October 12, 2017

CS 361: Probability & Statistics

Markov chains & simulation

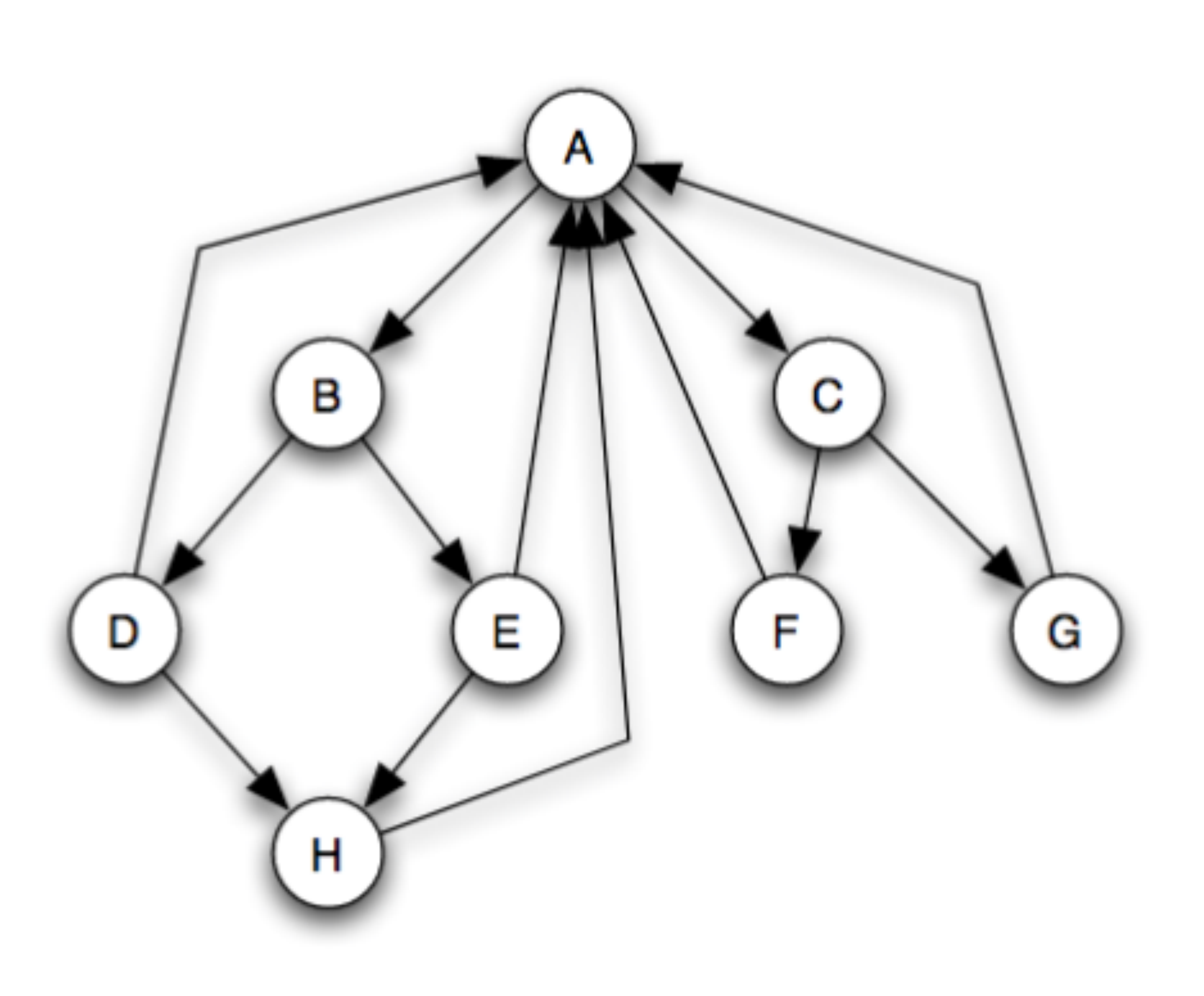
Linking and importance

The goal of a ranking system is to sort a set of objects by their importance

The web, academic publications, blog posts, etc form a network where the links or citations from one node to another can be viewed as an endorsement of the importance of the linked node

The key principle behind PageRank is recursive: a page is important if important pages link to it

