Graph Centrality

J.T. Liso, Sean Whalen
4/12/2018

Test Questions

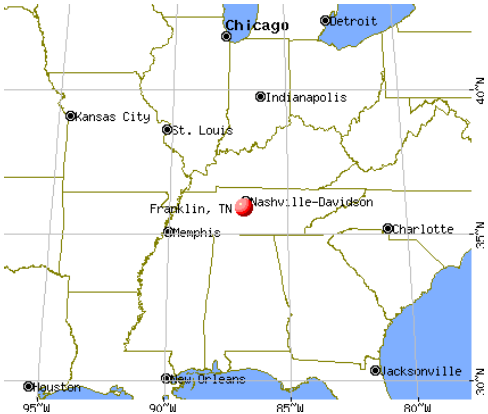
1. What are the two ways we can define importance in terms of a graph?

2. What is the output of PageRank?

3. What state(s) are considered the center of the contiguous USA based on degree centrality?

About J.T.

- 5-year BS/MS under Jens Gregor
  - Intelligent Modeling of Network Traffic
- River Edge, NJ -> Franklin, TN
- Hobbies:
  - Hiking/Camping
  - All things music
  - Weightlifting
  - Cooking

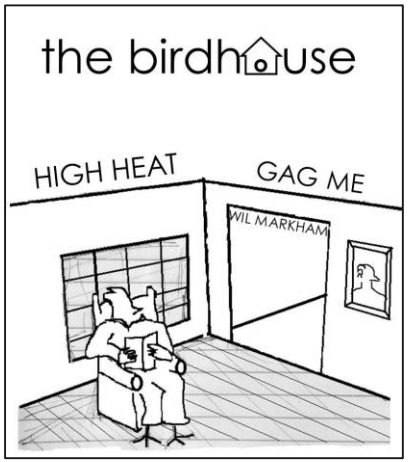


About Sean

At Rhythm ‘n Blooms with WUTK

Garage in Soddy-Daisy

- 5-year BS/MS
  - GTA for CS130
- Clinton, IL -> Ontario, NY -> Soddy, TN
- Interests:
  - Drumming
  - Sci-fi books & movies
- Like High Heat on facebook!
  - Next show is **TONIGHT** at the Birdhouse



Outline

1. Overview
2. History
3. Algorithms
    a. Degree Centrality
    b. Closeness Centrality
    c. Betweenness Centrality
    d. Eigenvector Centrality
    e. PageRank
4. Applications
5. Implementations
6. Open Issues
7. Discussion

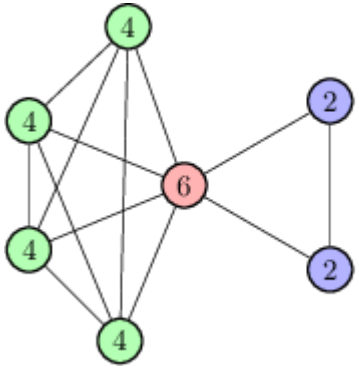
Overview

Graph Basics

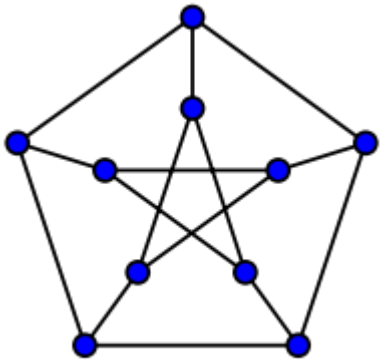
- Simple, finite, undirected/directed, weighted/unweighted
- Vertices (V) and edges (E)
- **Adjacency matrix:** V x V matrix where a 1 represents an edge between vertex u and vertex v
    - Use weight instead of 1 for weighted graphs
- **Degree:** number of edges incident upon a vertex
- **Path:** sequence of edges between two vertices
- **Distance:** number of edges in a path

Examples

- Diameter: 2
- Radius: 1
- Girth: 3



- Each vertex has degree 3
- Max distance of 2 between all pairs of vertices
- Girth: 5



What is centrality?

- Identifying most “important” vertices in a graph
- Importance has many different meanings when applied to graph theory
  - Characterization by flow through a vertex
  - Characterization by cohesiveness (focus of our talk)
- Different algorithms reveal different properties based on centrality relating to these definitions

Characterization by Network Flow

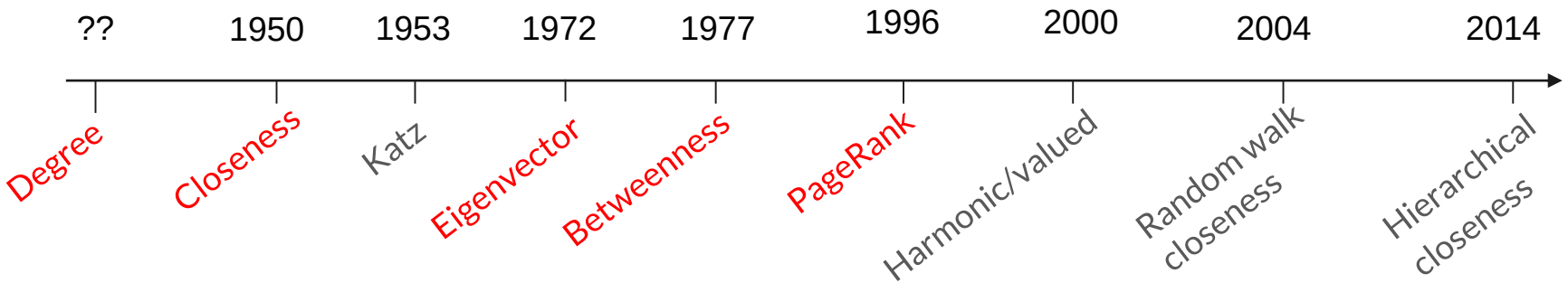
- Can find the centrality of weighted networks based on
  - **Geodesics:** shortest paths
  - **Paths:** vertex visited no more than once
  - **Trails:** edge traversed no more than once, vertices can be revisited
  - **Walks**: vertices and edges can be visited multiple times
- Depending on the type of flow, these different measures of centrality can be used

Characterization by Cohesiveness

- Ranks vertices by the number of walks from a given vertex
- **Radial:** count walks that start/end at a given vertex
  - Degree Centrality
  - Closeness Centrality
  - Eigenvector Centrality
  - PageRank
- **Medial :** count walks that pass through vertex
  - Betweenness Centrality
- **Volume:** total number of walks (all except Closeness)
- **Length:** distance of walks (Closeness)

History

Centrality Timeline

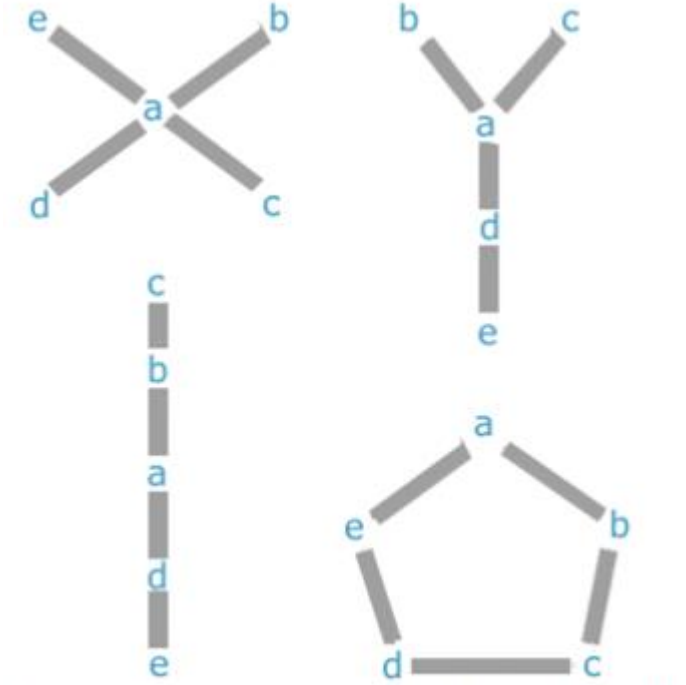


*Implemented these methods

*are many other centrality measures not listed...

Alex Bavelas (1920-)

- Developed the first formal measure of centrality: **closeness**
- Psychologist interested in the interaction between groups
    - Hypothesized the communication structure affects the performance of a group



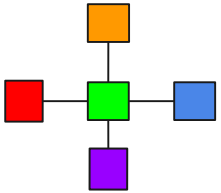
Four communicative dispositions of groups of five elements analyzed by Bavelas.

