Lab 8: Measuring Graph Centrality - PageRank

Monday, November 5

CompSci 531, Fall 2018

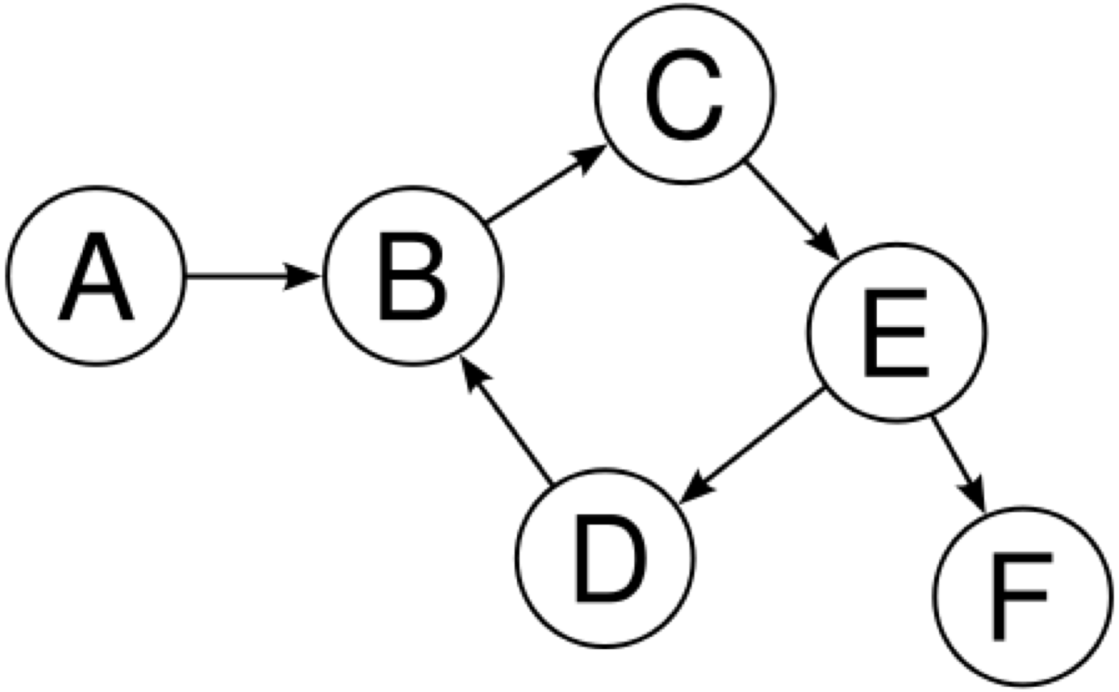
Outline

- Measuring Graph Centrality: Motivation

- Random Walks, Markov Chains, and Stationarity Distributions

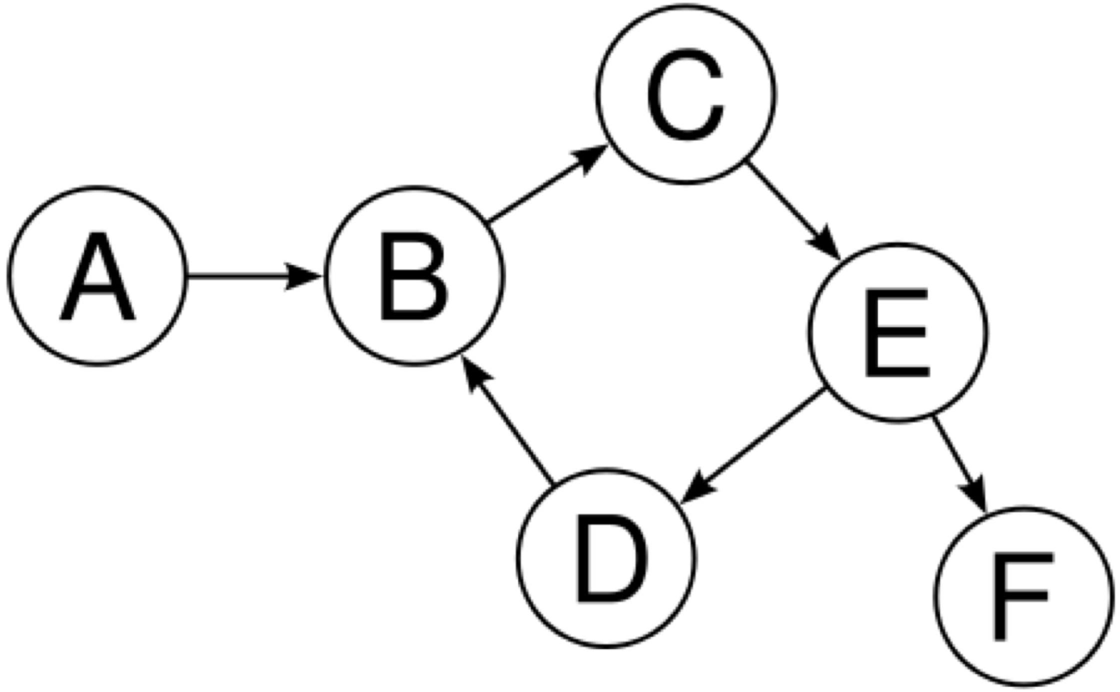
- Google’s PageRank Algorithm

Directed Graphs



	A	B	C	D	E	F
A	0	1	0	0	0	0
B	0	0	1	0	0	0
