Python for Complex Network Analysis

Sajeev G P

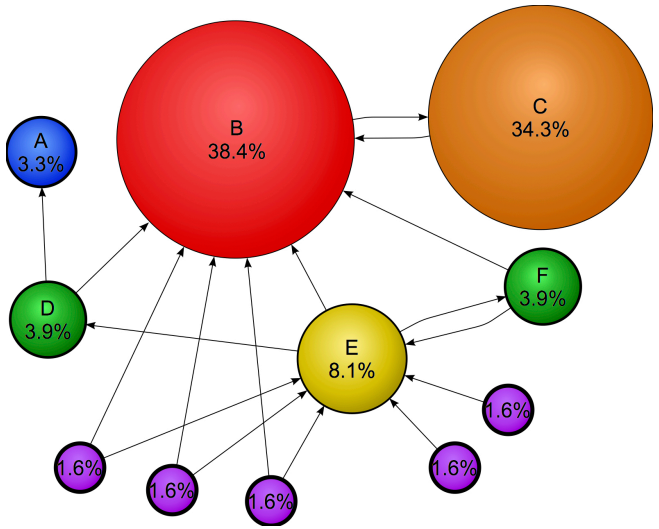
February 17, 2020

PageRank

Developed by Google founders to measure the importance of webpages from the hyperlink network structure.

PageRank assigns a score of importance to each node. Important nodes are those with many in-links from important pages.

PageRank can be used for any type of network, but it is mainly useful for directed networks.

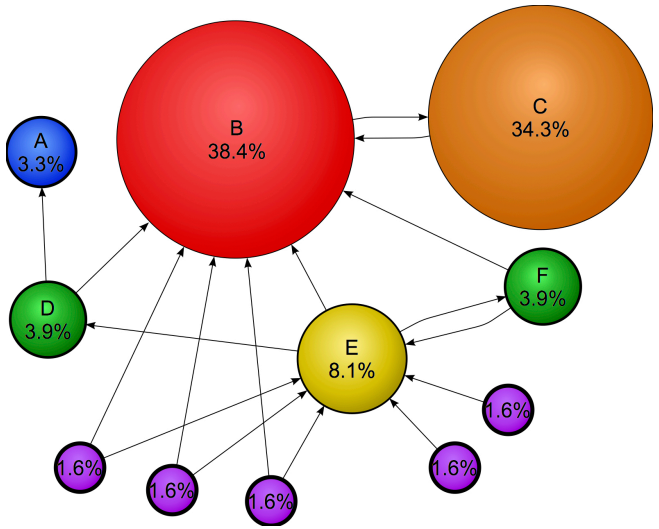


A node's PageRank depends on the PageRank of other nodes (Circular definition?).

PageRank

n = number of nodes in the network
k = number of steps

1. Assign all nodes a PageRank of $1/n$
2. Perform the *Basic PageRank Update Rule* k times.



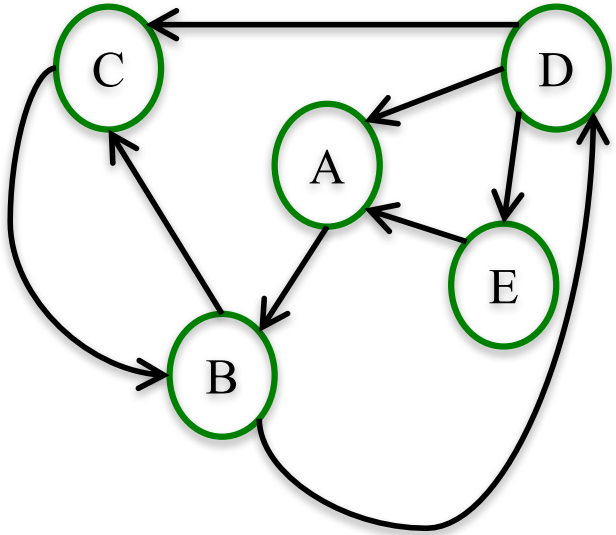
Basic PageRank Update Rule: Each node gives an equal share of its current PageRank to all the nodes it links to.

The new PageRank of each node is the sum of all the PageRank it received from other nodes.

PageRank

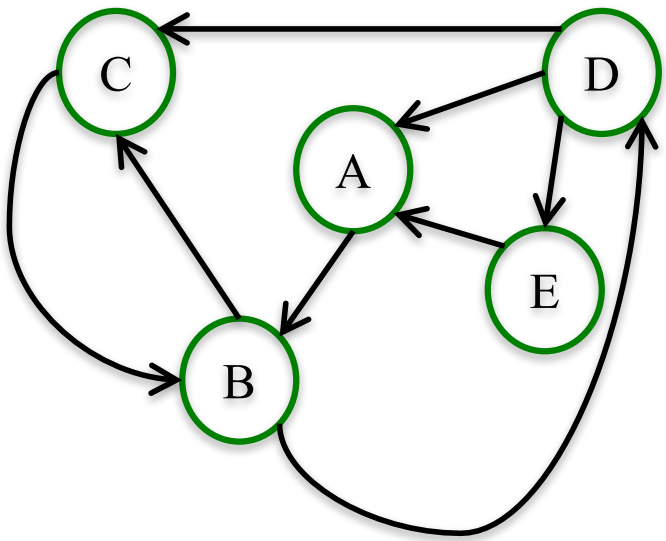
Who should be the most "important" node in this network?

Calculate the PageRank of each node after 2 steps of the procedure (k = 2).



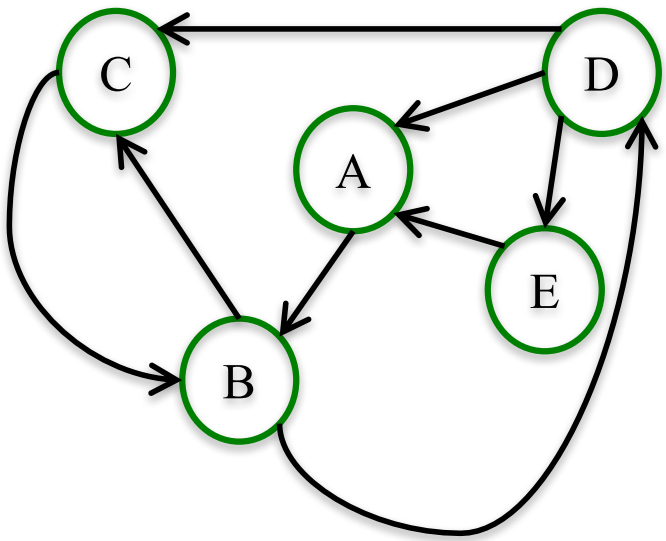
PageRank

Page Rank					
	A	B	C	D	E
	1/5	1/5	1/5	1/5	1/5



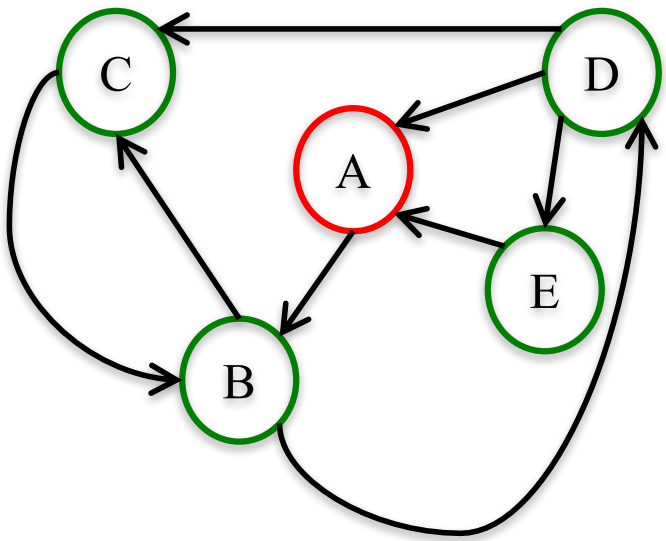
PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New					



PageRank – Step 1

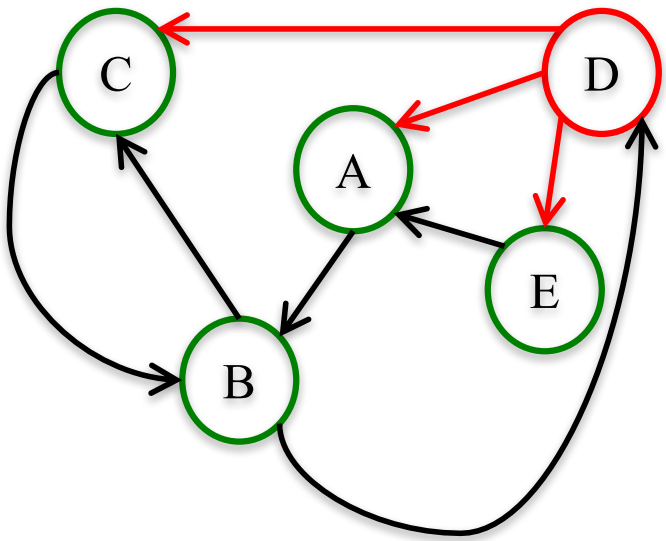
Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New					