The Bavelas-Leavitt Experiment

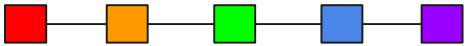
5 people play a game where they must solve a puzzle

- Each person has a unique bit of information
- The solution requires all the information to be pooled
- Every player must get the answer
- Communication is restricted

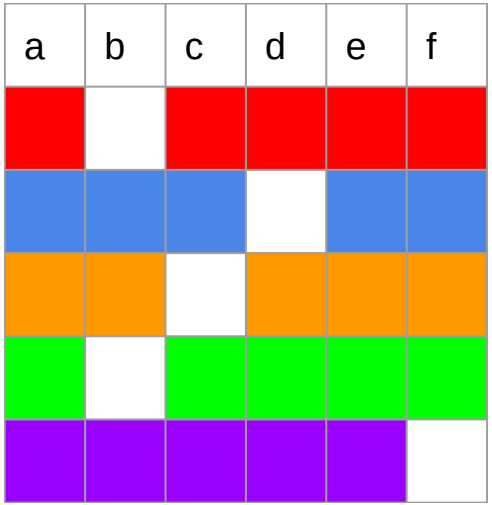
The Bavelas-Leavitt Experiment - Setup

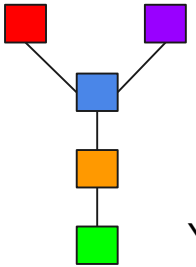


Star

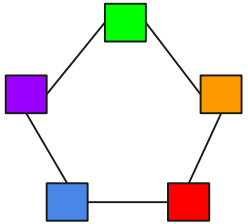


Line





Y



Circle

The Bavelas-Leavitt Experiment - Results

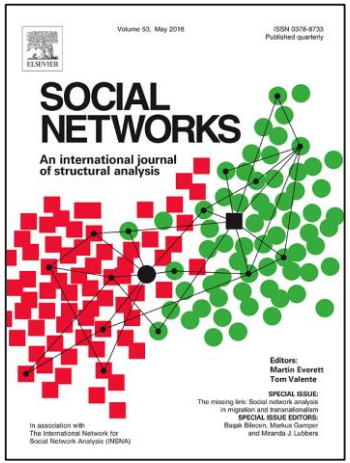
- **Time**. The star and Y were considerably faster, on average, than the line and circle.
- **Messages.** The star and Y used the least number of messages. The line was next, then the circle (which used quite a bit more).
- **Errors.** An error was defined as the throwing of an incorrect switch before the end of a game. The star, the Y and line made the fewest errors, while the circle made the most (however, the circle had the most error corrections).
- **Satisfaction**. The subjects in a the circle network enjoyed themselves the most, followed by the line, the Y and finally the star.

Conclusion:

According to Bavelas and Leavitt the more centralized a structure is, the better it performs.

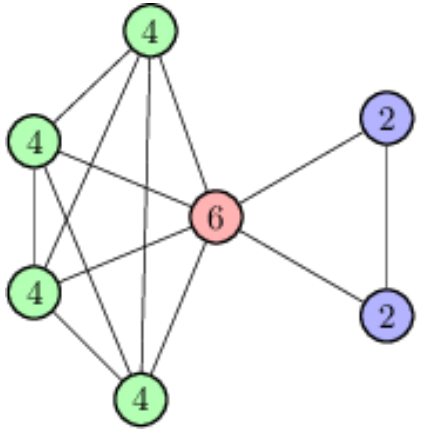
Linton Freeman

- University of California, Irvine
  - Researches social network analysis
- Compiled the previous literature on centrality into a formal mathematical framework
  - "Centrality in Social Networks: I. Conceptual Clarification." *Social Networks*, 1, 1979, 215-239.
  - "Centrality in Social Networks: II. Experimental Results." (L. C. Freeman, D. Roeder, R. Mulholland). *Social Networks*, 2, 1980, 119-142.



Algorithms

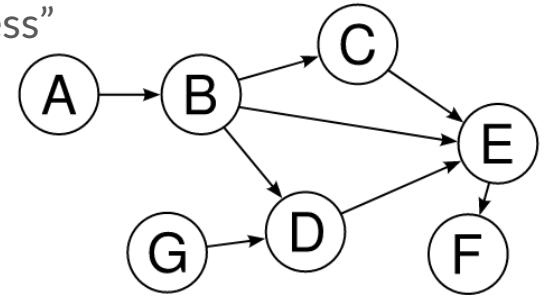
Degree Centrality



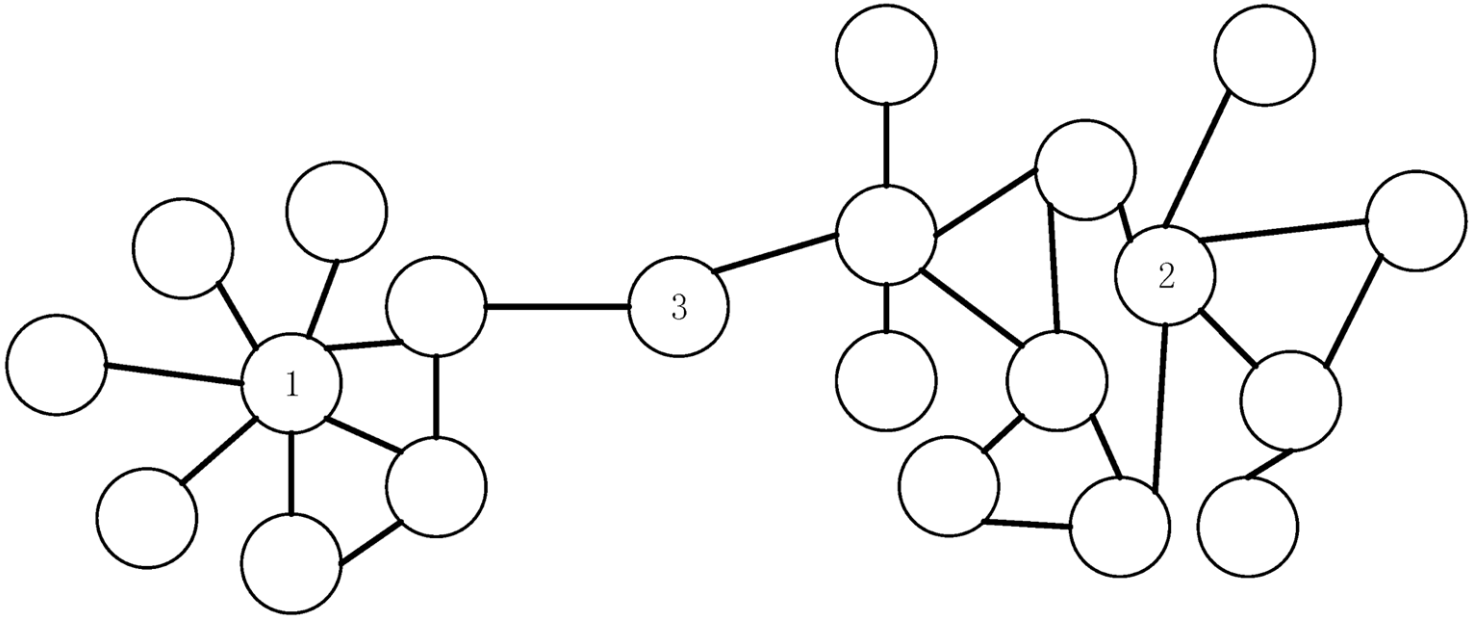
- Most conceptually simple centrality measure
- Simply rank vertices by the degree of the vertex
  - Highest degree vertex considered center
- Represents immediate risk of a node for catching whatever is flowing through the network (e.g. virus)
- O(V) time complexity

Degree Centrality of Directed Graphs

- **Indegree:** number of edges going into a vertex
- **Outdegree:** number of edges leaving a vertex
- We can also find degree centrality based on indegree or outdegree
  - Indegree centrality can be seen as “popularity”
  - Outdegree centrality can be seen as “gregariousness”



Limitations of Degree Centrality



Closeness Centrality

- Sum of the length of the shortest paths from a node to all others
  - More central → closer to other nodes
- Usually normalized
  - Take the average instead of the sum
  - Now nodes in graphs of different sizes are comparable
- Time complexity depends on the finding-shortest-paths subroutine...

Closeness - Mathematical Definition

From Bavelas' "Communication patterns in task-oriented groups" (1950)

$$C(x) = \frac{1}{\sum_y d(y,x)}$$

Reciprocal of sum of distances

$$C(x) = \frac{N-1}{\sum_y d(y,x)}$$

Normalized form

Closeness - Algorithm

1. Calculate the shortest paths between a vertex and all others
2. Sum and take the reciprocal
3. Repeat for all vertices

Closeness - Modern Modifications

What to do when a graph is disconnected?

Try the harmonic mean instead of the arithmetic mean and let $1/\infty = 0$

$$H(x) = \sum_{y \neq x} \frac{1}{d(y,x)}$$

Harmonic or “Valued” closeness

Betweenness Centrality

- Based on Bavelas' method
  - Every pair of vertices has some number of shortest paths
  - Betweenness counts the number of these paths that a vertex touches

"..degree to which a point falls on the shortest path between others and therefore has a potential for control of communication."