C	0	0	0	0	1	0
D	0	1	0	0	0	0
E	0	0	0	1	0	1
F	0	0	0	0	0	0

Graph Centrality



- Which vertex is "the most important" in this graph?
- What do we even mean by important?
- In this class, we will focus on importance as *centrality* as measured by a random walk.

Motivation – Social Media

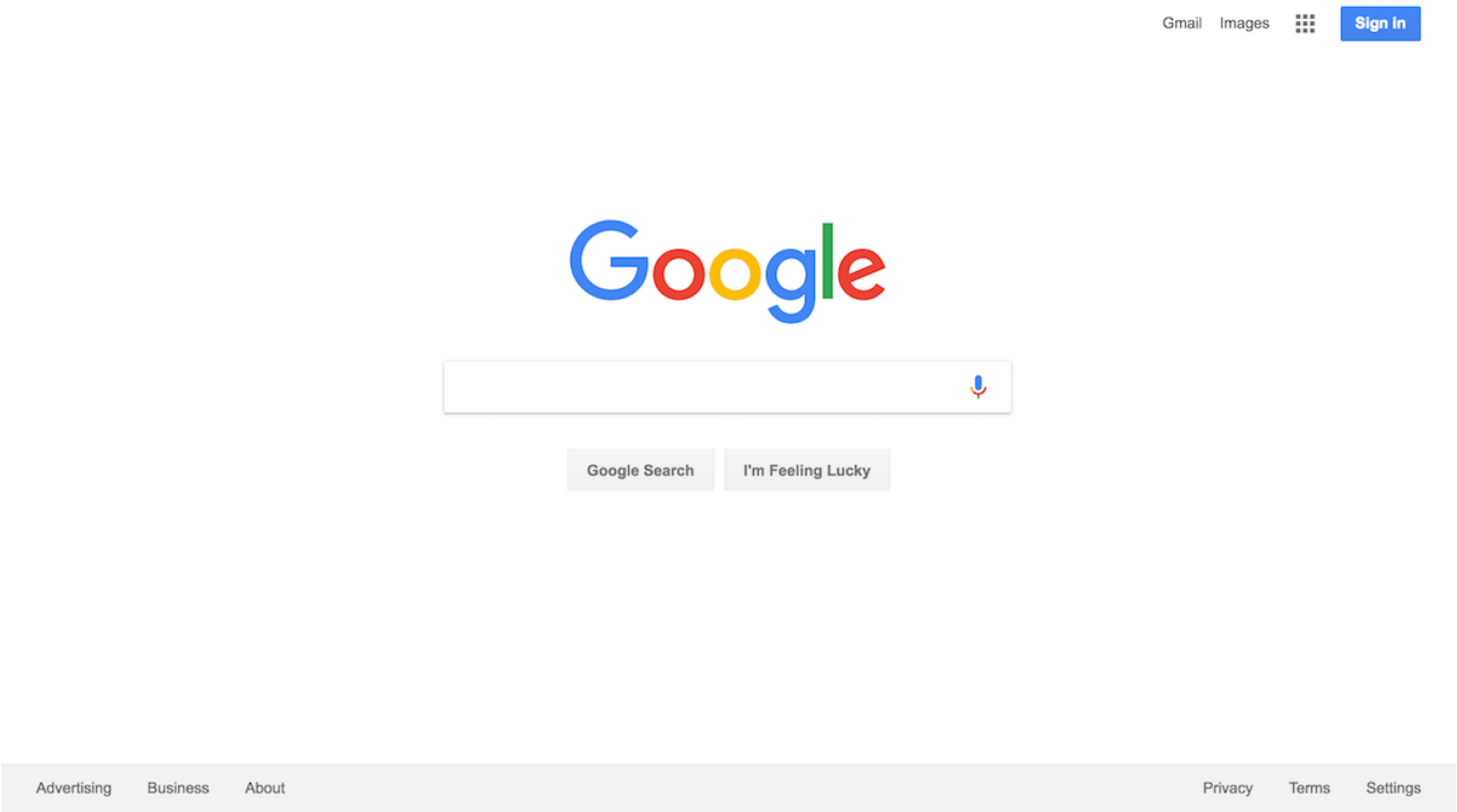
- Who is “important” in the Twitter network?