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New					

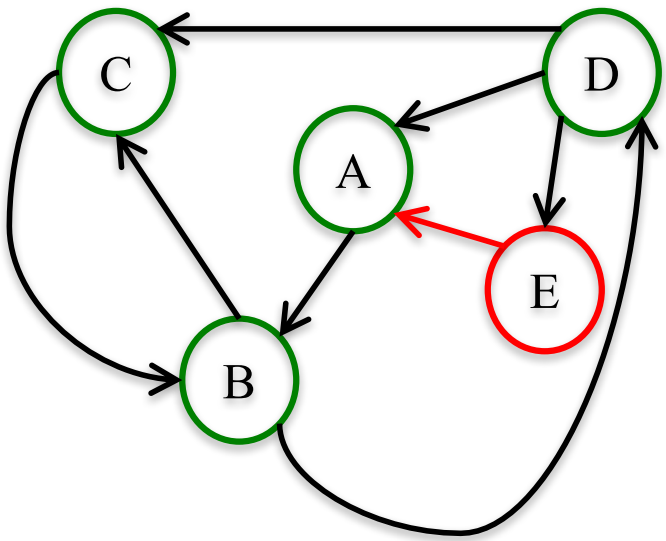
A: (1/3)*(1/5)



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New					

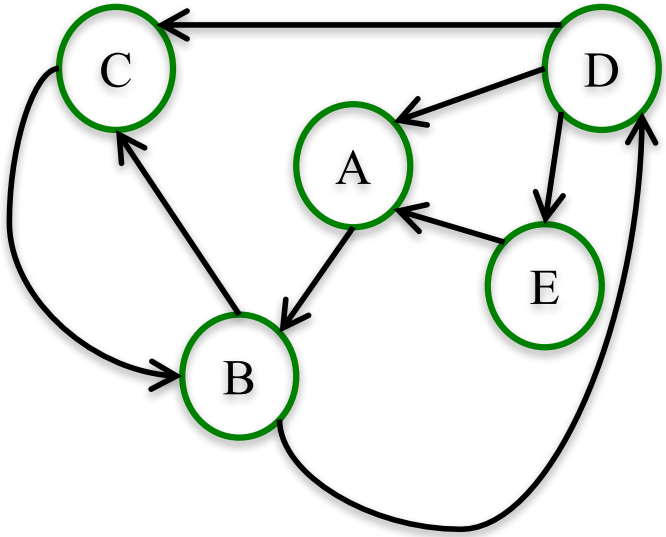
A: (1/3)*(1/5) + 1/5



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New					

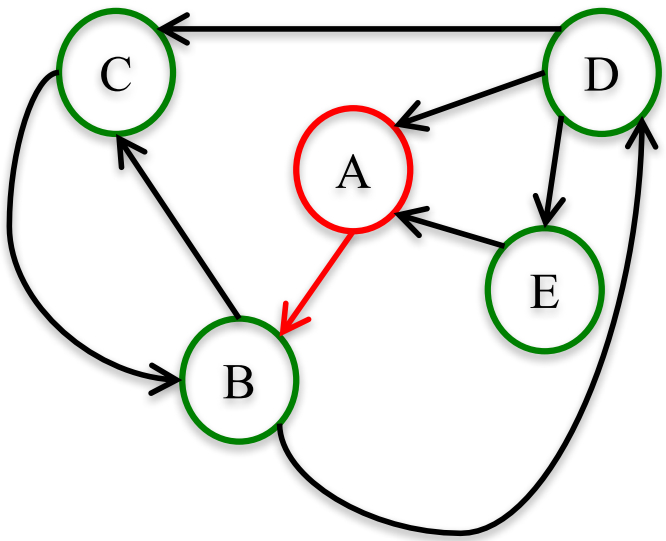
A: (1/3)*(1/5) + 1/5 = 4/15



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15				

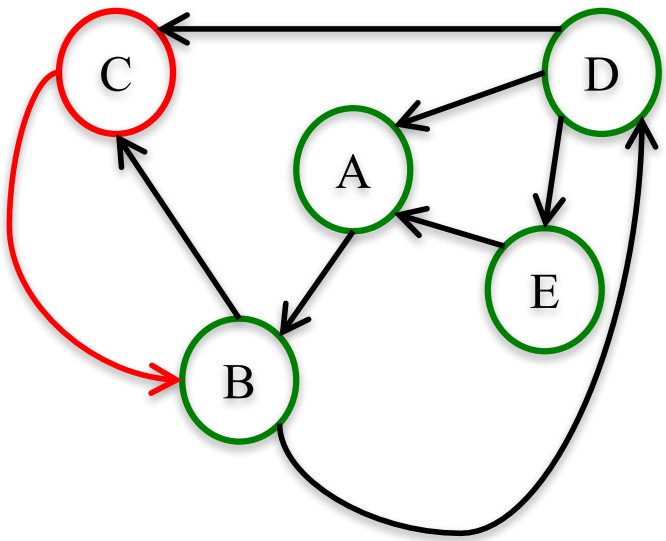
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15				

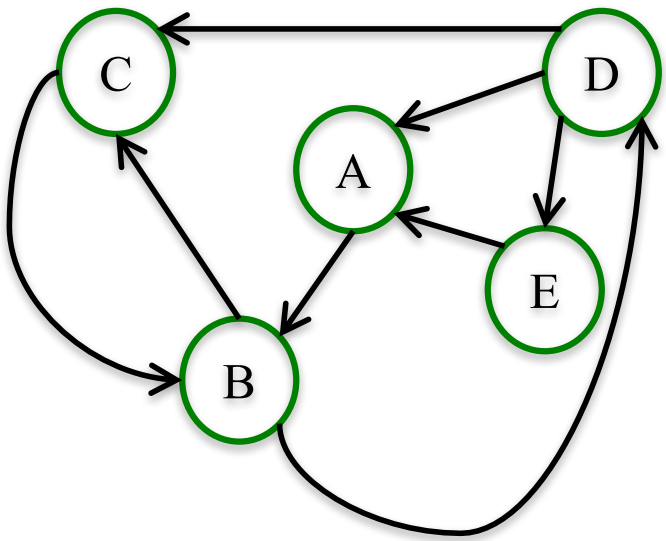
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5			

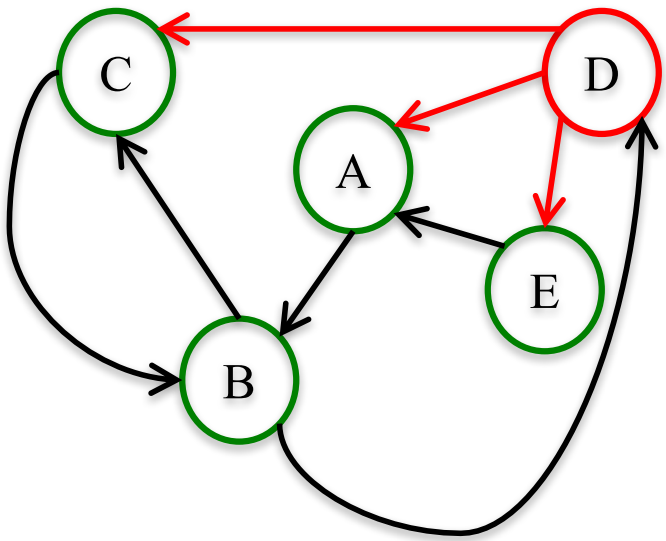
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5			

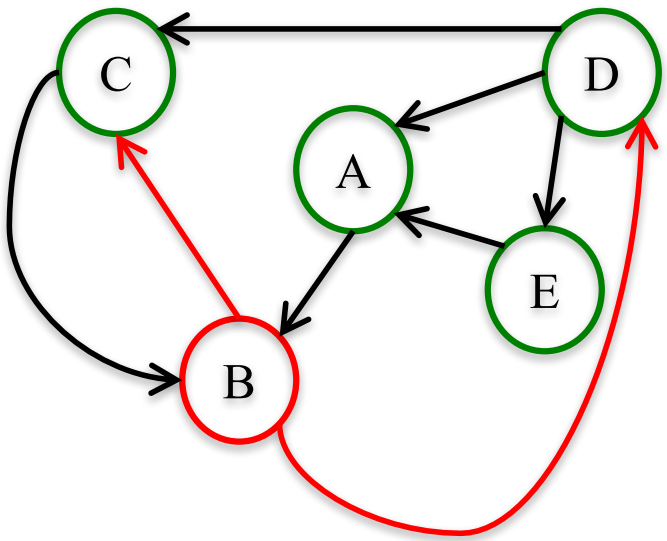
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5)



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5			

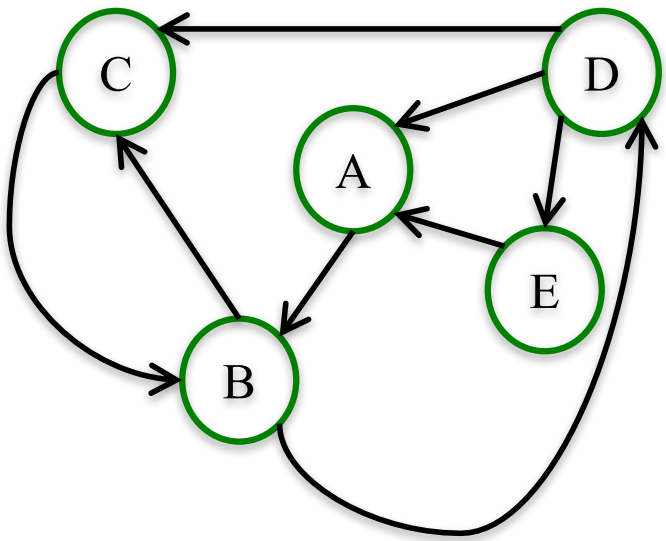
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5)