Linton C. Freeman, 1977, "A Set of Measures of Centrality Based on Betweenness"

Betweenness Centrality - Formalized

$$C(v_k) = \sum_i^N \sum_j^N \frac{g_{ij}(v_k)}{g_{ij}}$$

Example

- P1 and P3 have two geodesics
  - Then neither P2 or P4 have total control of their communication
  - But they do have some potential

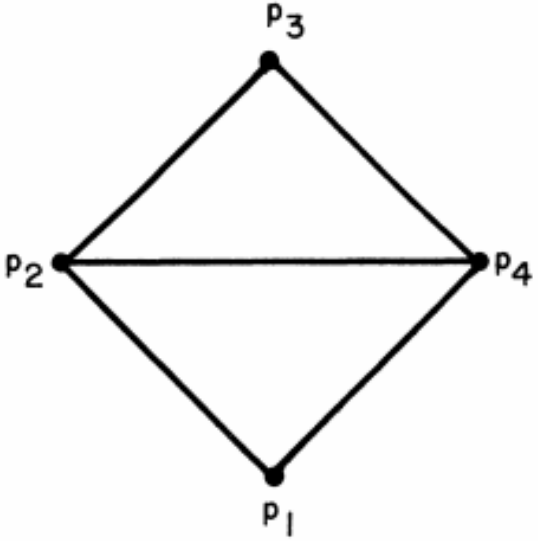


FIGURE 1
A Graph with Four Points and Five Edges

Linton C. Freeman, 1977, “A Set of Measures of Centrality Based on Betweenness”

Maximum Betweenness

When a point falls on all geodesics of length greater than one

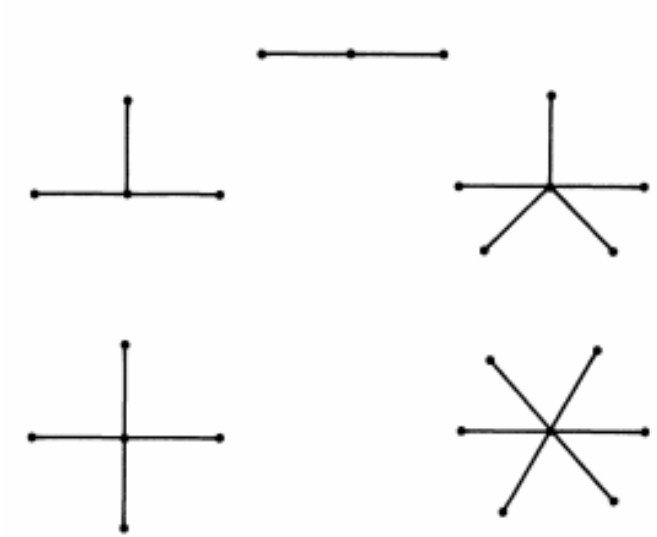


FIGURE 4
Graphs of Stars or Wheels for n = 3, 4, 5, 6, 7

Linton C. Freeman, 1977, “A Set of Measures of Centrality Based on Betweenness”

Betweenness - Algorithm

1. Compute the length and number of shortest paths between all pairs (s,t)
2. For each vertex v, calculate every possible pairwise dependency and sum them

2 dominates the time complexity O(n^3)

Augment your favorite shortest path finding algorithm (Floyd-Warshall, Dijkstra) to compute the pairwise dependencies as the paths are found

Eigenvector Centrality

- Ranks vertices based on the concept that high scoring vertices contribute more to the score of a vertex than low or equal scoring
- Important vertices are connected to important vertices
- The relative centrality score for vertex *v* is defined as

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

where *M(V)* is the set of vertices adjacent to *v*

Eigenvector Centrality Cont.

- However, this can be reduced to

$$A \times x = \lambda \cdot x$$

- This equals the eigenvector, eigenvalue pair of the adjacency matrix
- By the **Perron-Frobenius Theorem**, we can use the corresponding eigenvector for the largest eigenvalue to get the centrality of the graph
- Typically use **Power Iteration** to solve the eigenvalue problem
  - $O(V^3)$

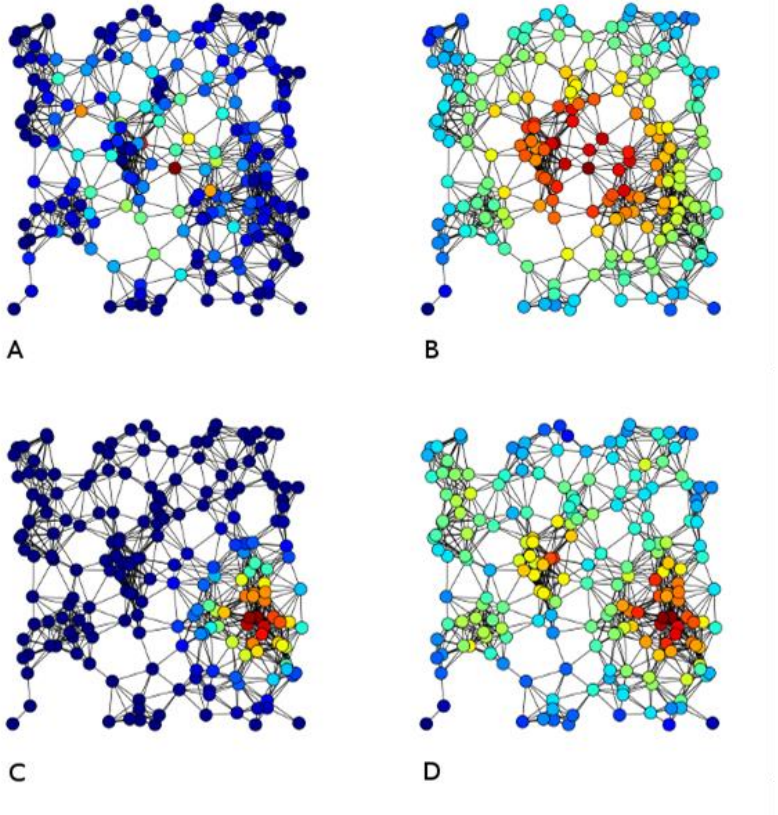
```python
def power_iteration(A, num_simulations):
    # Ideally choose a random vector
    # To decrease the chance that our vector
    # Is orthogonal to the eigenvector
    b_k = np.random.rand(A.shape[0])

    for _ in range(num_simulations):
        # calculate the matrix-by-vector product Ab
        b_k1 = np.dot(A, b_k)

        # calculate the norm
        b_k1_norm = np.linalg.norm(b_k1)

        # re normalize the vector
        b_k = b_k1 / b_k1_norm
```

Wikipedia Power Method code



(A) Betweenness (B) Closeness (C) Eigenvector (D) Degree

Image from Wikipedia

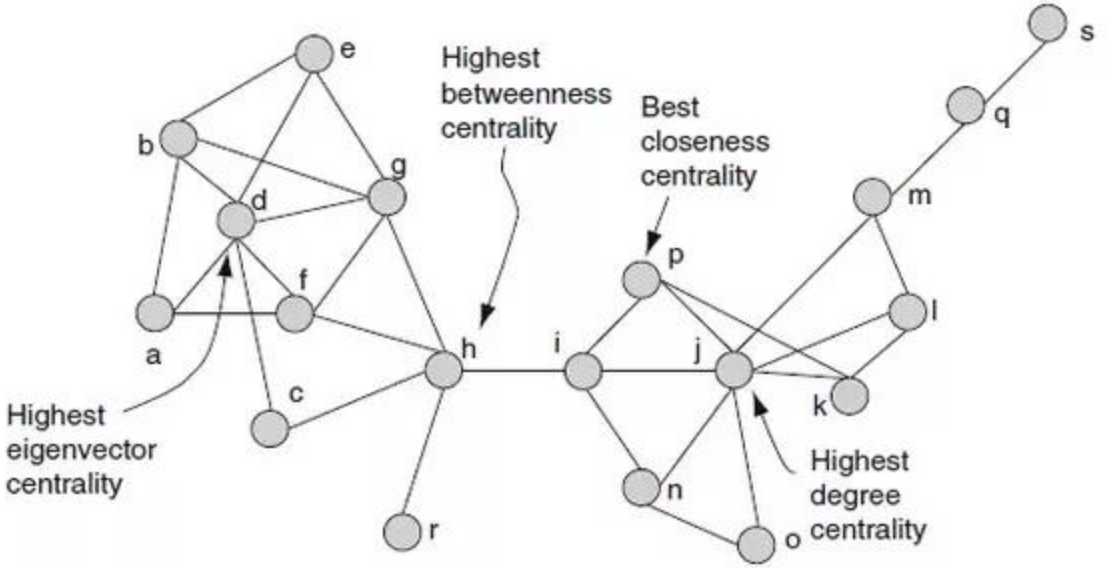


Image from https://www.quora.com/What-are-the-limitations-of-graph-centrality-measures

PageRank

- Modified version of eigenvector centrality applied to the internet
- **INPUT** a directed web-graph
  - Nodes: webpages
  - Edges: hyperlinks
    - Link on a page to another (outbound) is a forward edge
    - Link off a page to it (inbound) is a backward edge
- **OUTPUT** a probability distribution that a person randomly clicking links arrives at any page
- Developed by Larry Page, Sergey Brin, and others at Stanford University in 1996

“The intuition behind PageRank is that it uses information which is external to the Web pages themselves - their backlinks, which provide a kind of peer review.”