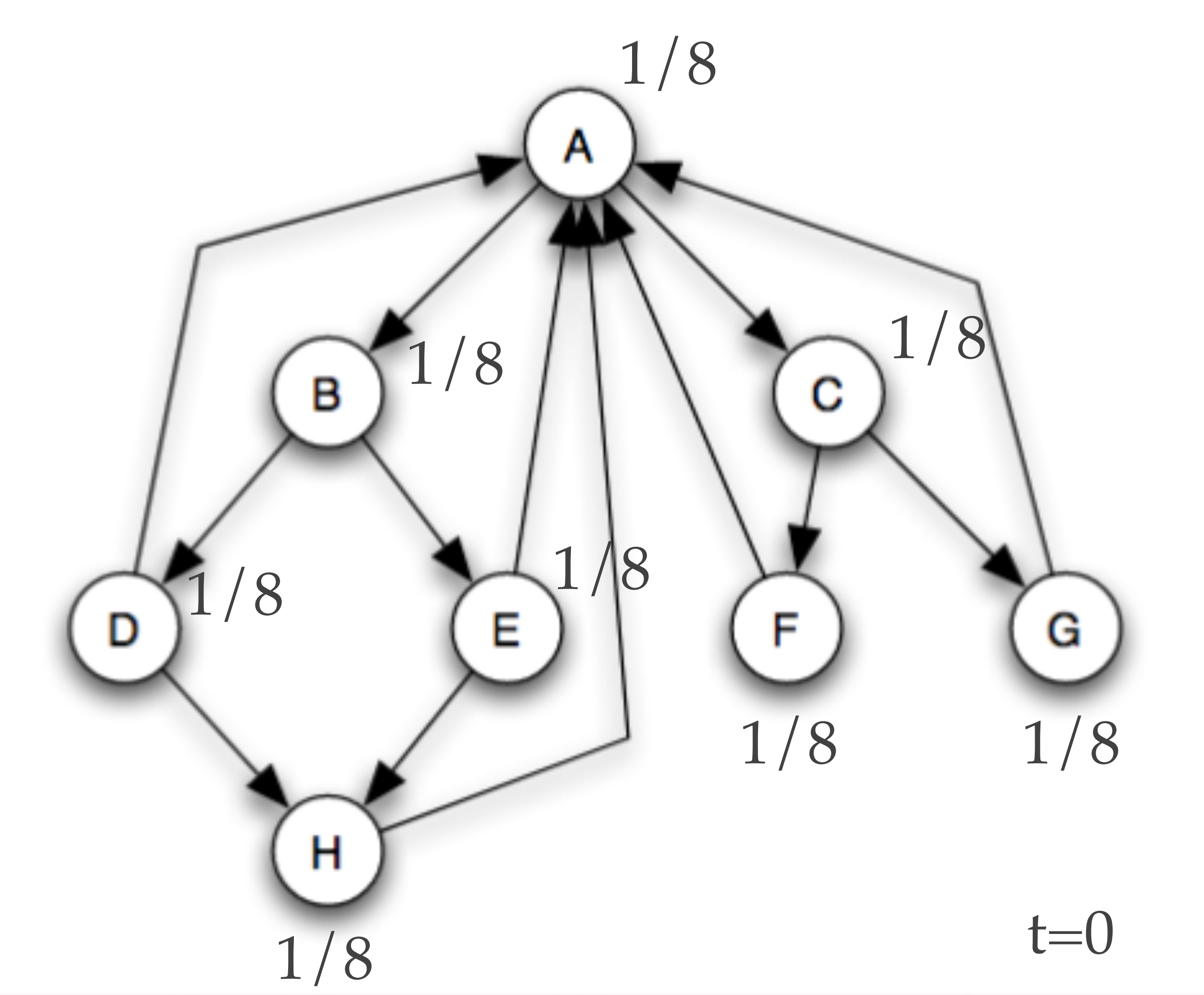
PageRank