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5	1/6		

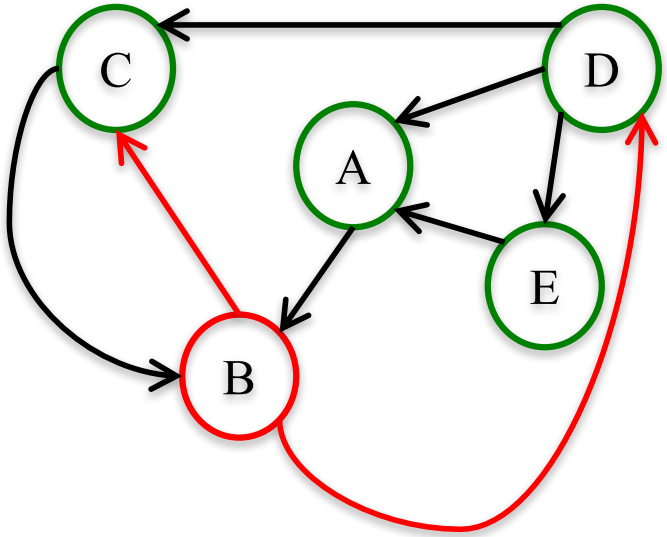
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5) = 5/30 = 1/6



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5	1/6		

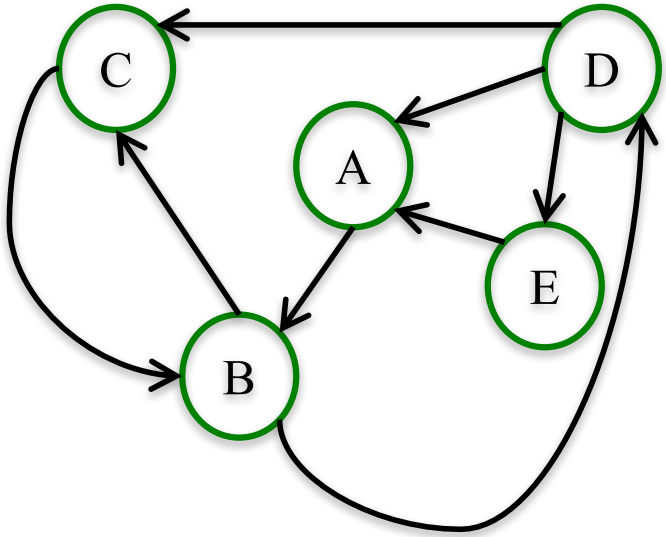
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5) = 5/30 = 1/6
D: (1/2)*(1/5)



PageRank – Step 1

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5	1/6	1/10	

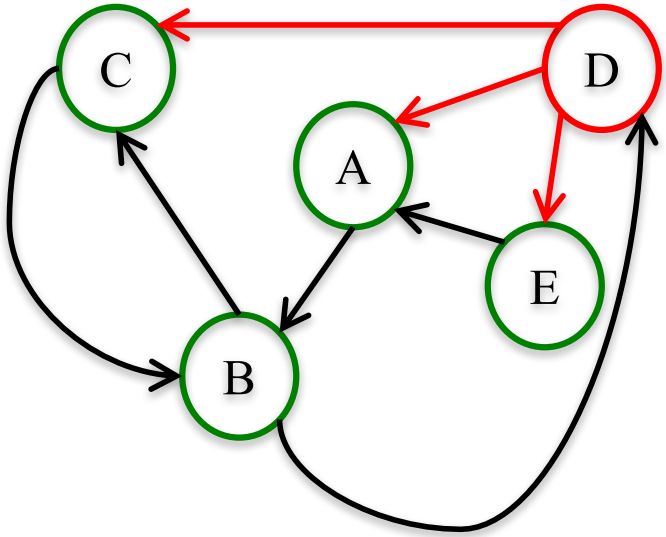
A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5) = 5/30 = 1/6
D: (1/2)*(1/5) = 1/10



PageRank – Step 1

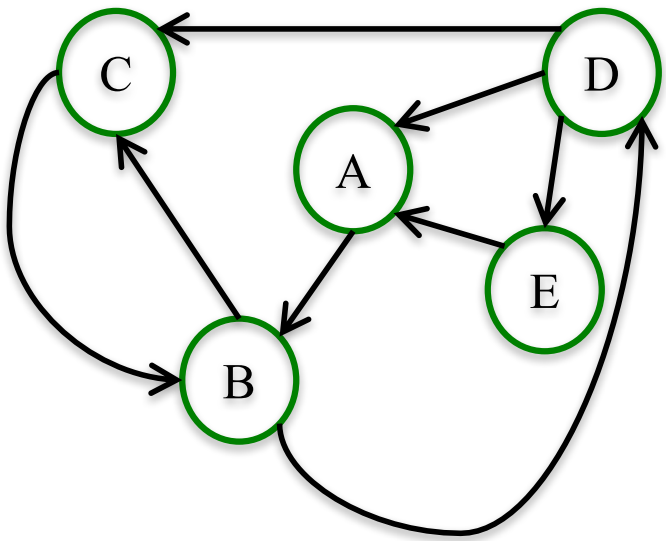
Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5	1/6	1/10	

A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5) = 5/30 = 1/6
D: (1/2)*(1/5) = 1/10
E: (1/3)*(1/5)



PageRank – Step 1

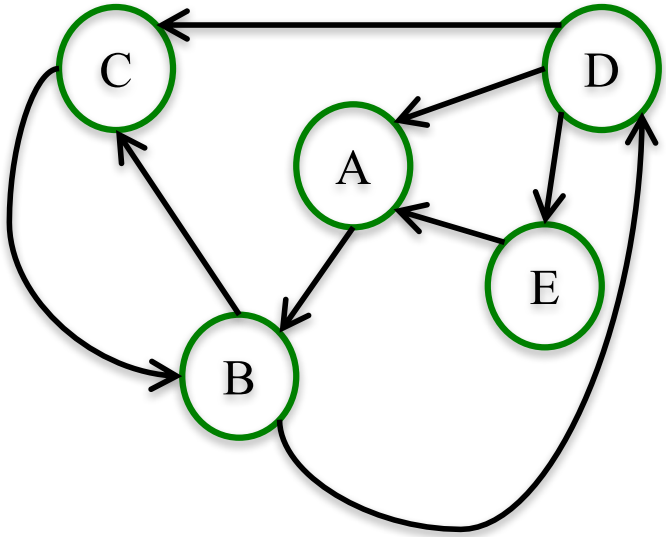
Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5	1/6	1/10	1/15



A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5) = 5/30 = 1/6
D: (1/2)*(1/5) = 1/10
E: (1/3)*(1/5) = 1/15

PageRank – Step 2

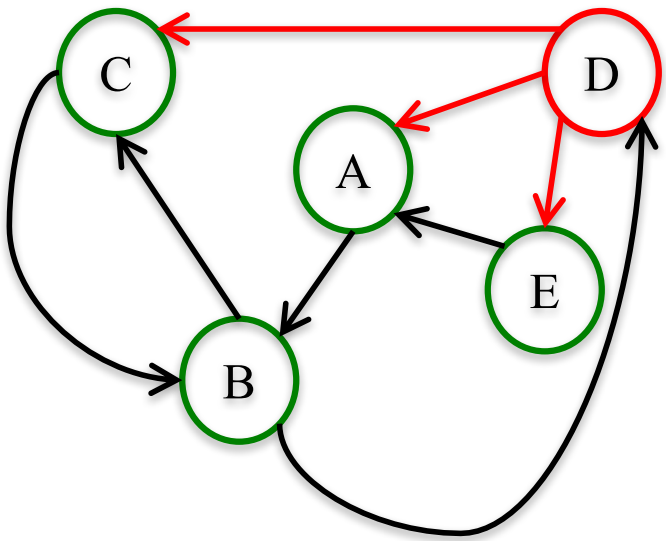
Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New					



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New					

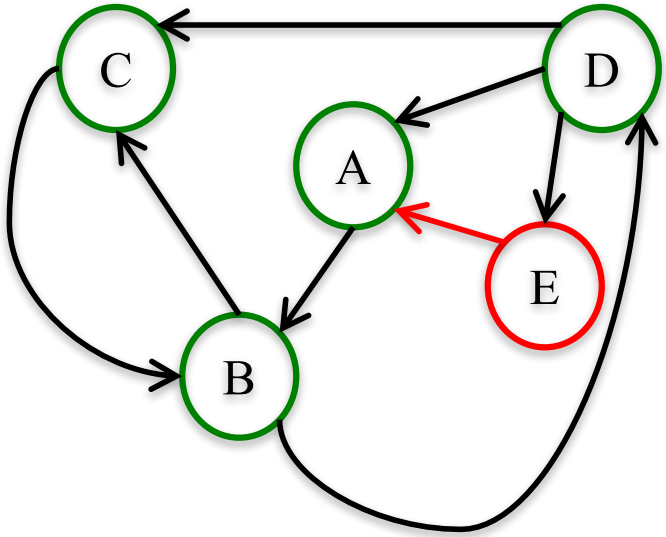
A: (1/3)*(1/10)