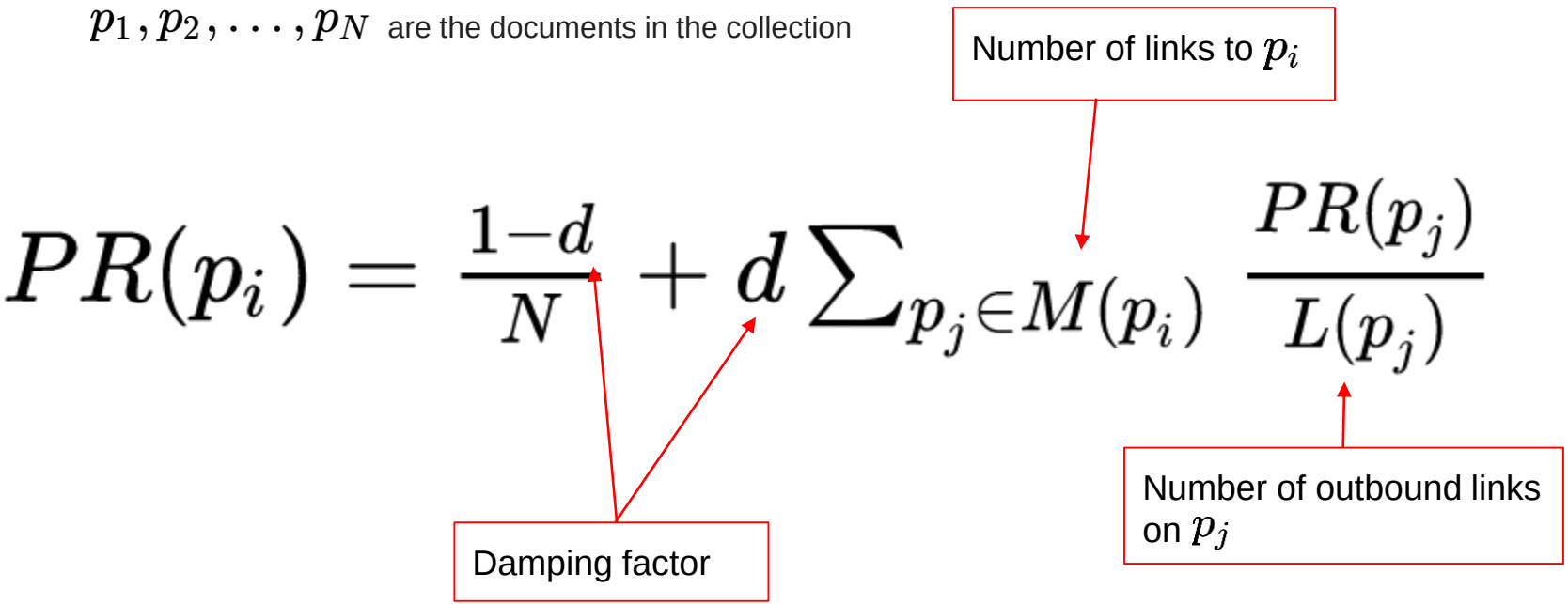
Larry Page & Sergey Brin

PageRank - Some Terms

- Damping factor (d) - probability the websurfer stops clicking links and chooses a new **random** page
  - Empirically determined to be ~ 0.85
- Document - a webpage
- Collection - all the N documents
- Sink - a page that has no outbound links
  - A new random page is chosen when the websurfer gets stuck in a sink

PageRank - Construction

$p_1, p_2, \ldots, p_N$ are the documents in the collection



PageRank - Calculation

- Where to start?
  - PR of a page depends on the PR of the pages that link to it..
  - Those pages depend on others…
  - Cycles of links exist too!
- Instead take a guess, evaluate and iterate from there!
  - Power Iteration method

“*PageRank or PR(A) can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the web.*”

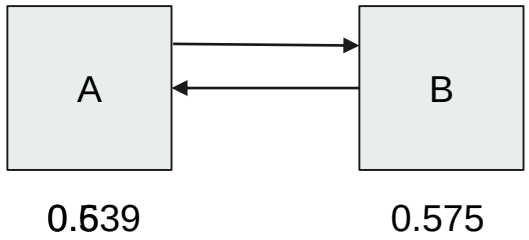
Simple Example

Step 0
Initialize PR(A) to 1/N = 0.5, d=0.85

Step 1
PR(B) = 1-d + d*PR(A)/1
= 1-0.85 + 0.85*(0.5)/1
= 0.575

Step 2
PR(A) = 0.15 + 0.85(0.575)/1
= 0.639...

...Step ~20
Converges at PR(A) = PR(B) = 1.0



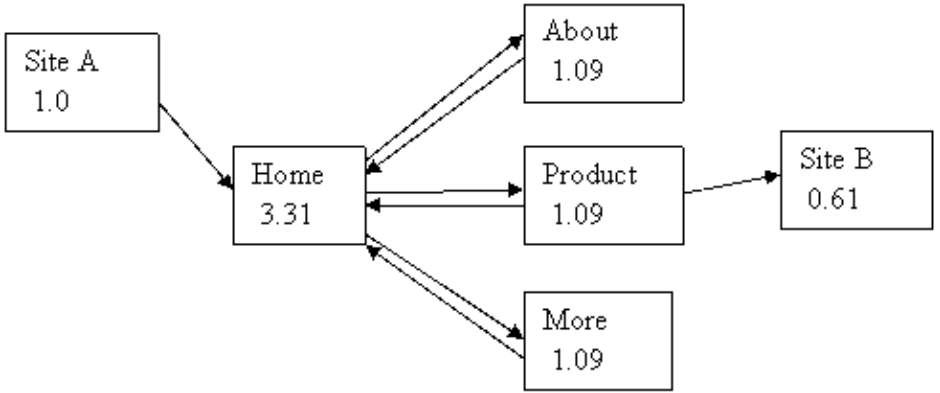
Convergence: the PR of pages doesn't change after an iteration

A typical website

Hierarchical, but with links to external sites

The links back from the subdirectories of the website boost the PR of the home page

A well structured website will amplify the PR of the home page

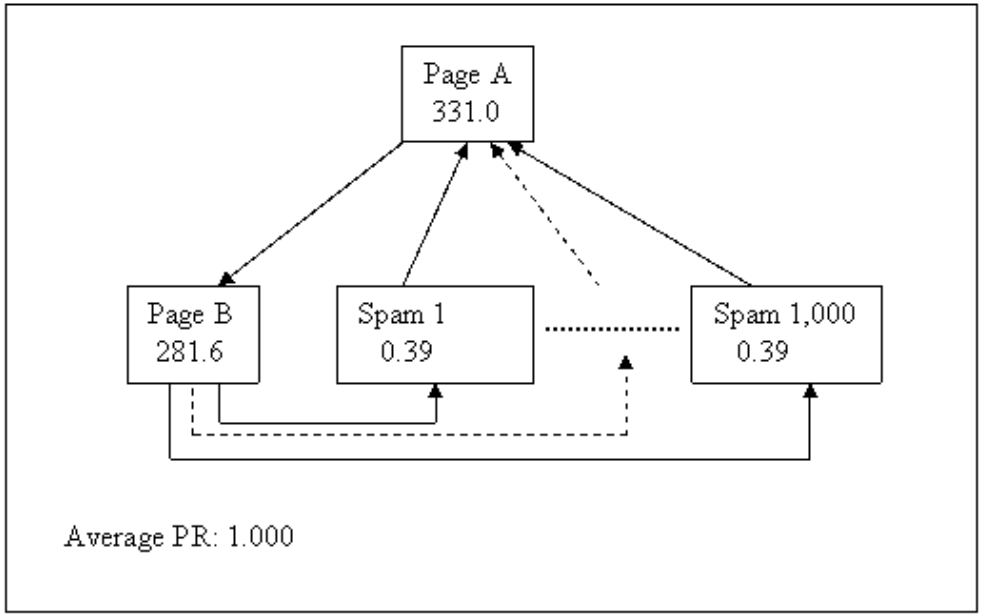


A disreputable website

Only one link leaving the home page, but 1,000 spam pages link back.

Common practice by adult websites to inflate search results

Hasn’t been effective for years… Google caught on



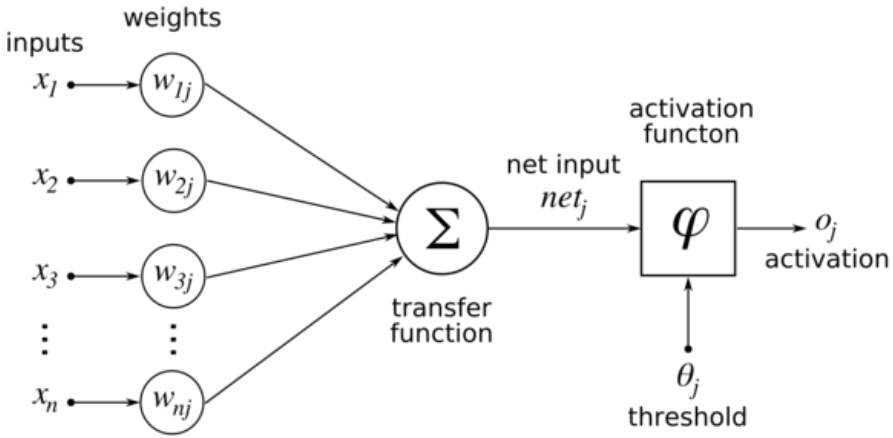
Applications

General Applications

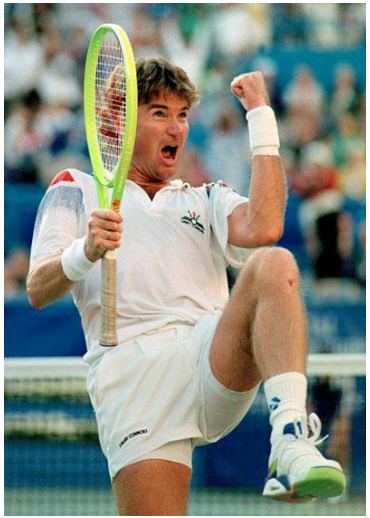
- Finding most susceptible patients to a disease
- Exploring how information flows through a social network
- Analyzing network traffic
- Computational diagnostics in chemistry
    - Graph of pressure, temperature, and species concentrations to detect local chemistry

Eigenvector Centrality Application

- Eigenvector centrality of a neuron in a neural network correlates with relative firing rate.