Motivation – Academic Publishing

- How impactful is a scientific publication?

Motivation – Web Search

- Which webpages are most important for displaying after a search query? (The original motivation).

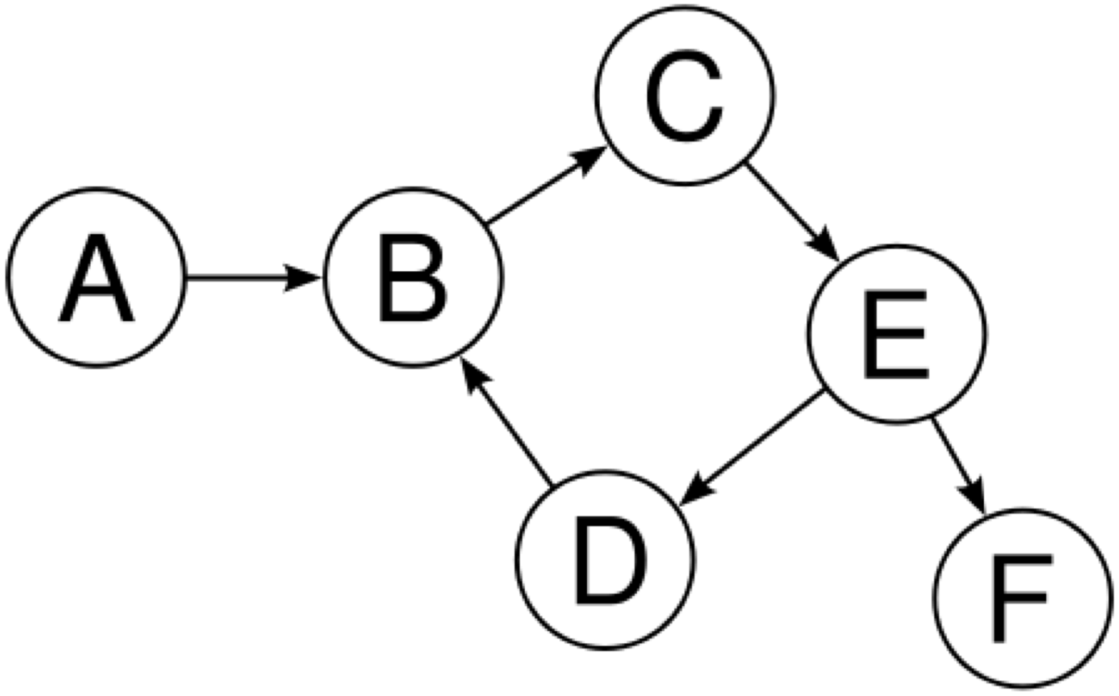


Outline

- ~~Measuring Graph Centrality: Motivation~~
- Random Walks, Markov Chains, and Stationarity Distributions
- Google's PageRank Algorithm

Formalizing “Graph Centrality”