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New					

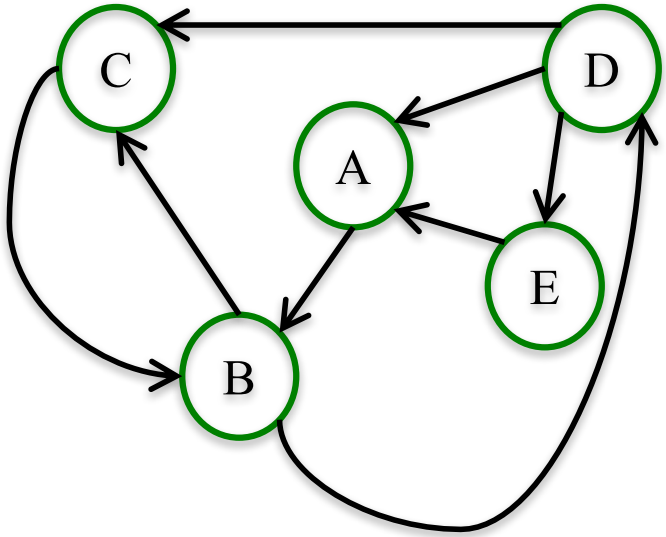
A: (1/3)*(1/10) + 1/15



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10				

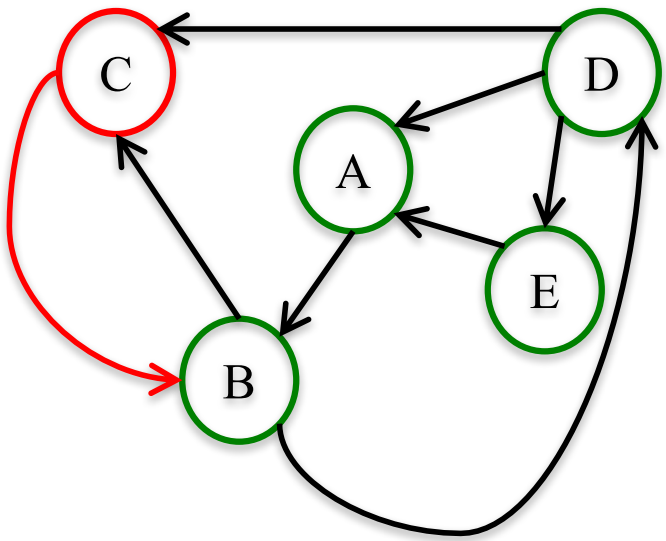
A: (1/3)*(1/10) + 1/15 = 1/10



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10				

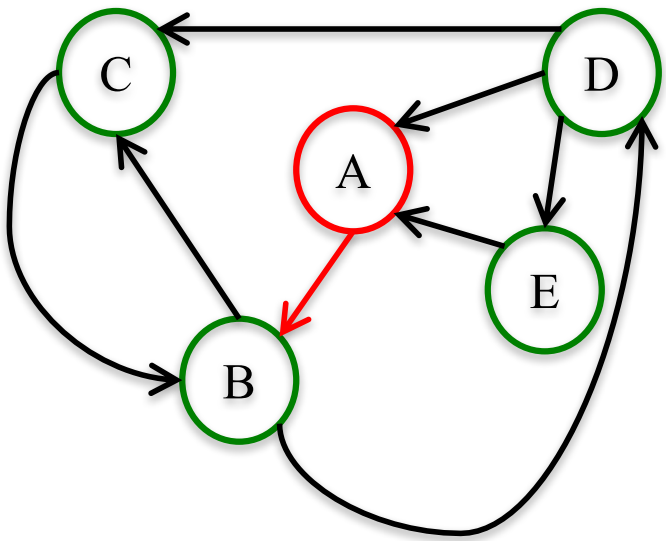
A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10				

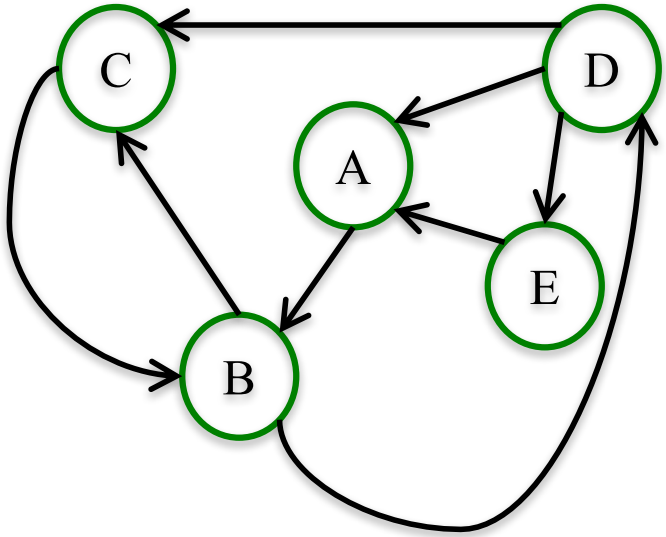
A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30			

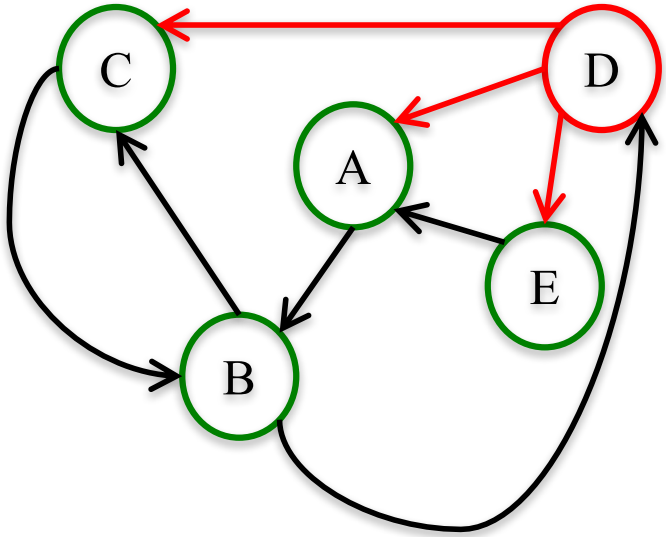
A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30			

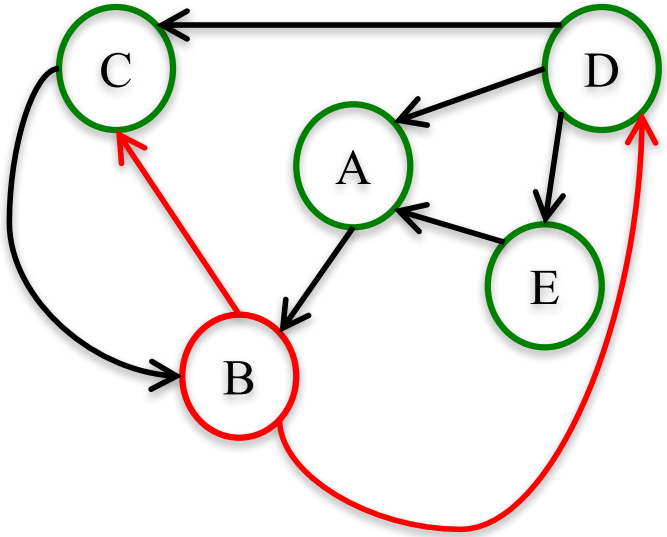
A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10)



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30			

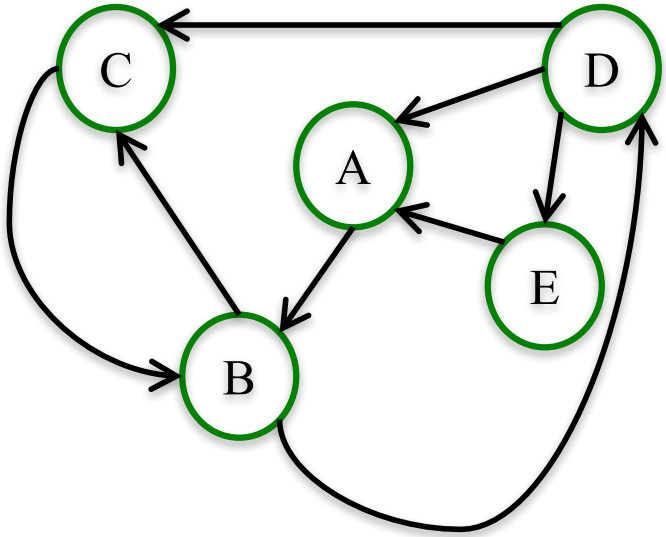
A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10) + (1/2)*(2/5)



PageRank – Step 2

Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30	7/30		

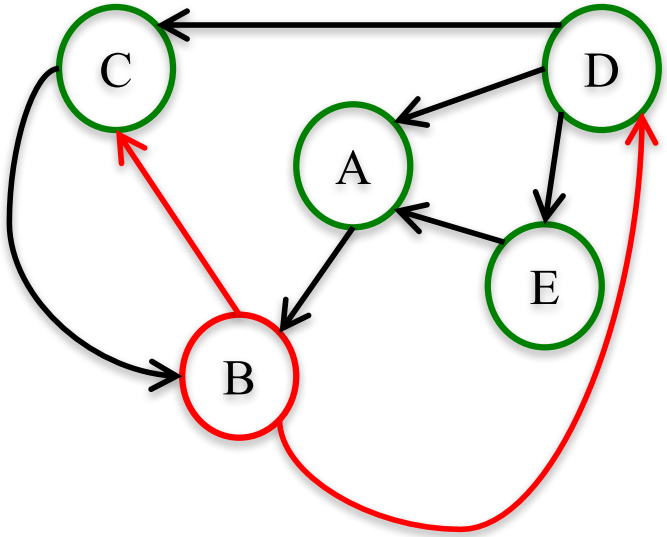
A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10) + (1/2)*(2/5) = 7/30