PageRank Applications



- Ever heard of Google?
- Find best athletes using graphs of tennis players
  - Jimmy Connors ranked best
- Literature
  - Jane Austen and Walter Scott found to be most original authors of 19th century

See reference 5 for a complete list of fun PageRank applications

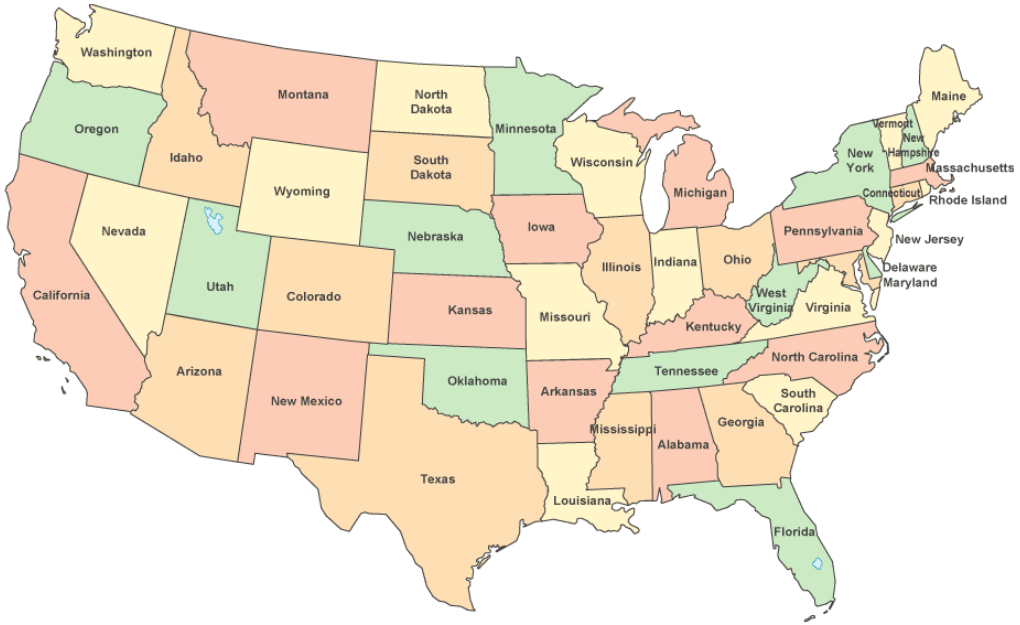
Implementations

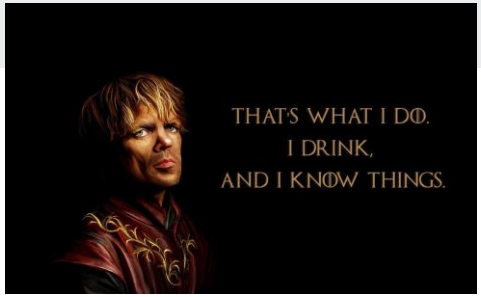
General information

- Python 2.7
  - **networkx** provides easy storage of graph information
  - **matplotlib** to create graphs
- Generated random graphs, varying number of vertices
  - Sparse and Dense
  - Directed and Undirected

Contiguous USA (Unweighted Undirected)

Indegree	TN/MO (8)
Closeness	MO
Betweenness	MO
Eigenvector	MO
PageRank	TN





Game of Thrones (Weighted Undirected)

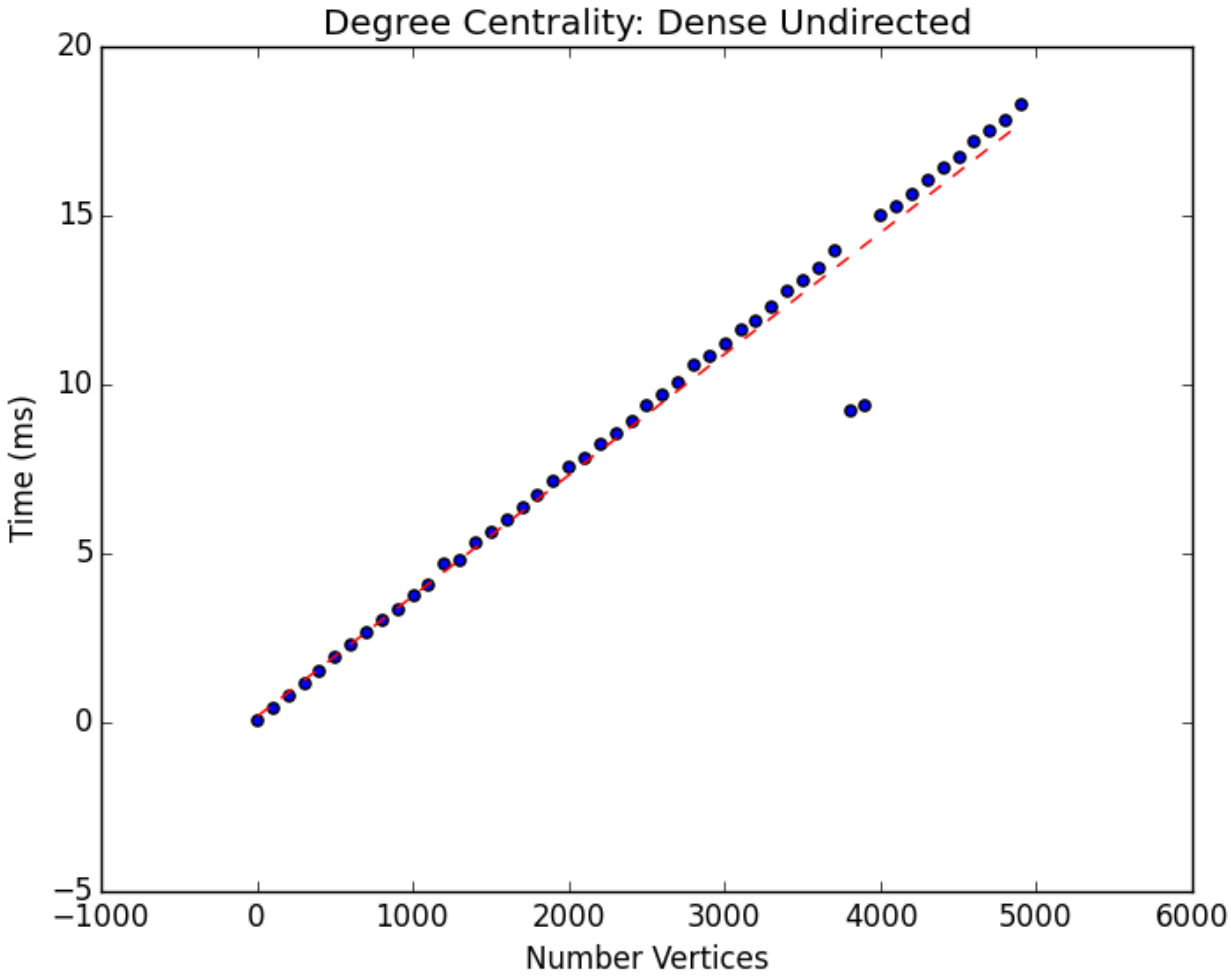
- **Nodes:** characters
- **Edges:** weighted by the number of times the two characters' names appeared within 15 words of each other in the text (see citation 7)

