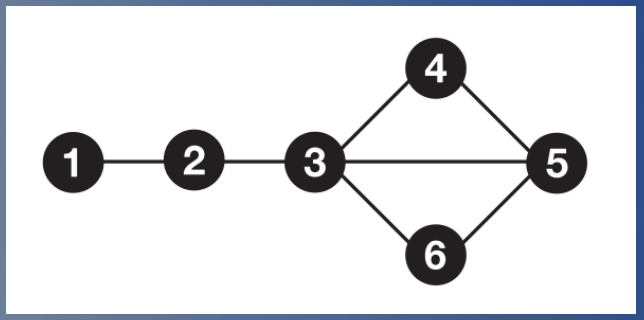
$$\mathbf{d}_1 = \begin{bmatrix} 0.577 \\ 0.577 \\ 0.577 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{d}_2 = \begin{bmatrix} 0.447 \\ 0.671 \\ 0.447 \\ 0.224 \\ 0.224 \\ 0.224 \end{bmatrix}, \quad \mathbf{d}_3 = \begin{bmatrix} 0.358 \\ 0.501 \\ 0.573 \\ 0.286 \\ 0.358 \\ 0.286 \end{bmatrix}$$



$$\mathbf{d} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$B = A + I = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{d}_1 = \begin{bmatrix} 0.577 \\ 0.577 \\ 0.577 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{d}_2 = \begin{bmatrix} 0.447 \\ 0.671 \\ 0.447 \\ 0.224 \\ 0.224 \\ 0.224 \end{bmatrix}, \quad \mathbf{d}_3 = \begin{bmatrix} 0.358 \\ 0.501 \\ 0.573 \\ 0.286 \\ 0.358 \\ 0.286 \end{bmatrix}, \quad \mathbf{d}_{30} = \begin{bmatrix} 0.092 \\ 0.249 \\ 0.581 \\ 0.405 \\ 0.514 \\ 0.405 \end{bmatrix}$$



$$\mathbf{d} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$B = A + I = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

$$\mathbf{d}_1 = \begin{bmatrix} 0.577 \\ 0.577 \\ 0.577 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{d}_2 = \begin{bmatrix} 0.447 \\ 0.671 \\ 0.447 \\ 0.224 \\ 0.224 \\ 0.224 \end{bmatrix}, \quad \mathbf{d}_3 = \begin{bmatrix} 0.358 \\ 0.501 \\ 0.573 \\ 0.286 \\ 0.358 \\ 0.286 \end{bmatrix}, \quad \mathbf{d}_{30} = \begin{bmatrix} 0.092 \\ 0.249 \\ 0.581 \\ 0.405 \\ 0.514 \\ 0.405 \end{bmatrix}$$

$$\mathbf{v_0} = \begin{bmatrix} 0.092 \\ 0.249 \\ 0.581 \\ 0.405 \\ 0.514 \\ 0.405 \end{bmatrix}$$

Math!

Perron-Frobenius Theorem:

If M is a “nice” matrix then M has a principal eigenvalue such that all entries in its corresponding eigenvector are positive.

Gould’s Index:

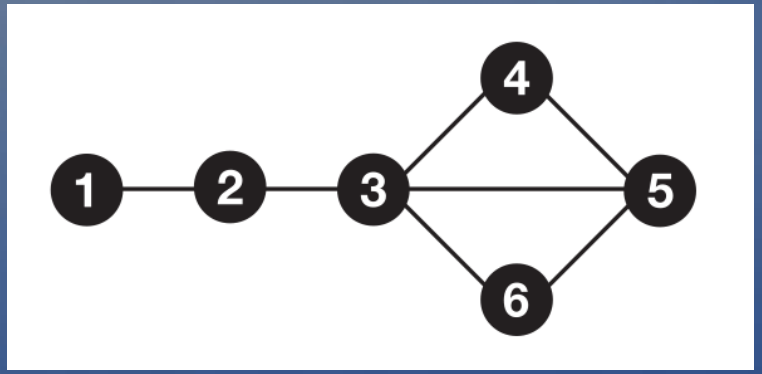
The principal eigenvector, once normalized, gives an accessibility rank to each vertex.