PageRank – Step 2

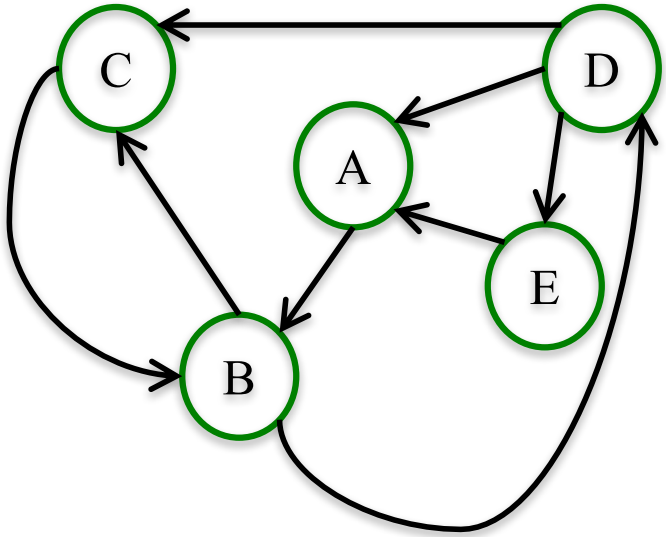
Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30	7/30		

A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10) + (1/2)*(2/5) = 7/30
D: (1/2)*(2/5)



PageRank – Step 2

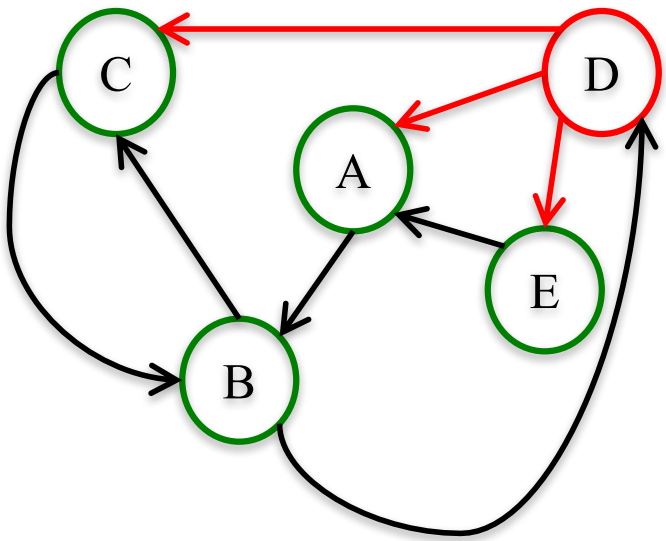
Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30	7/30	2/10	



A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10) + (1/2)*(2/5) = 7/30
D: (1/2)*(2/5) = 2/10

PageRank – Step 2

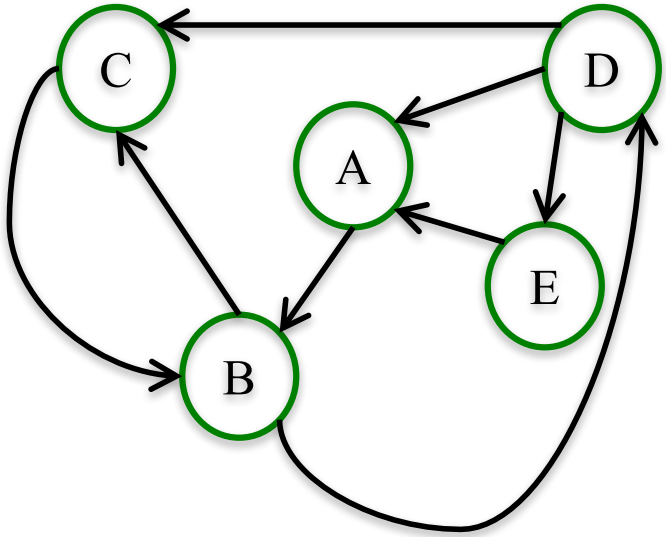
Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30	7/30	2/10	



A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10) + (1/2)*(2/5) = 7/30
D: (1/2)*(2/5) = 2/10
E: (1/3)*(1/10)

PageRank – Step 2

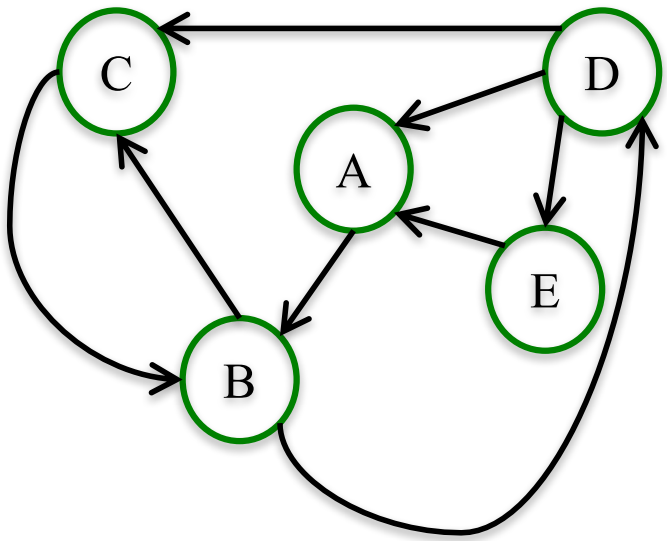
Page Rank (k = 2)					
	A	B	C	D	E
Old	4/15	2/5	1/6	1/10	1/15
New	1/10	13/30	7/30	2/10	1/30



A: (1/3)*(1/10) + 1/15 = 1/10
B: 1/6 + 4/15 = 13/30
C: (1/3)*(1/10) + (1/2)*(2/5) = 7/30
D: (1/2)*(2/5) = 2/10
E: (1/3)*(1/10) = 1/30

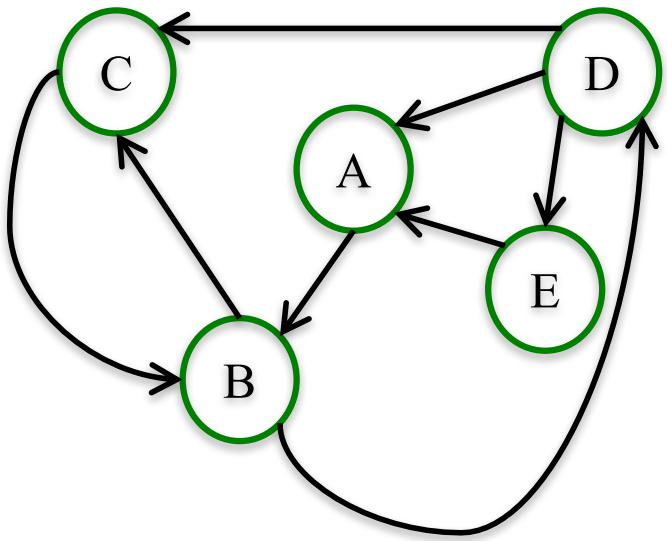
PageRank

	Page Rank				
	A	B	C	D	E
k=2	1/10	13/30	7/30	2/10	1/30
k=2	.1	.43	.23	.20	.03



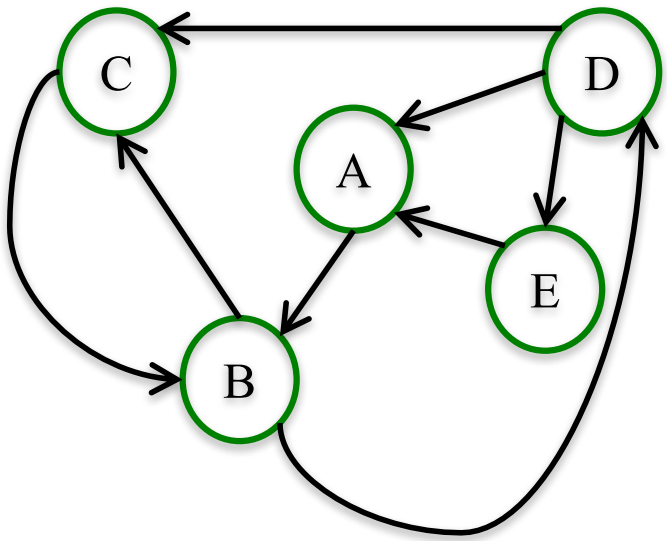
PageRank

	Page Rank				
	A	B	C	D	E
k=2	1/10	13/30	7/30	2/10	1/30
k=2	.1	.43	.23	.20	.03
k=3	.1	.33	.28	.22	.06



PageRank

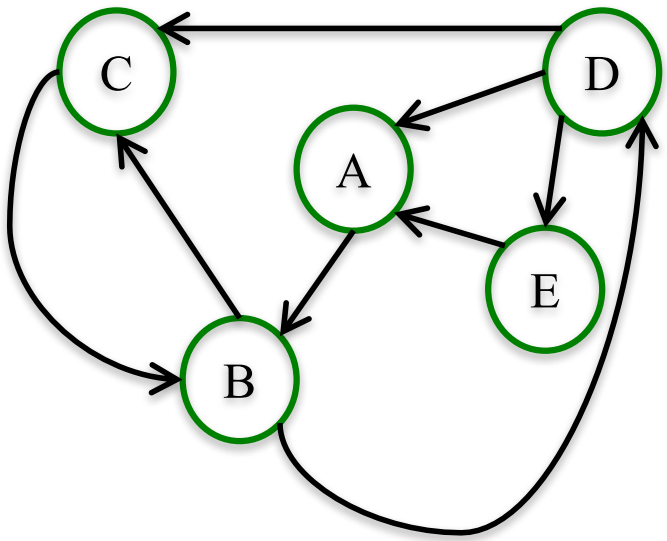
	Page Rank				
	A	B	C	D	E
k=2	1/10	13/30	7/30	2/10	1/30
k=2	.1	.43	.23	.20	.03
k=3	.1	.33	.28	.22	.06



What if continue with k = 4,5,6,...?