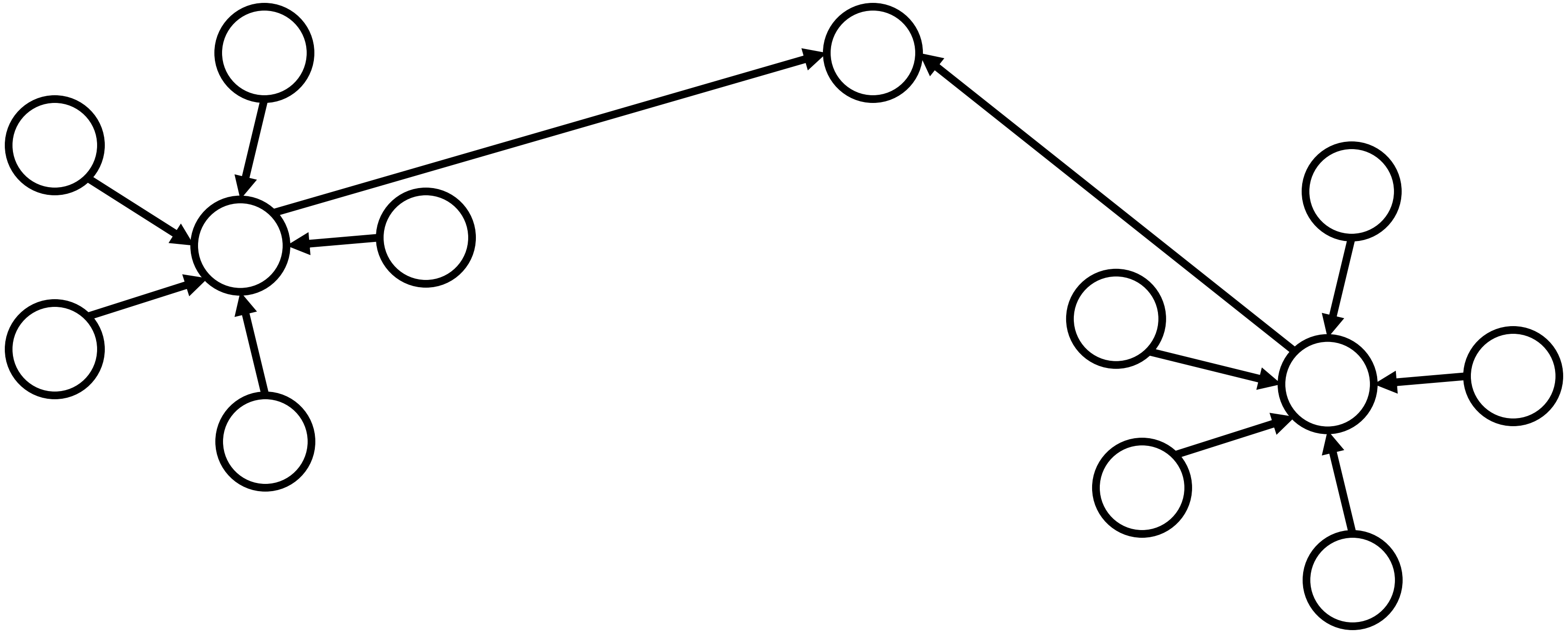
- **Attempt 1.** Measure the *in-degree* (number of incoming directed edges) of every node.