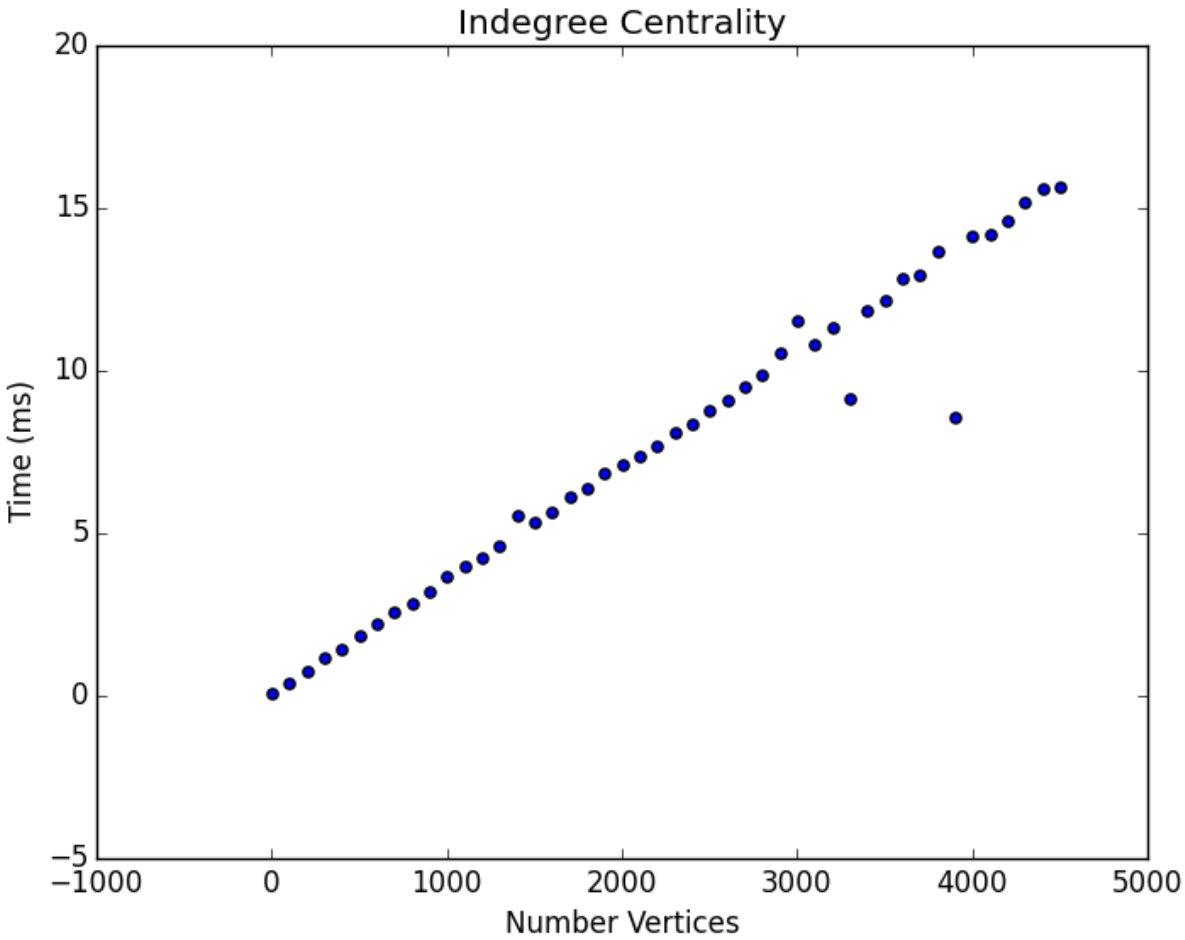
Degree	Tyrion
Closeness	Tyrion
Betweenness	Robert Baratheon
Eigenvector	Tyrion
PageRank	Tyrion

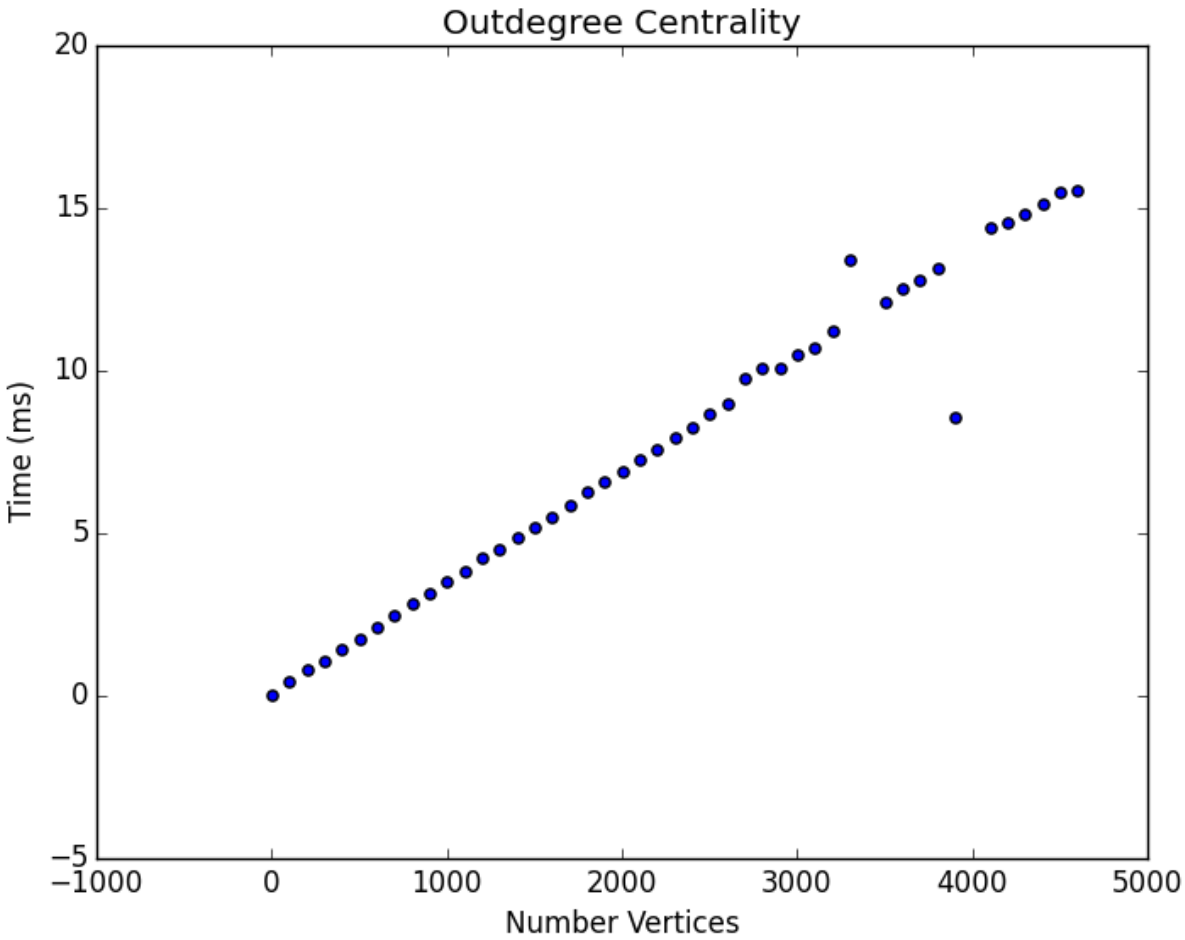
NFL 2009 Season (Weighted Directed)

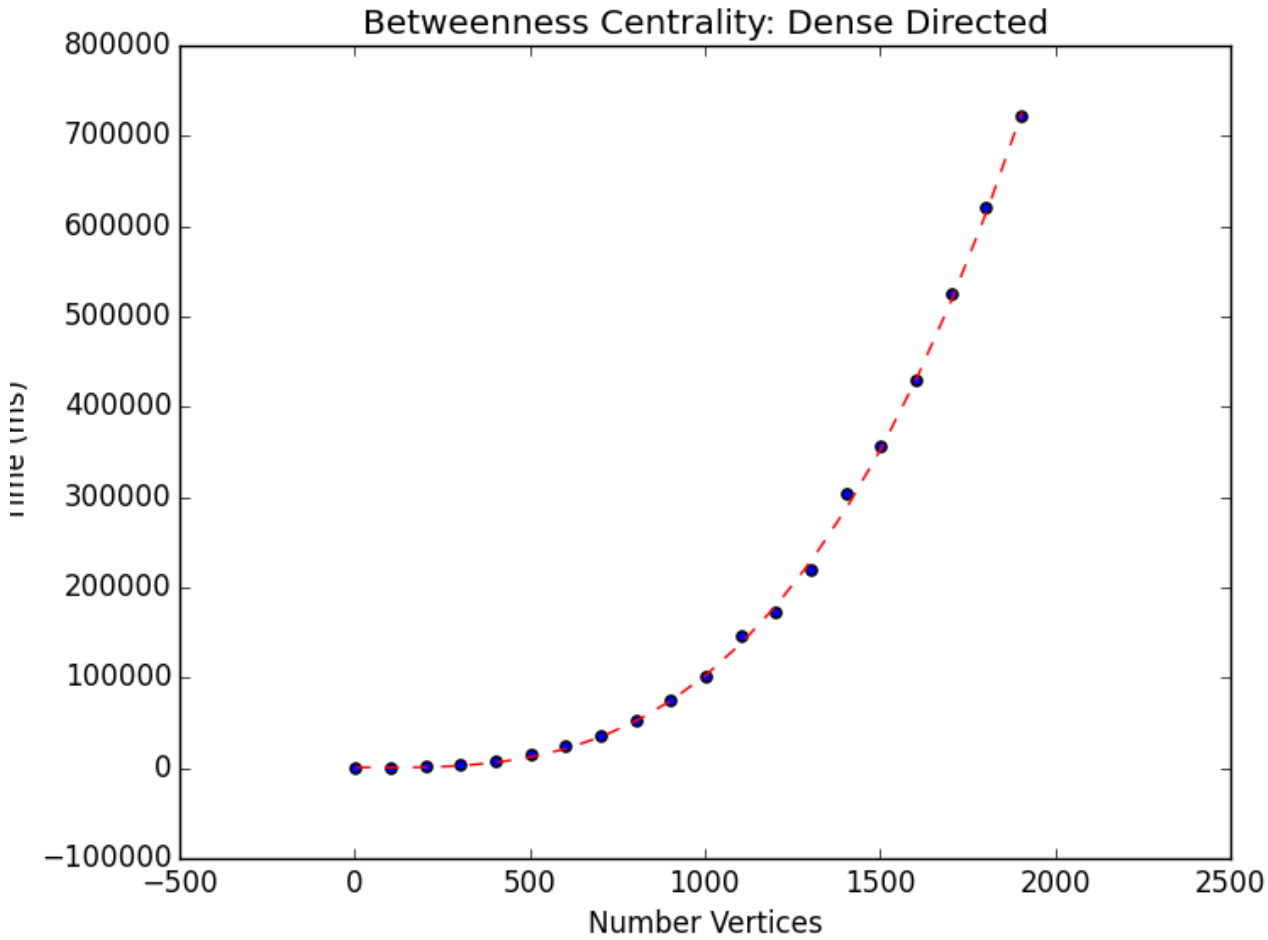
- **Nodes:** teams
- **Edges:** score difference (winning team directed to losing team)
- New Orleans Saints Super Bowl Champ