PageRank

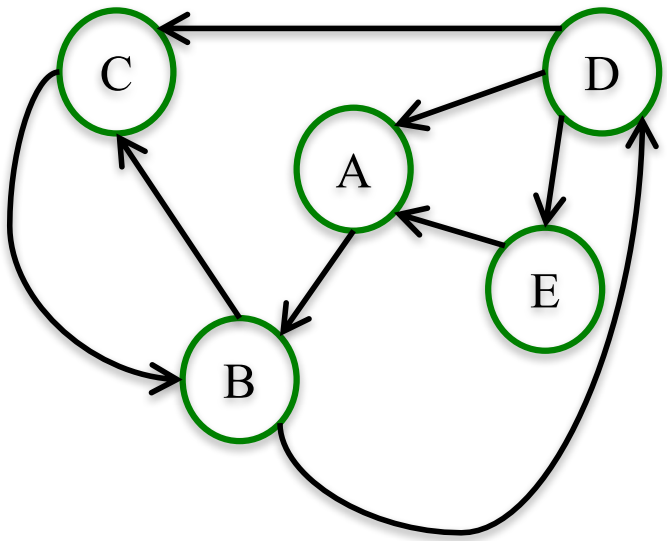
	Page Rank				
	A	B	C	D	E
k=2	1/10	13/30	7/30	2/10	1/30
k=2	.1	.43	.23	.20	.03
k=3	.1	.33	.28	.22	.06
k=∞	.12	.38	.25	.19	.06



What if continue with k = 4,5,6,...?

PageRank

	Page Rank				
	A	B	C	D	E
k=2	1/10	13/30	7/30	2/10	1/30
k=2	.1	.43	.23	.20	.03
k=3	.1	.33	.28	.22	.06
k=∞	.12	.38	.25	.19	.06



What if continue with k = 4,5,6,...? For most networks, PageRank values converge.

Summary

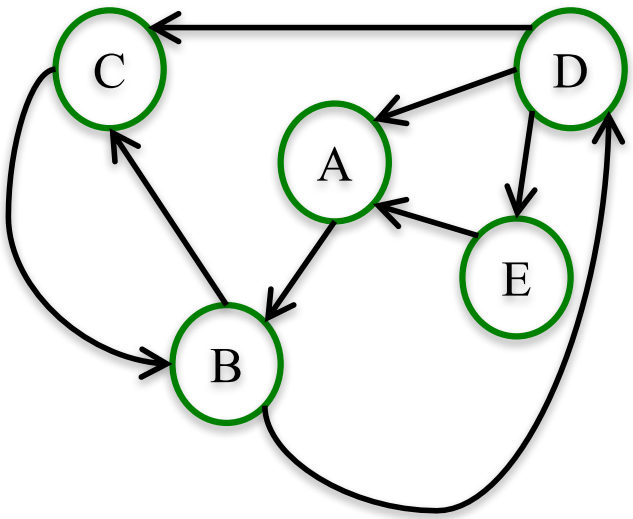
Steps of Basic PageRank:

1. All nodes start with PageRank of $1/n$
2. Perform the *Basic PageRank Update Rule* k times:
   - **Basic PageRank Update Rule**: Each node gives an equal share of its current PageRank to all the nodes it links to.
   - The new PageRank of each node is the sum of all the PageRank it received from other nodes.

For most networks, PageRank values converge as k gets larger ($k \to \infty$)

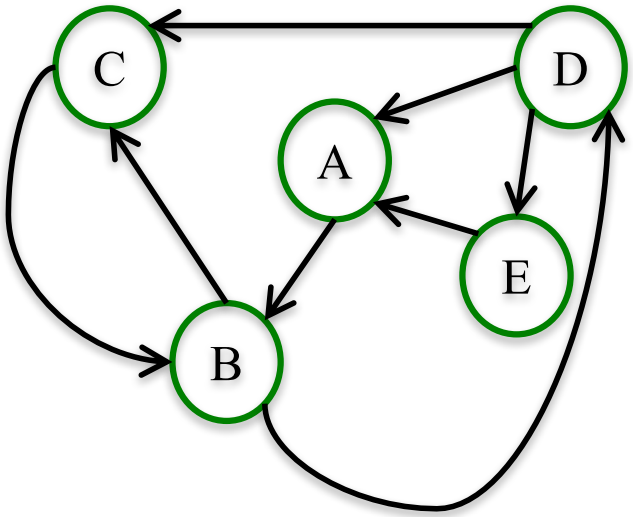
Interpreting PageRank

The PageRank of a node at step k is the probability that a ***random walker*** lands on the node after taking k steps.



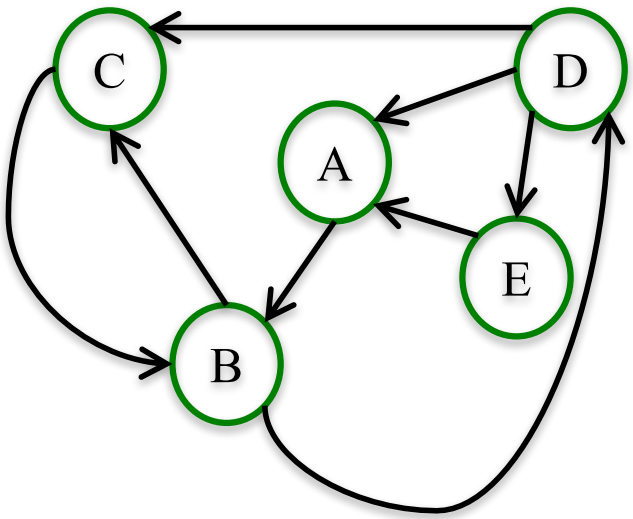
Interpreting PageRank

Random walk of k steps: Start on a random node.



Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node.

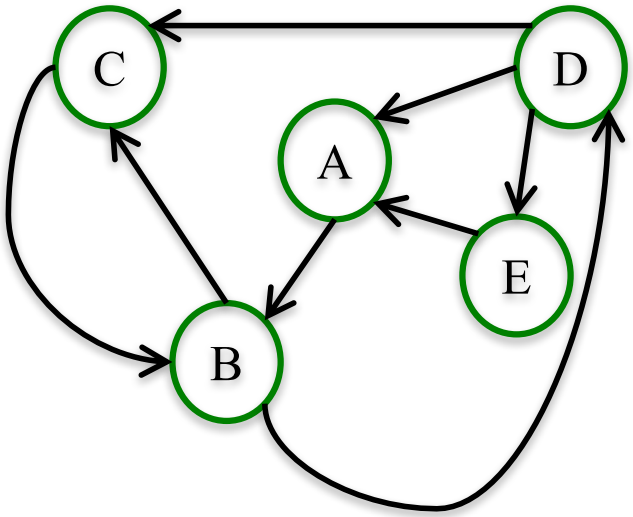


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 1: Choose a random node.

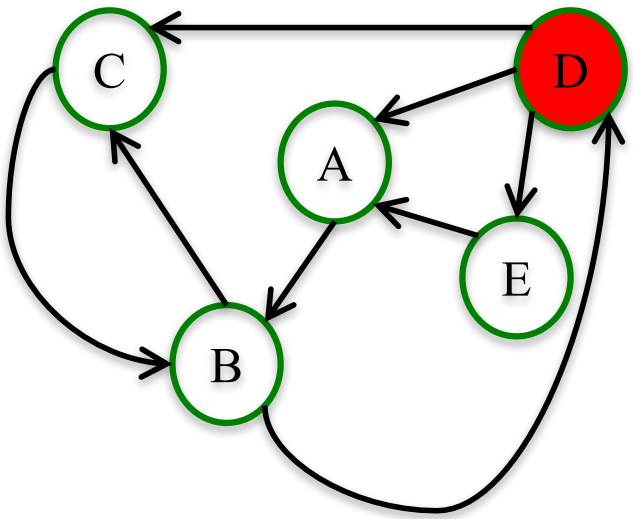


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 1: Choose a random node.

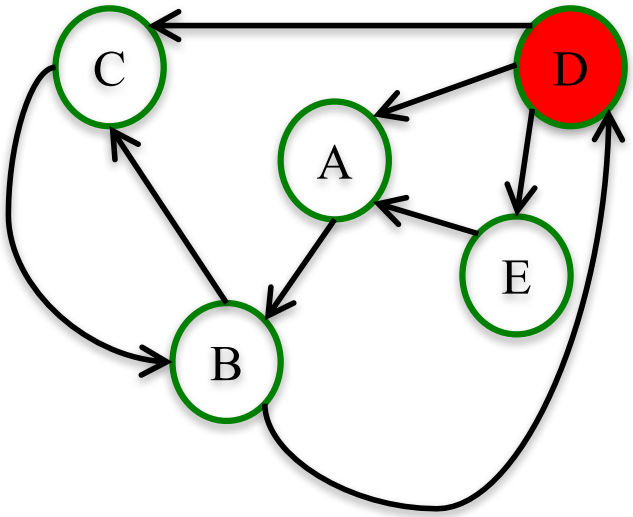


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 1: Choose a random outgoing edge.