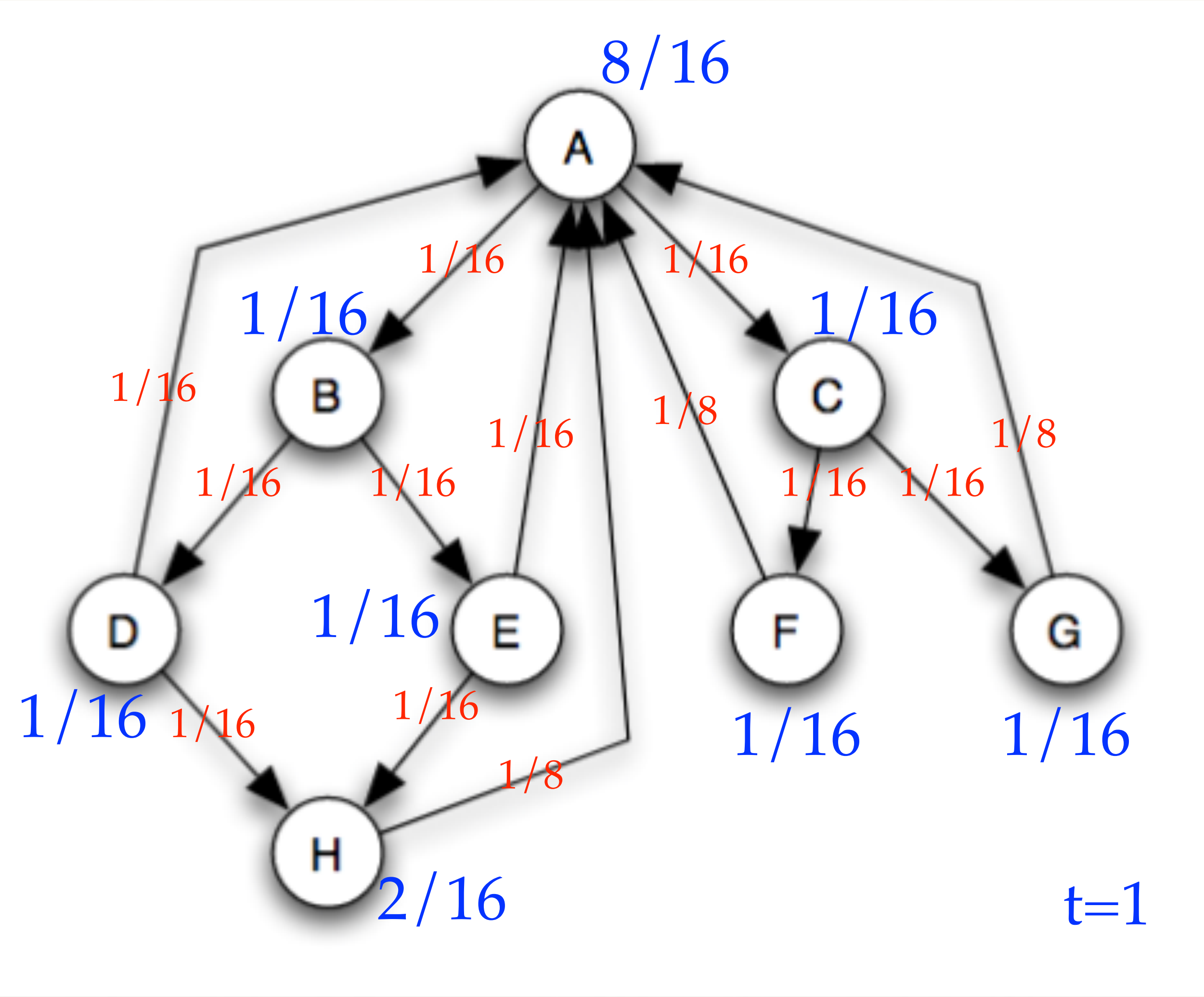
Initialize: Each node starts with 1/n importance