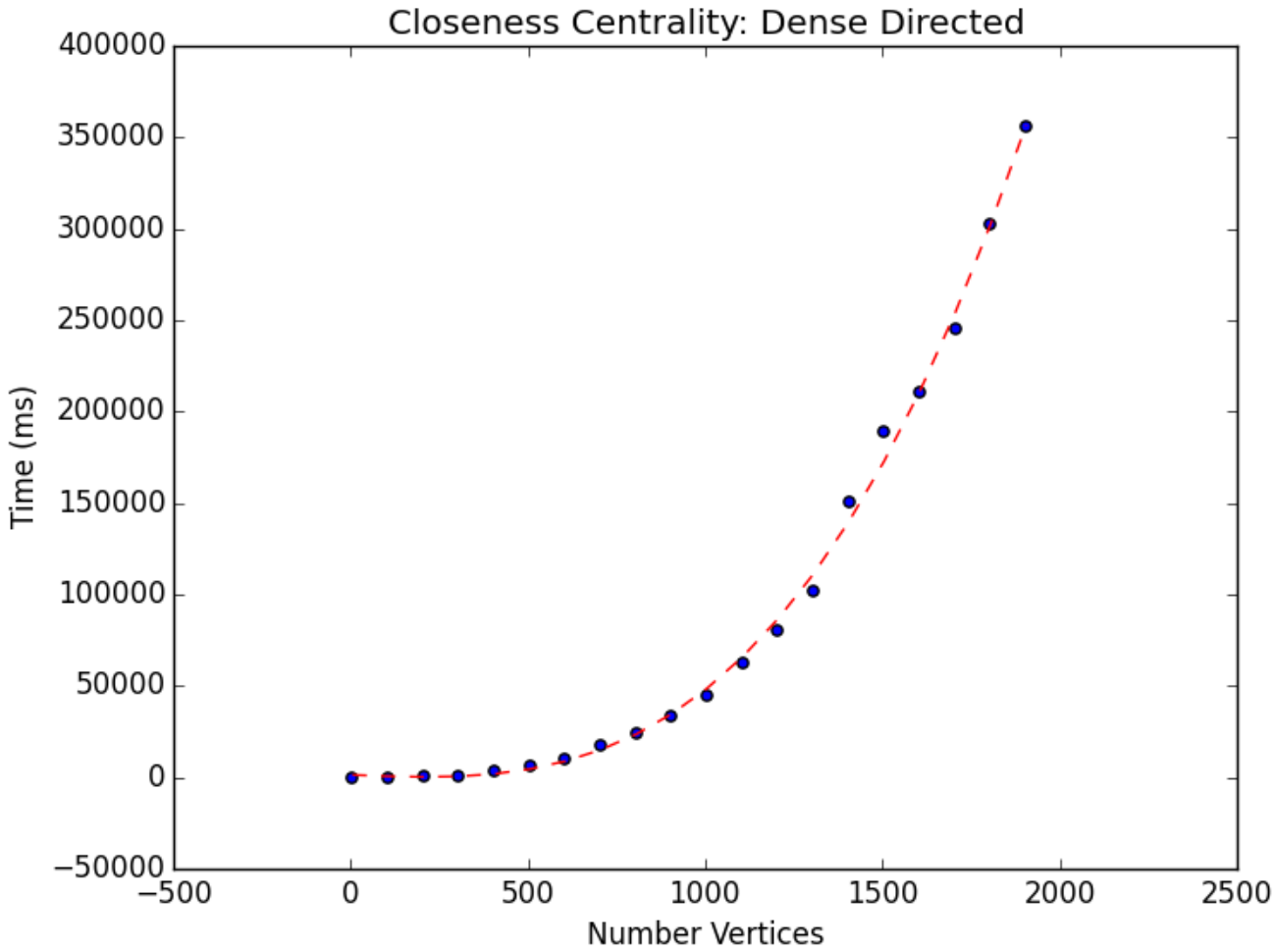
Indegree	St. Louis Rams
Outdegree	New Orleans Saints
Closeness	Dallas Cowboys
Betweenness	Tennessee Titans
Eigenvector	Oakland Raiders
PageRank	Oakland Raiders

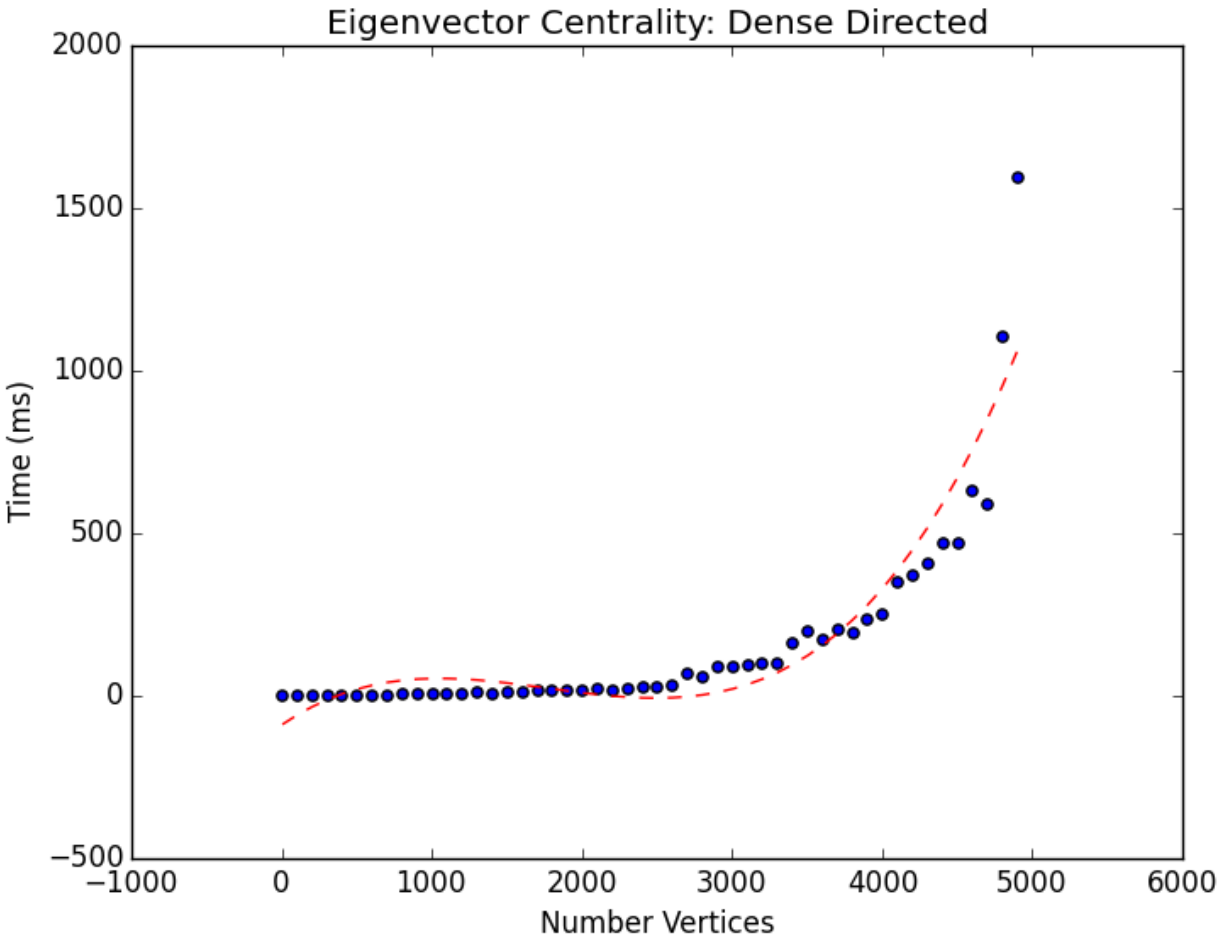

Degree Centrality: Dense Undirected
Time (ms)
Number Vertices
20
15
10
5
0
−5
−1000
0
1000
2000
3000
4000
5000
6000