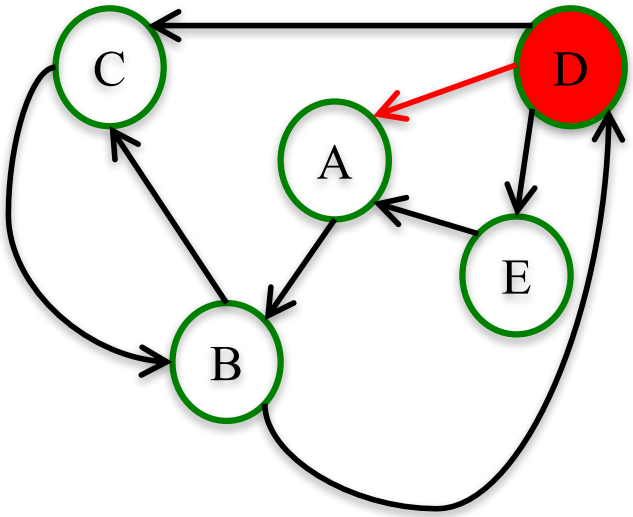


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 1: Choose a random outgoing edge.

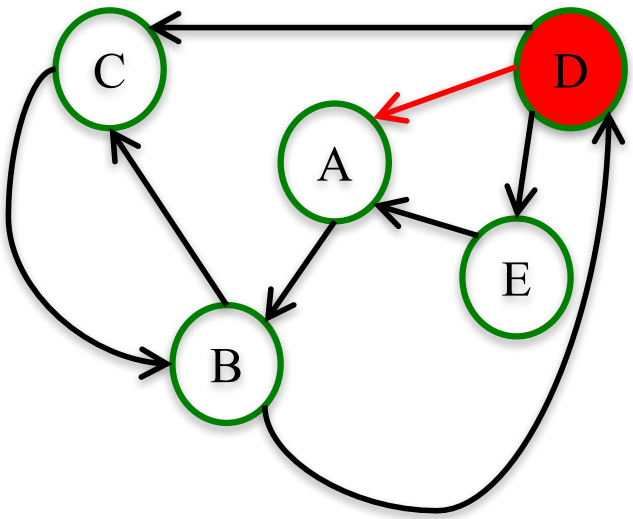


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 1: Follow the edge to the next node.

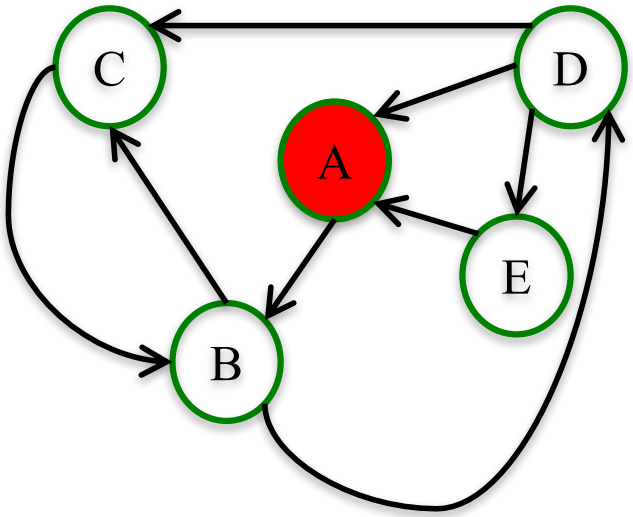


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 1: Follow the edge to the next node.

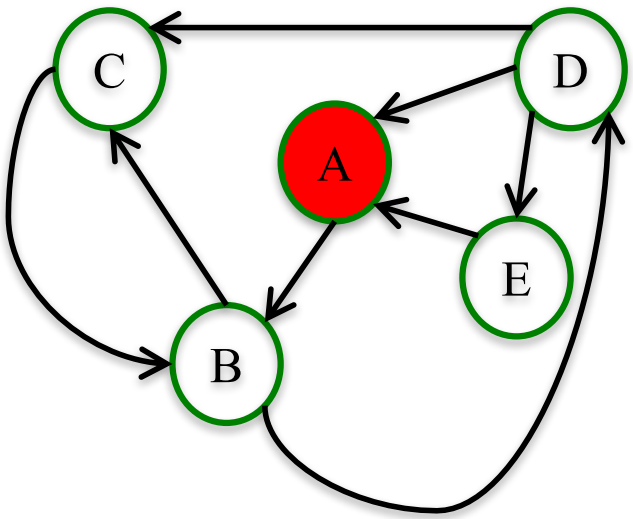


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 2: Choose a random outgoing edge and follow it.

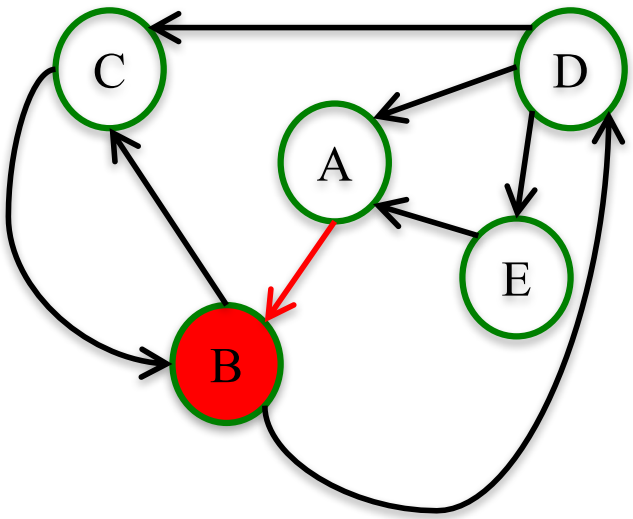


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 2: Choose a random outgoing edge and follow it.

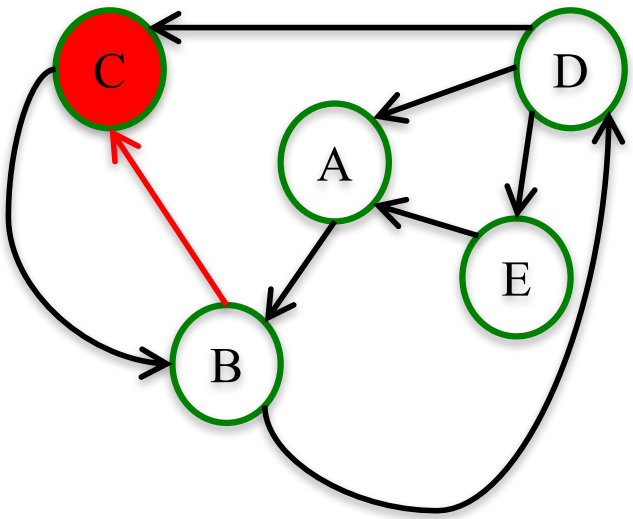


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 3: Choose a random outgoing edge and follow it.

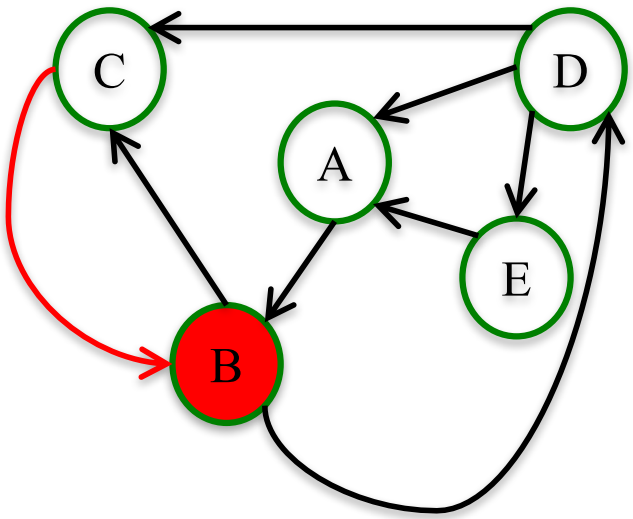


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

Step 4: Choose a random outgoing edge and follow it.