Update: And at each step, propagate importance evenly across outgoing links

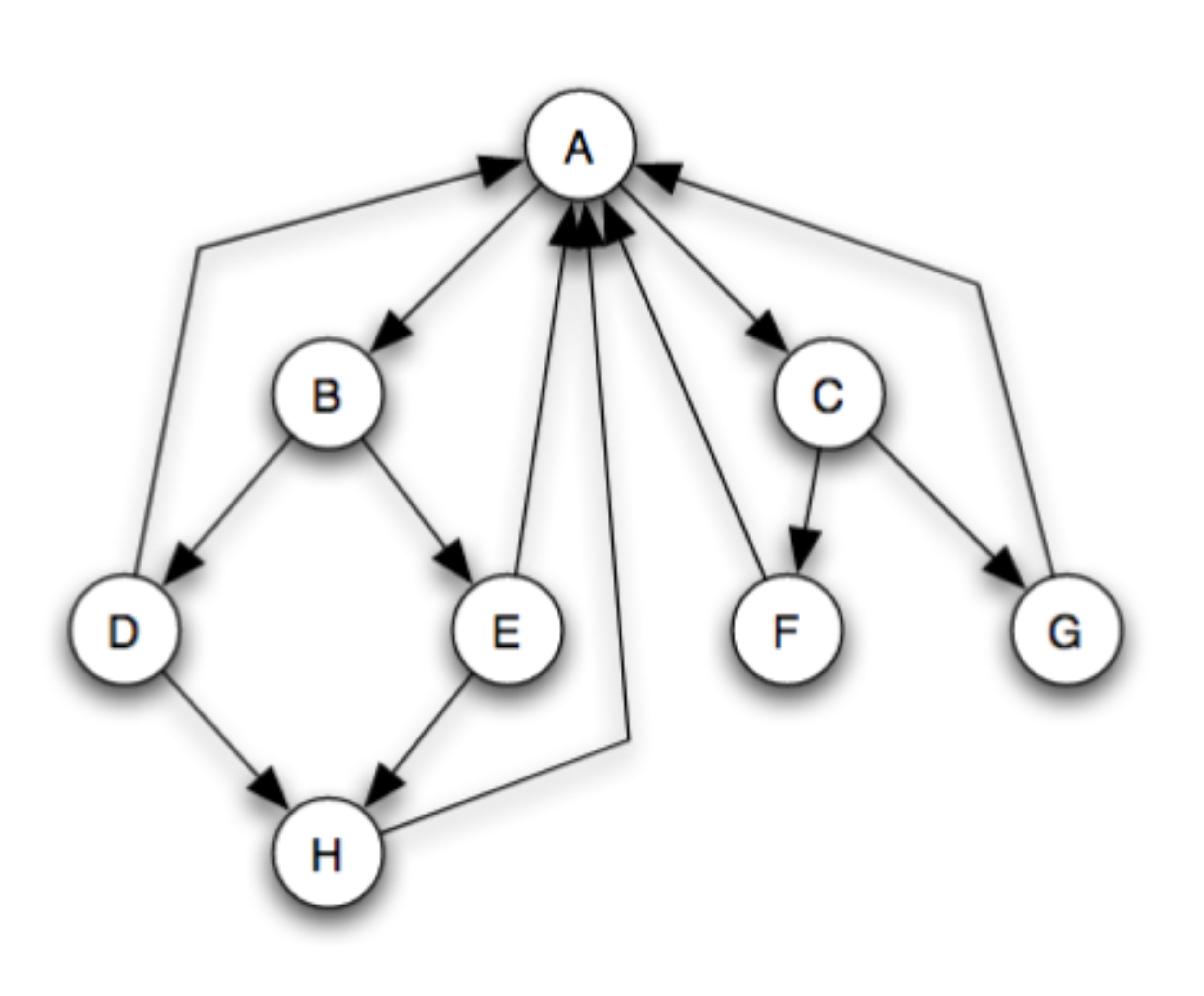
PageRank

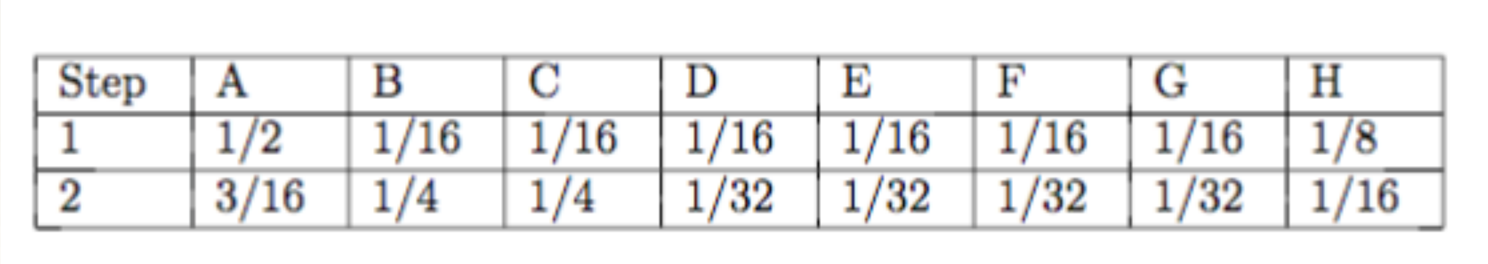


PageRank


8/16
A
1/16
1/16
1/16
B
1/16
C
1/16
1/16
1/8
1/8
1/16
1/16
1/16
1/16
D
1/16
E
F
G
1/16
1/16
1/16
1/16
1/16
1/16
1/8
H
2/16
t=1