Eigenvector Centrality

- The eigenvector centrality is defined by the normalized principal eigenvector, a.k.a. Gould's Index, which has all positive entries.
- The principal eigenvector can be computed relatively easily with a power iteration method.
- In practice, the power iteration method is accomplished with edge lists, and can be multiplied in parallel.

Back to Our Example

- The maximum centrality belongs to vertex 3, followed by 5.
- Vertex 2 is 5th highest, only exceeding that of 1.



$$\mathbf{v_0} = \begin{bmatrix} 0.092 \\ 0.249 \\ 0.581 \\ 0.405 \\ 0.514 \\ 0.405 \end{bmatrix}$$

- Note that betweenness centrality of vertex 2 was second highest.

Google PageRank

- Google does not publicize their exact algorithm, but it initially used a variant of eigenvalue centrality.
- It is believed that the original PageRank randomly crawled the web, following random links on each page, and jumped to another page with probability ~10% .
- Pages visited more often have higher scores. Higher scores means higher on search results.

Advantages

- Unimportant nodes, like Alaskan Airports and teenage girls, do not count for much each.
- Only having high out-degree does not boost the score.
- No human input is needed. This means no college football voting, no celebrity lists, no lists of suspected drug lords.
- Edge weights are allowed in the computation. This is huge for airports, and often for calls as well.

New Results

ATL	Atlanta	Georgia	0.427686
ORD	Chicago	Illinois	0.248052
LAX	Los Angeles	California	0.228303
DEN	Denver	Colorado	0.204601
DFW	Dallas	Texas	0.190705
PHX	Phoenix	Arizona	0.168223
JFK	New York	New York	0.163633
MSY	New Orleans	Louisiana	0.150766
LAS	Las Vegas	Nevada	0.149357
PHL	Philadelphia	Pennsylvania	0.148795
MSP	Minneapolis	Minnesota	0.148038
MCO	Orlando	Florida	0.145595
CLT	Charlotte	North Carolina	0.144368
SFO	San Fransisco	California	0.143491
BOS	Boston	Massachusetts	0.130848
DTW	Detroit	Michigan	0.129700
MIA	Miami	Florida	0.128554
DCA	Washington, DC	Washington, DC	0.126660
SEA	Seattle	Washington	0.122803
IAH	Houston	Texas	0.118719
BWI	Baltimore	Maryland	0.115619

New Results - Worldwide

ATL	Atlanta	United States	0.271987
LHR	London	United Kingdom	0.209556
ORD	Chicago	United States	0.202435
JFK	New York	United States	0.192145
LAX	Los Angeles	United States	0.187259
CDG	Paris	France	0.158700
DFW	Dallas	United States	0.143737
FRA	Frankfurt	Germany	0.143401
SFO	San Fransisco	United States	0.127507
YYZ	Toronto	Canada	0.124394
AMS	Amsterdam	Netherlands	0.119998
PEK	Beijing	China	0.119029
MIA	Miami	United States	0.118796
DEN	Denver	United States	0.116140
PVG	Shanghai	China	0.111080
ICN	Seoul	South Korea	0.111004
NRT	Tokyo	Japan	0.108700
FCO	Rome	Italy	0.106034
MAD	Madrid	Spain	0.105972
PHL	Philadelphia	United States	0.105972

New Results

- College football uses a version of this algorithm, but it still incorporates voting. While they do not make their methods public, it is believed that voting weights the random steps when they are taken.
- Google still uses a later version of PageRank, eventually they shifted to make physical location a huge part of the score.

Twitter

- The most influential tweeter ...