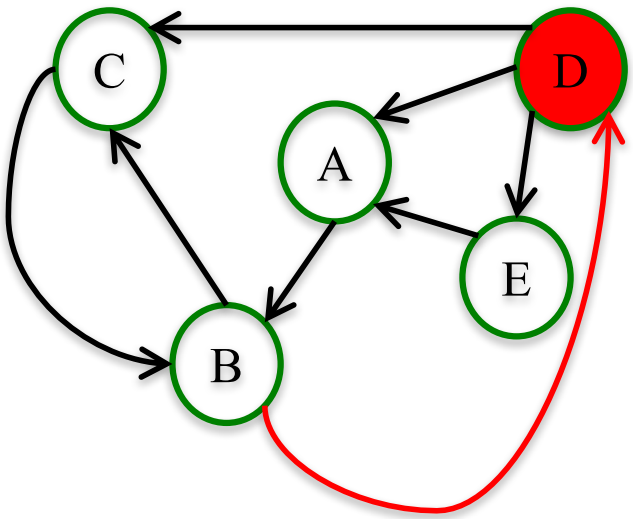


Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

For example, a random walk of 5 steps on this graph looks like this:

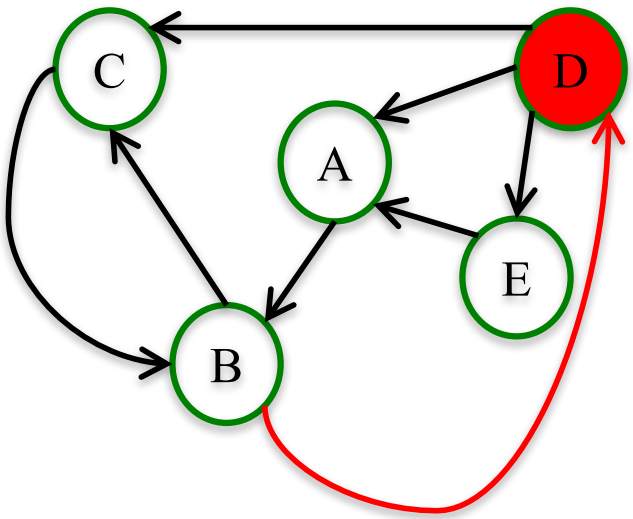
Step 5: Choose a random outgoing edge and follow it.



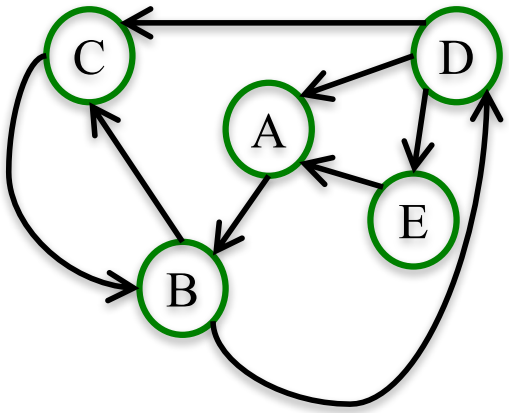
Interpreting PageRank

Random walk of k steps: Start on a random node. Then choose an outgoing edge at random and follow it to the next node. Repeat k times.

	Page Rank				
	A	B	C	D	E
k=∞	.12	.38	.25	.19	.06

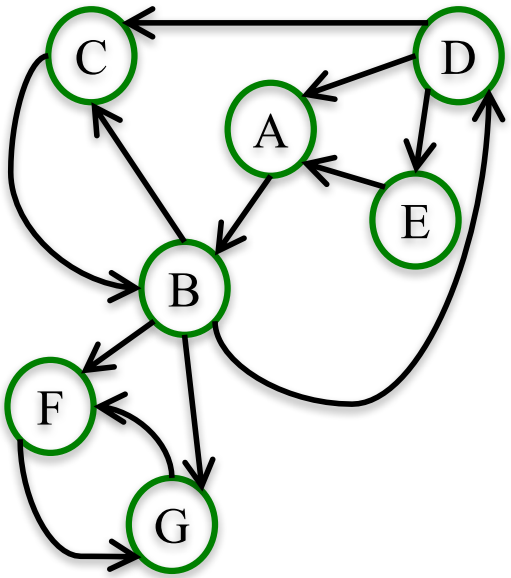


PageRank Problem



PageRank Problem

What's the PageRank of the nodes in this network?
[Hint: think about the random walk interpretation]



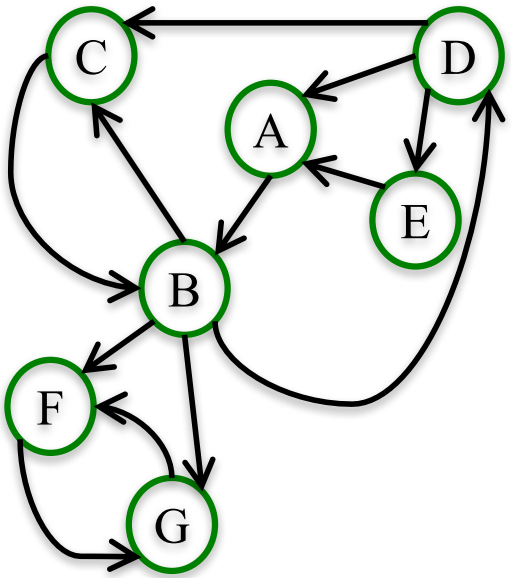
For a large enough k: F and G each have PageRank of ½ and all the other nodes have PageRank 0.

Why? Imagine a random walk on this network. Whenever the walk lands on F or G, it is "stuck" on F and G.

This seems problematic!

PageRank Problem

To fix this, we introduce a "damping parameter" α.



Random walk of k steps with damping parameter α: Start on a random node. Then:

- With probability α: choose an outgoing edge at random and follow it to the next node.
- With probability $1 - \alpha$: choose a node at random and go to it.

Repeat k times.

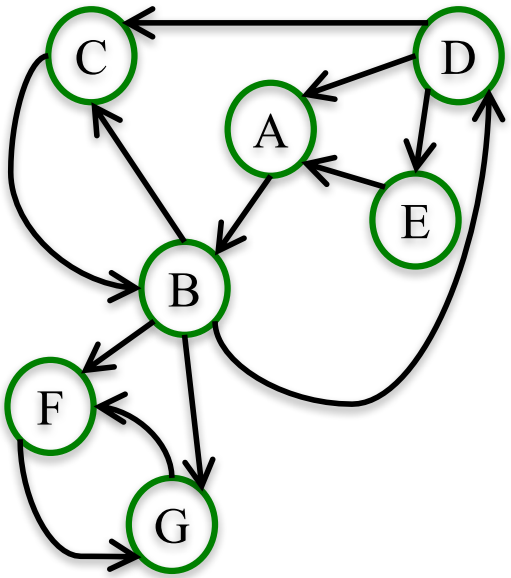
The random walk is no longer "stuck" on nodes F and G.

Scaled PageRank

The **Scaled PageRank** of k steps and damping factor α of a node n is the probability that a random walk with damping factor α lands on a n after k steps.

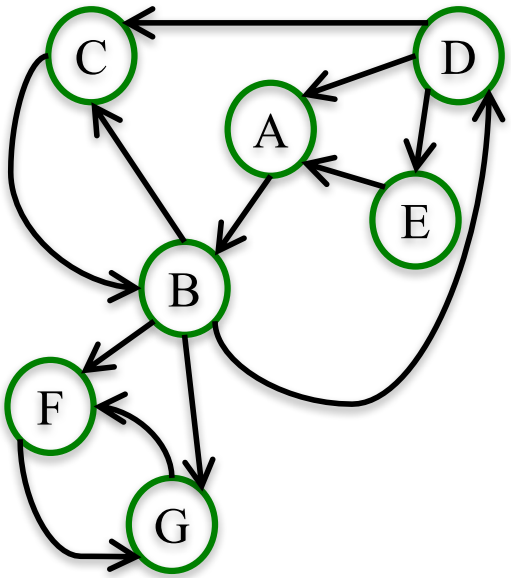
For most networks, as k gets larger, Scaled PageRank converges to a unique value, which depends on α.

In practice, we use a parameter of α between 0.8 and 0.9.



Scaled PageRank

Scaled PageRank ($\alpha = .8, k$ large)						
A	B	C	D	E	F	G
.08	.17	.1	.08	.05	.27	.25



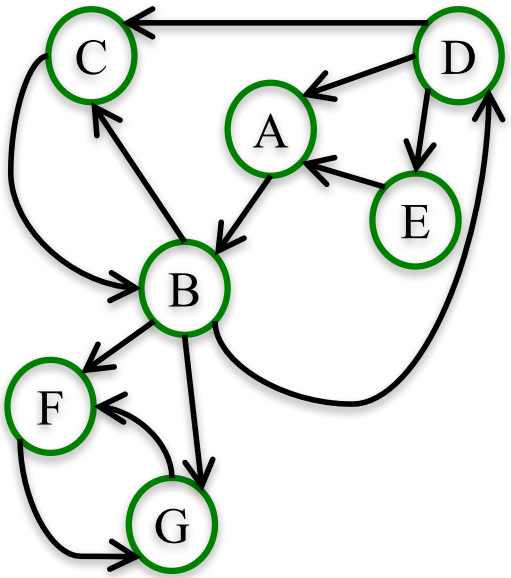
F and G still have high PageRank, but not all the PageRank.

Damping factor works better in very large networks like the Web or large social networks.

You can use NetworkX function pagerank(G, alpha=0.8) to compute Scaled PageRank of network G with damping parameter alpha.

Summary

- The Basic PageRank of a node can be interpreted as the probability that a random walk lands on the node after k random steps.
- Basic PageRank has the problem that, in some networks, a few nodes can "suck up" all the PageRank from the network.
- To fix this problem, Scaled PageRank introduces a parameter α, such that the random walker chooses a random node to jump to with probability $1 - \alpha$.
- Typically we use α between 0.8 and 0.9
- NetworkX function pagerank(G, alpha=0.8) computes Scaled PageRank of network G with damping parameter α=0.8.



Resources

- NetworkX Docs
  http://networkx.lanl.gov/tutorial/index.html
- NetworkX Tutorial
  http://networkx.lanl.gov/contents.html
- Matplotlib Docs
  http://matplotlib.sourceforge.net/contents.html
- Matplotlib Tutorial
  http://matplotlib.sourceforge.net/users/pyplot_tutorial.html
- Numpy Docs
  http://numpy.scipy.org/
- MacPorts
  http://macports.org