Node	in-degree
A	0
B	2
C	1
D	1
E	1
F	1

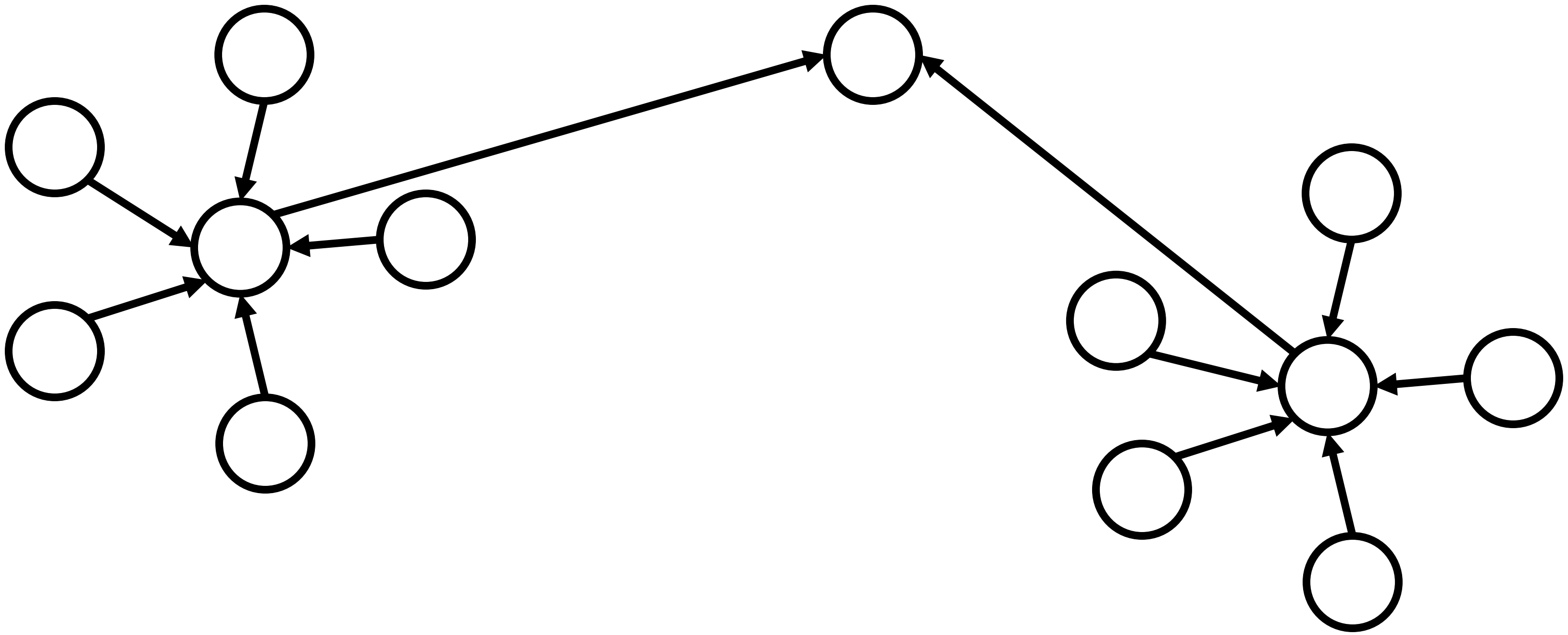
Formalizing Graph Centrality

- **Problem.** Why do edges from unimportant and important nodes contribute equally?
- What is the most important and central vertex in this graph?



Formalizing Graph Centrality

- **Attempt 2.** Say that a node is “central” in so far as we are likely to arrive at the node while traversing the graph.
- For example, in this graph, *all* traversals end at the same place.



Random Walk

- **Question.** What do we mean by "likely" in a traversal? Where is the probability coming from?
- **Answer.** We consider a *random walk*.

- Start at a random vertex
- For t from 1 to T steps:
  - Choose an outgoing edge uniformly at random and follow it

- Let π_i^t be the probability that we are at node i at time t. Then the centrality of node i is $\lim_{t\to\infty} \pi_i^t$.

Transition Probabilities

- Note that $\overrightarrow{\pi^{t+1}}$ only depends on $\overrightarrow{\pi^{t}}$. In particular, let d_i denote the out-degree of vertex i. Then

$$\pi_j^{t+1} = \sum_{i:(i,j)\in E} \frac{\pi_i^t}{d_i}.$$

- For convenience, let P be the transition matrix defined below. For now, assume that $d_i \geq 1$ for all i.

$$P_{ij} = \begin{cases} \dfrac{1}{d_i}, & A_{ij} = 1 \\ 0, & A_{ij} = 0 \end{cases}$$

Markov Chain

- Each row represents a conditional probability distribution: we can interpret P_{ij} as the probability that we move to j given we are at i.
- We can rewrite the updates in terms of the transition matrix.

$$\overrightarrow{\pi^{t+1}} = \overrightarrow{\pi^{t}}\, P$$

- Note that $\overrightarrow{\pi^{t+1}}$ is independent the history, conditional on $\overrightarrow{\pi^{t}}$, i.e.,

$$\left(\overrightarrow{\pi^{t+1}} \,\middle|\, \overrightarrow{\pi^{1}}, \overrightarrow{\pi^{2}}, \dots, \overrightarrow{\pi^{t}}\right) = \left(\overrightarrow{\pi^{t+1}} \,\middle|\, \overrightarrow{\pi^{t}}\right).$$

- Thus, this random walk is a *Markov Chain.*

Stationary Distribution

- $\lim_{t\to\infty} \overrightarrow{\pi^t}$, our measure of graph centrality, is the *stationary distribution* of the Markov chain.

Questions.

1. Does the limit even exist?
2. Does the limit depend on the starting state $\overrightarrow{\pi^1}$?
3. Can we compute $\lim_{t\to\infty} \overrightarrow{\pi^t}$ efficiently?