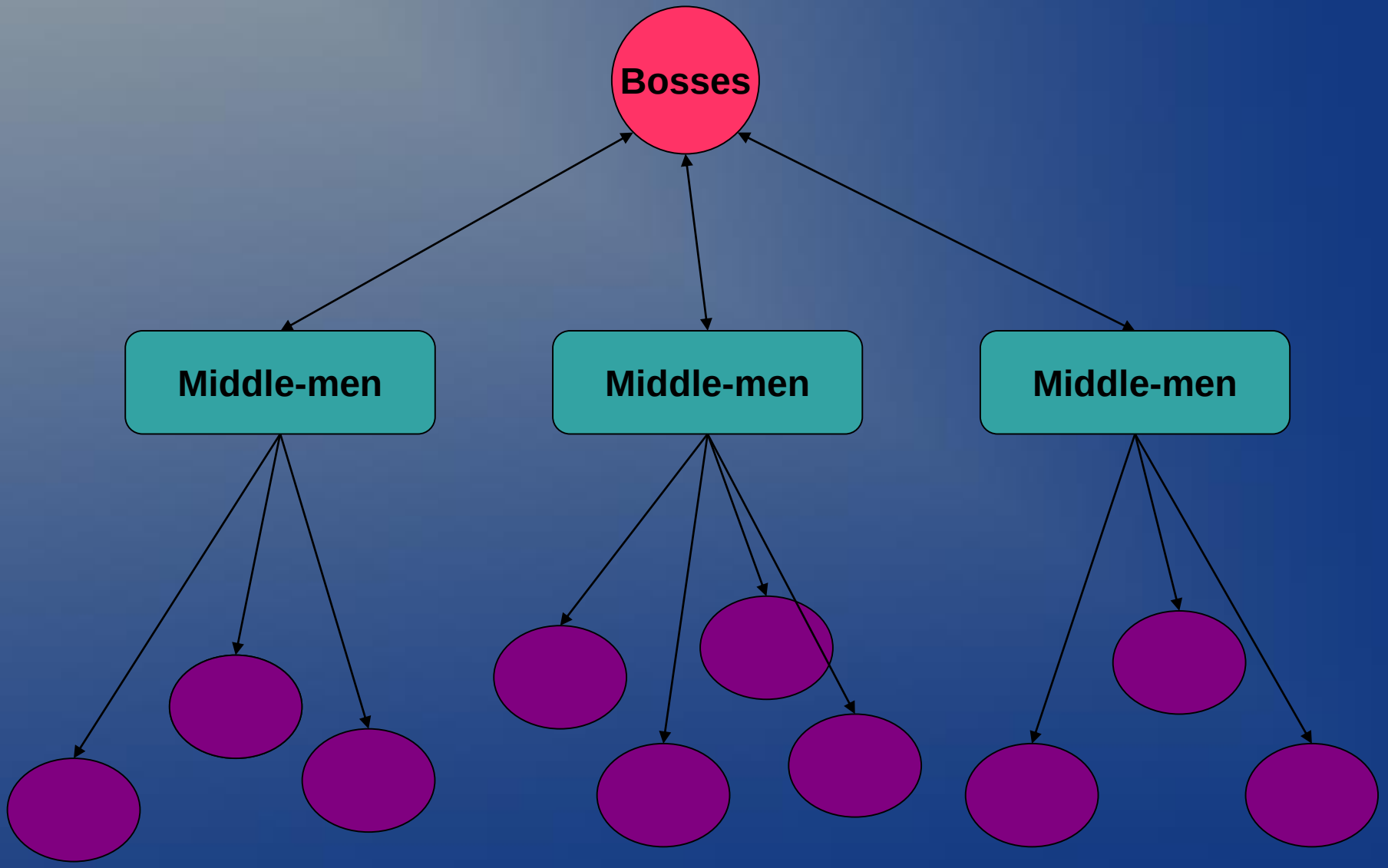
Twitter

- The most influential tweeter ... is still Justin Beiber. A lot of celebrities do tweet to him, albeit not always positive.
- The insight we provided is that teenage girls make up a majority on twitter. Other groups definitely use it, but in the minority.
- Analytics provide huge insights into our data. We don’t always love those insights.

Twitter-like Data?

- The exact same algorithm was conducted on 2 data sets, one email and one telephony data.
- These data sets are far from complete, which made well-known clustering algorithms impossible to implement without restricting to a much smaller subgraph.
- Eigenvector centrality provided lists of possible seeds which led to much different starting groups, and many reports of much higher effectiveness on clustering algorithms.

Did We Find the Drug Bosses?



Thank you very much!

Questions?