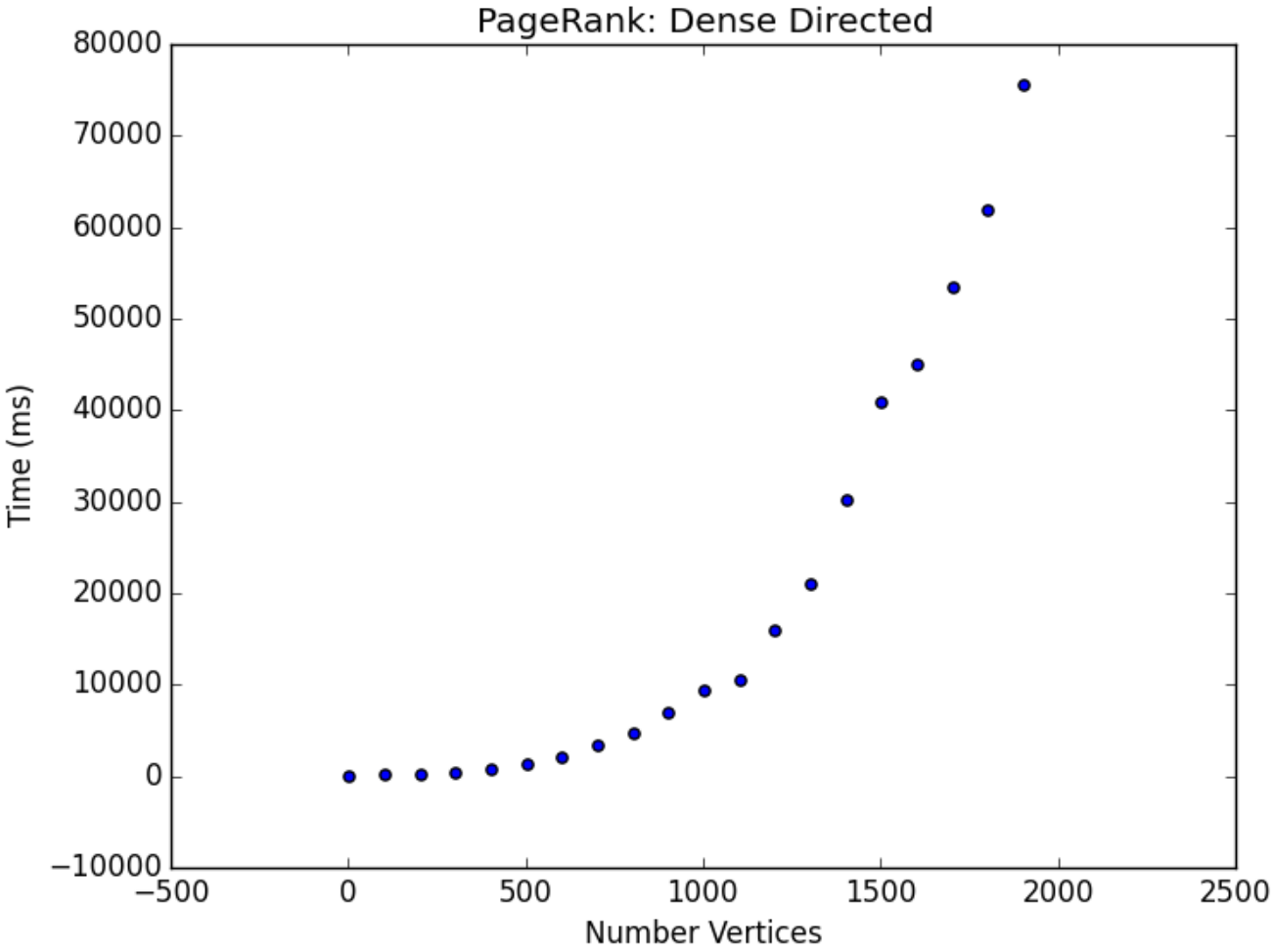

Indegree Centrality
Time (ms)
Number Vertices
20
15
10
5
0
−5
−1000
0
1000
2000
3000
4000
5000

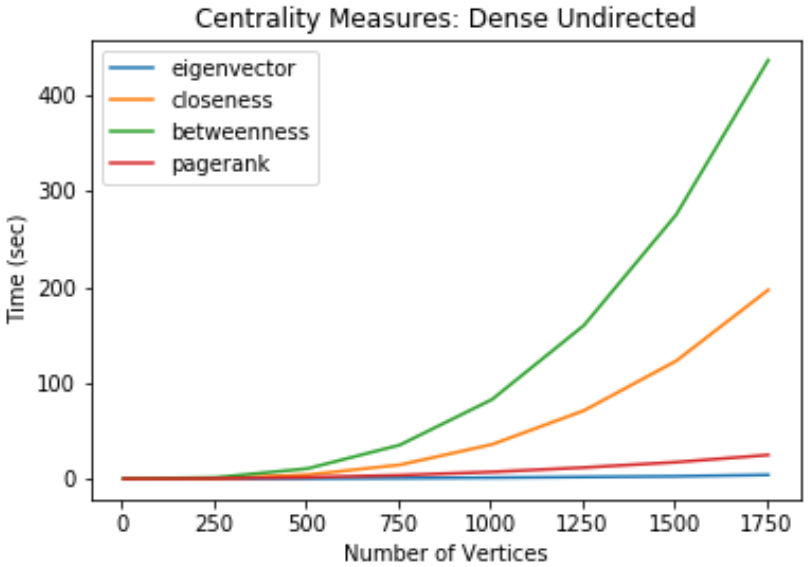

Outdegree Centrality
Time (ms)
Number Vertices
20
15
10
5
0
−5
−1000
0
1000
2000
3000
4000
5000

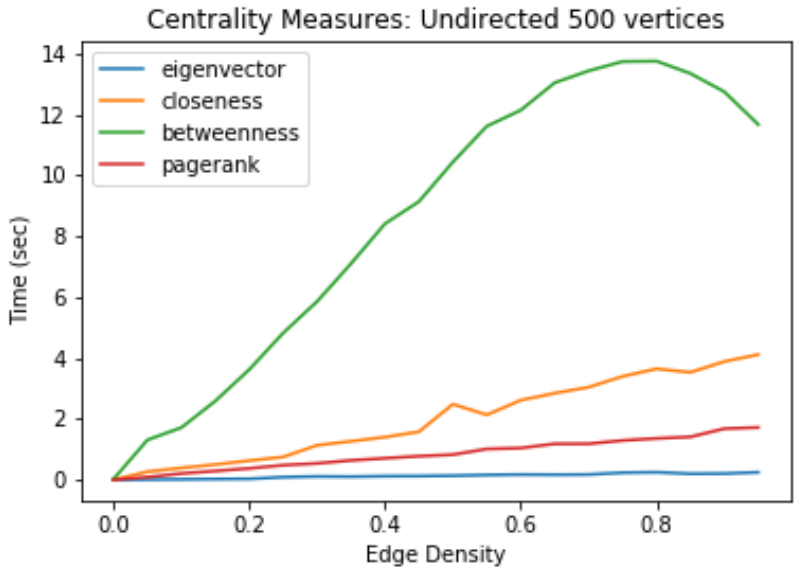

Betweenness Centrality: Dense Directed
800000
700000
600000
500000
400000
300000
200000
100000
0
−100000
−500
0
500
1000
1500
2000
2500
Number Vertices


Closeness Centrality: Dense Directed
Time (ms)
Number Vertices
400000
350000
300000
250000
200000
150000
100000
50000
0
−50000
−500
0
500
1000
1500
2000
2500


Eigenvector Centrality: Dense Directed
Time (ms)
Number Vertices
2000
1500
1000
500
0
−500
−1000
0
1000
2000
3000
4000
5000
6000


PageRank: Dense Directed
Time (ms)
Number Vertices
80000
70000
60000
50000
40000
30000
20000
10000
0
−10000
−500
0
500
1000
1500
2000
2500





Comparison of standard techniques

Open Issues

Open Issues / Limitations

- Centrality of “big” graphs
  - Are there more efficient approaches to centrality for large graphs?
  - Accurate approximations?
- Most centrality rankings are not useful after the first few ranks
  - Only the most important vertices are correctly recognized
  - The rest of the vertices relative importance is not learned
- Identifying combinations of nodes that are central
- Centrality of dynamic graphs
  - Are there methods faster than recomputation?

References

[1] Page, L., Brin, S., Montwani, R., and Winograd, T. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999.
[2] Borgatti, Stephen P. (2005). "Centrality and Network Flow". Social Networks. Elsevier. 27: 55–71.
[3] Borgatti, Stephen P.; Everett, Martin G. (2006). "A Graph-Theoretic Perspective on Centrality". Social Networks. Elsevier. 28: 466–484.
[4] Bisht, Jyant. “Degree Centrality” https://www.geeksforgeeks.org/degree-centrality-centrality-measure/
[5]https://blogs.cornell.edu/info2040/2014/11/03/more-than-just-a-web-search-algorithm-googles-pagerank-in-non-internet-contexts/
[6] Radicchi F (2011) Who Is the Best Player Ever? A Complex Network Analysis of the History of Professional Tennis. PLoS ONE 6(2): e17249. https://doi.org/10.1371/journal.pone.0017249
[7] A. Beveridge and J. Shan, "Network of Thrones." Math Horizons 23(4), 18-22 (2016) http://www.maa.org/sites/default/files/pdf/Mathhorizons/NetworkofThrones%20%281%29.pdf
[8] Francesco Bonchi, Gianmarco De Francisci Morales, and Matteo Riondato. 2016. Centrality Measures on Big Graphs: Exact, Approximated, and Distributed Algorithms. In Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 1017-1020. DOI: https://doi.org/10.1145/2872518.2891063

Questions Revisited

1. What are the two ways we can define importance in terms of a graph?

2. What is the output of PageRank?

3. What states are considered the center of the contiguous USA based on degree centrality?

Discussion!