Existence and Uniqueness

- Note that if $\lim_{t\to\infty} \overrightarrow{\pi^t}$ exists, then it must be some $\overrightarrow{\pi^*}$ such that
$$\overrightarrow{\pi^*} = \overrightarrow{\pi^*}\, P \to P^T\, \overrightarrow{\pi^*} = \overrightarrow{\pi^*}\,.$$
- That is, the stationary distribution $\overrightarrow{\pi^*}$ should be an *eigenvector* of the transposed transition matrix P^T, with eigenvalue 1.
  - (More to come next class on eigenvalues in graphs).

- Is it the only one? We need a theorem from linear algebra. Suppose for a moment that P has all strictly positive values.

Existence and Uniqueness

- **Perron-Frobenius Theorem** (abbreviated). Let A be a square matrix with real, strictly positive entries. Then the following hold.
  1. The largest eigenvalue (call it λ_1) of A is unique.
  2. There is a *unique* eigenvector (call it $\overrightarrow{v^*}$) corresponding to λ_1, all entries of which are positive, and this is the *only* eigenvector with all positive entries.
  3. The power iteration method that repeatedly applies $\overrightarrow{v^{t+1}} = A\overrightarrow{v^t}$ beginning from an initial vector $\overrightarrow{v^1}$ not orthogonal to $\overrightarrow{v^*}$ converges to $\overrightarrow{v^*}$ as $t \to \infty$.

- Every row of P is a probability distribution, so $P\,\vec{1} = \vec{1}$.
- By conditions 2 and 1, it must be that the largest eigenvalue of P is 1.
- Since P is square, P and P^T have the same eigenvalues, so 1 is the largest eigenvalue of P^T too!

Existence and Uniqueness

- Since 1 is the largest eigenvalue of P^T, the theorem implies that $\overrightarrow{\pi^*}$ exists and is the *unique* eigenvector of P^T with all positive entries.

- So we have answered questions 1 and 2: the stationary distribution exists, and it is unique.

- What about computation? The theorem tells us that the power iteration method converges in the limit…but how long does that take?

Computation

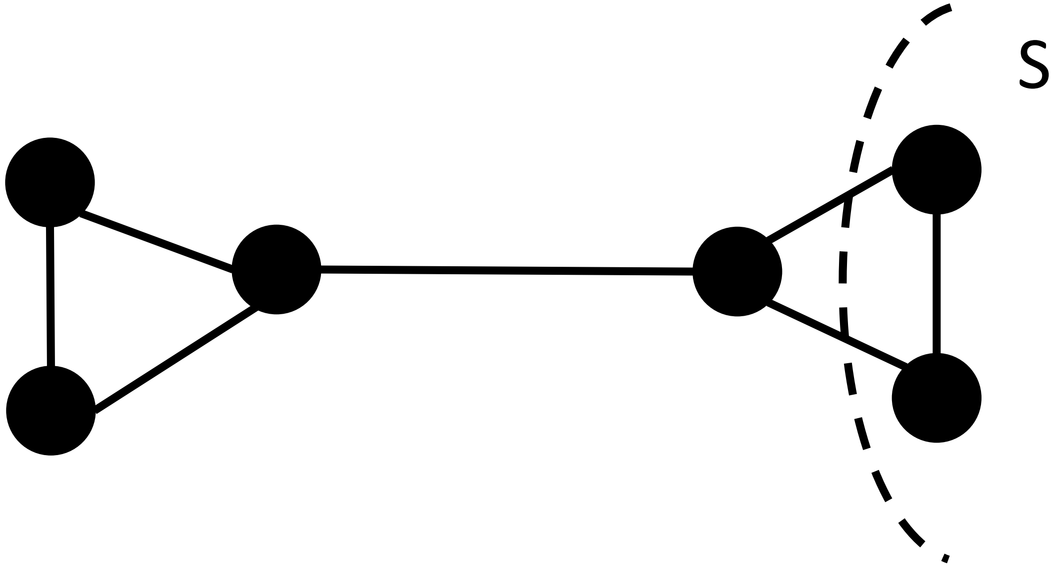
- In general, the convergence rate is determined by the *spectral gap*. If $\lambda_1 = 1$ is the largest eigenvalue of P^T, and λ_2 is the second largest eigenvalue of P^T, then the spectral gap is $\lambda_1 - \lambda_2$.

- As we will see next lab, the spectral gap is in turn related to the *conductance* of the underlying graph.

- Let $S \subseteq V$ be a cut in $G = (V, E)$. The *conductance* of the cut is $\phi(S) = \frac{|\{(i,j) \in E : i \in S, j \notin S\}|}{\min(\sum_{i \in S} d_i,\ \sum_{i \notin S} d_i)}$.

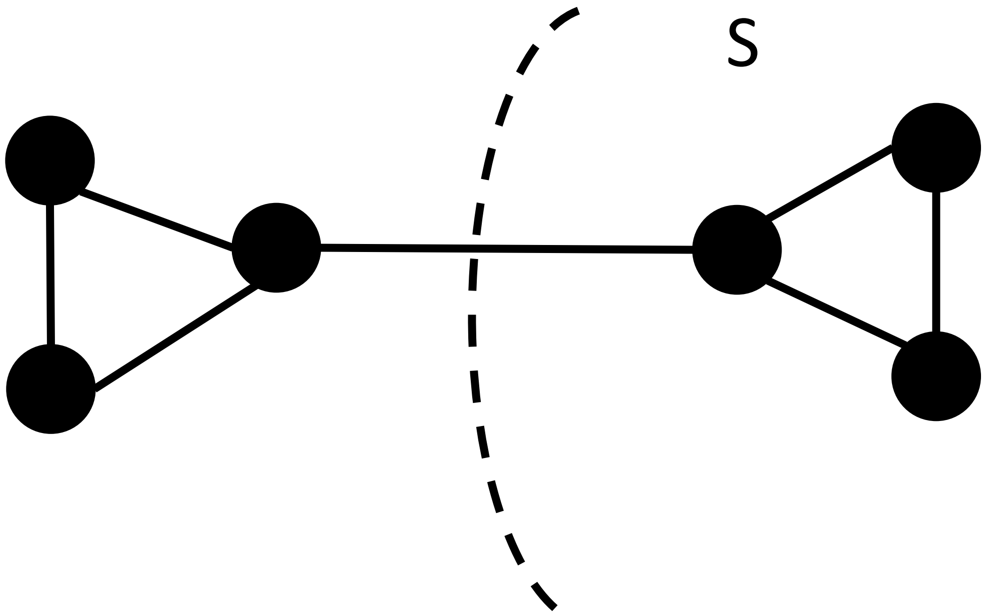
Computation

- The conductance of a graph is the minimum conductance of any cut.



$$\phi(S) = \frac{2}{\min(10, 4)} = \frac{1}{2}$$

Computation



$$\phi(S) = \frac{1}{\min(7,7)} = \frac{1}{7}$$

Computation

- So intuitively, lower conductance graphs have bottlenecks, and it may take a longer time for the random walk to traverse the cut.

- By contrast, power iteration converges rapidly on graphs with high conductance (e.g., complete graphs).

- To converge (to within some constant error term), one needs $O\left(\frac{\log(n)}{\phi^2}\right)$ iterations. What does that look like in practice?

Outline

- ~~Measuring Graph Centrality: Motivation~~

- ~~Random Walks, Markov Chains, and Stationarity Distributions~~

- Google's PageRank Algorithm

PageRank

• Page rank is named after Larry Page.

• He was doing a PhD at Stanford when he started working on the project of building a search engine.

• He didn’t finish his PhD, but he is currently the Alphabet CEO and worth around 53 billion USD.

PageRank

- PageRank treats the web as a huge graph, where webpages are vertices, and hyperlinks are directed edges.

- The PageRank algorithm simply applies the power iteration method to compute the stationary distribution of a random walk on the web.

- Recall that we needed *all* entries in P to be strictly positive to be guaranteed that this works.

- That means that from any vertex, there has to be nonzero probability of transitioning to *any* other vertex.

PageRank

- To satisfy this, PageRank assumes a slightly different random walk than we described. In particular:

- Start at a random vertex
- For t from 1 to T steps:
  - If current page has no links
    - Choose a page uniformly at random.
  - Else
    - With probability 0.15, choose a page uniformly at random.
    - With the remaining probability, choose a link from the current page uniformly at random and follow it.

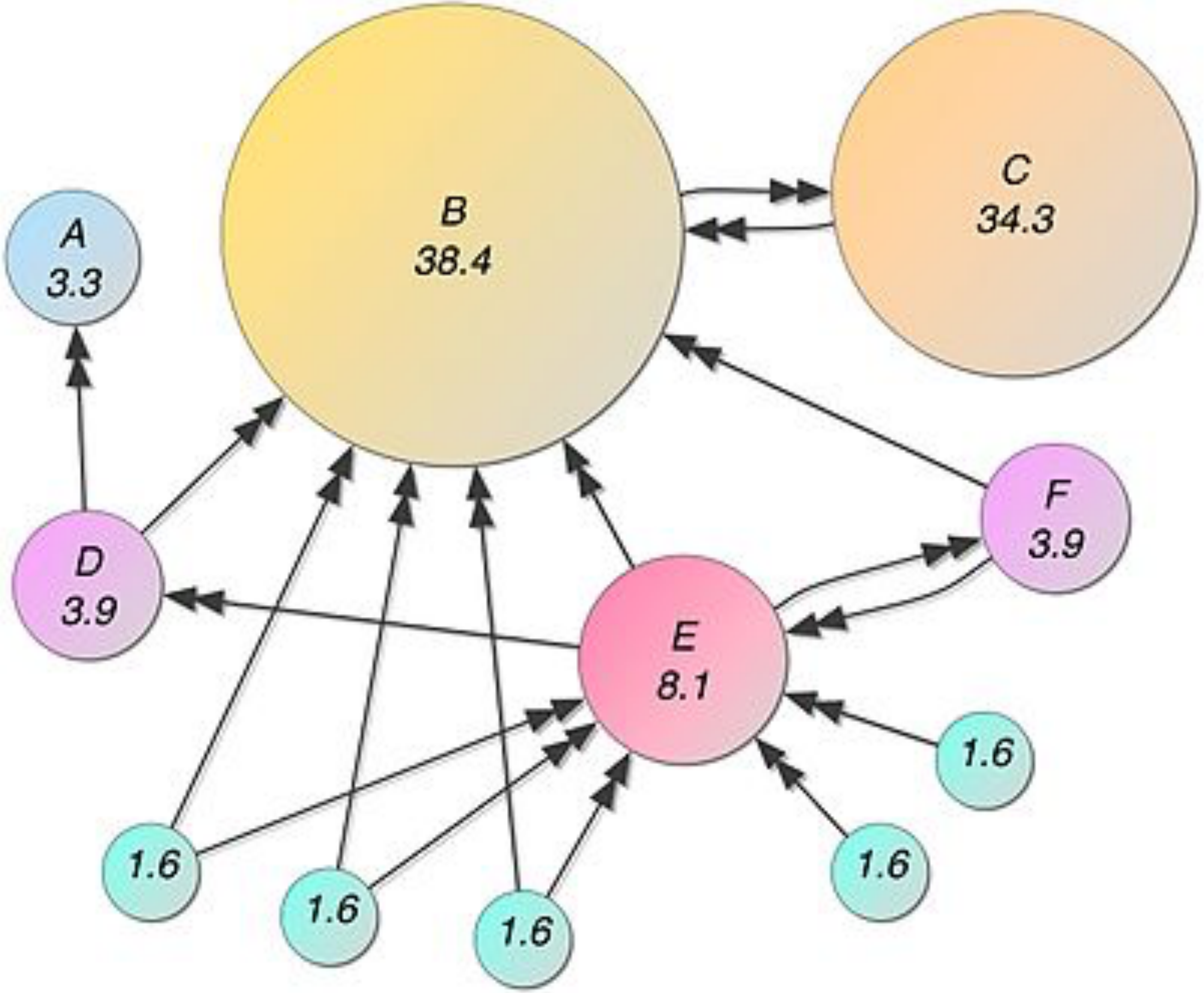
PageRank

- Thus, if there are n web pages in total, the transition matrix for this random walk is given by

$$P_{ij} = \begin{cases} \dfrac{0.85A_{ij}}{d_i} + \dfrac{0.15}{n}, & i\ has\ links \\ \dfrac{1}{n}, & i\ has\ no\ links \end{cases}$$

- Then we just compute the stationary distribution by the power iteration method.
- What kind results does this generate?

PageRank



PageRank

- Note that our modification also ensures that the conductance of the graph is not too small. In practice, 50 to 100 power iterations suffice for a reasonable approximation to the stationary distribution.

- This might seem hard for large n, but note that the graph itself is extremely sparse, so matrix – vector multiplication can be implemented efficiently.

- All other things equal, google search prefers to show results with higher PageRank.

- The #1 thing that increases your PageRank?
  - Having other important pages link to you.