PageRank



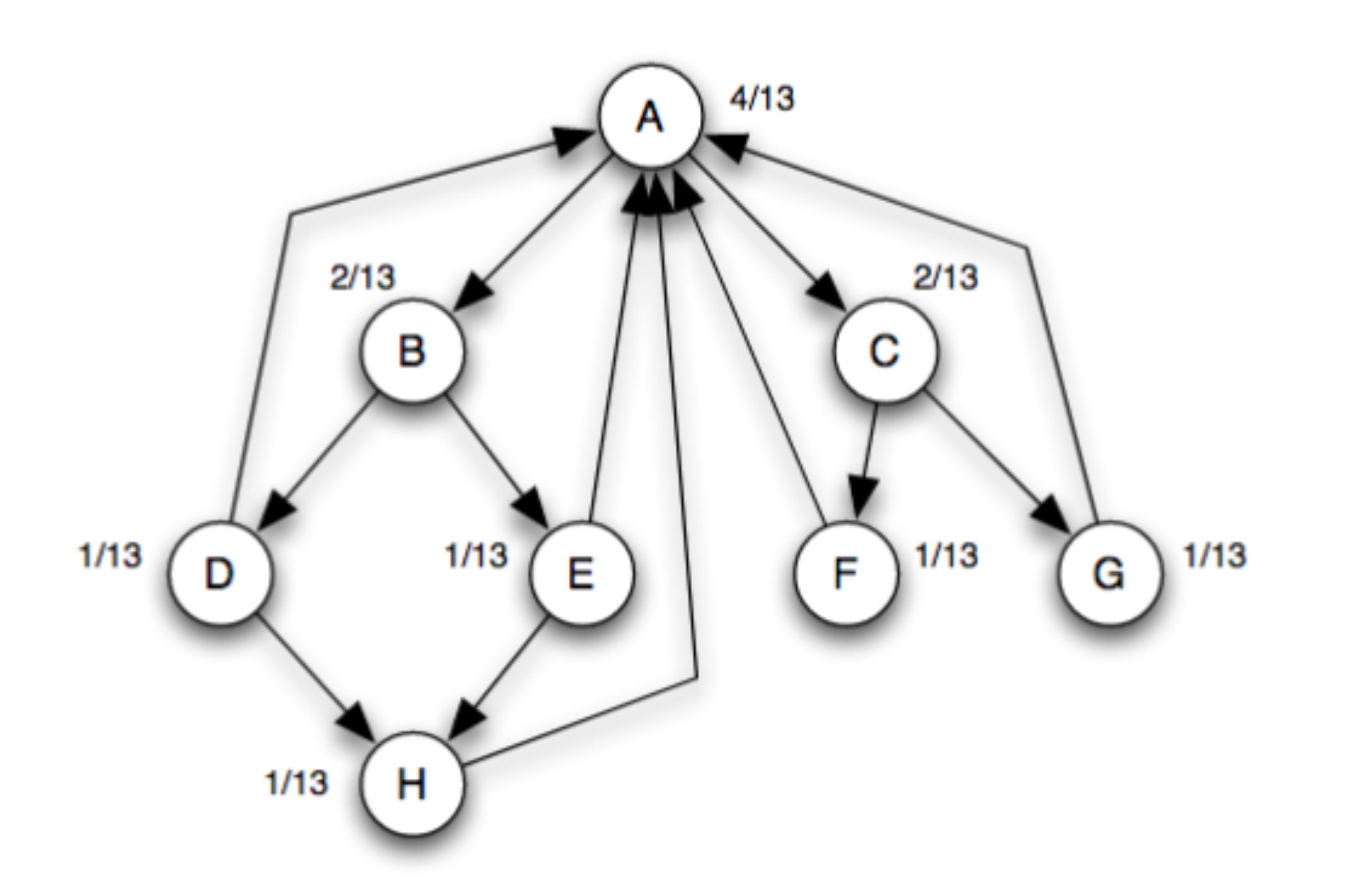


Step	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

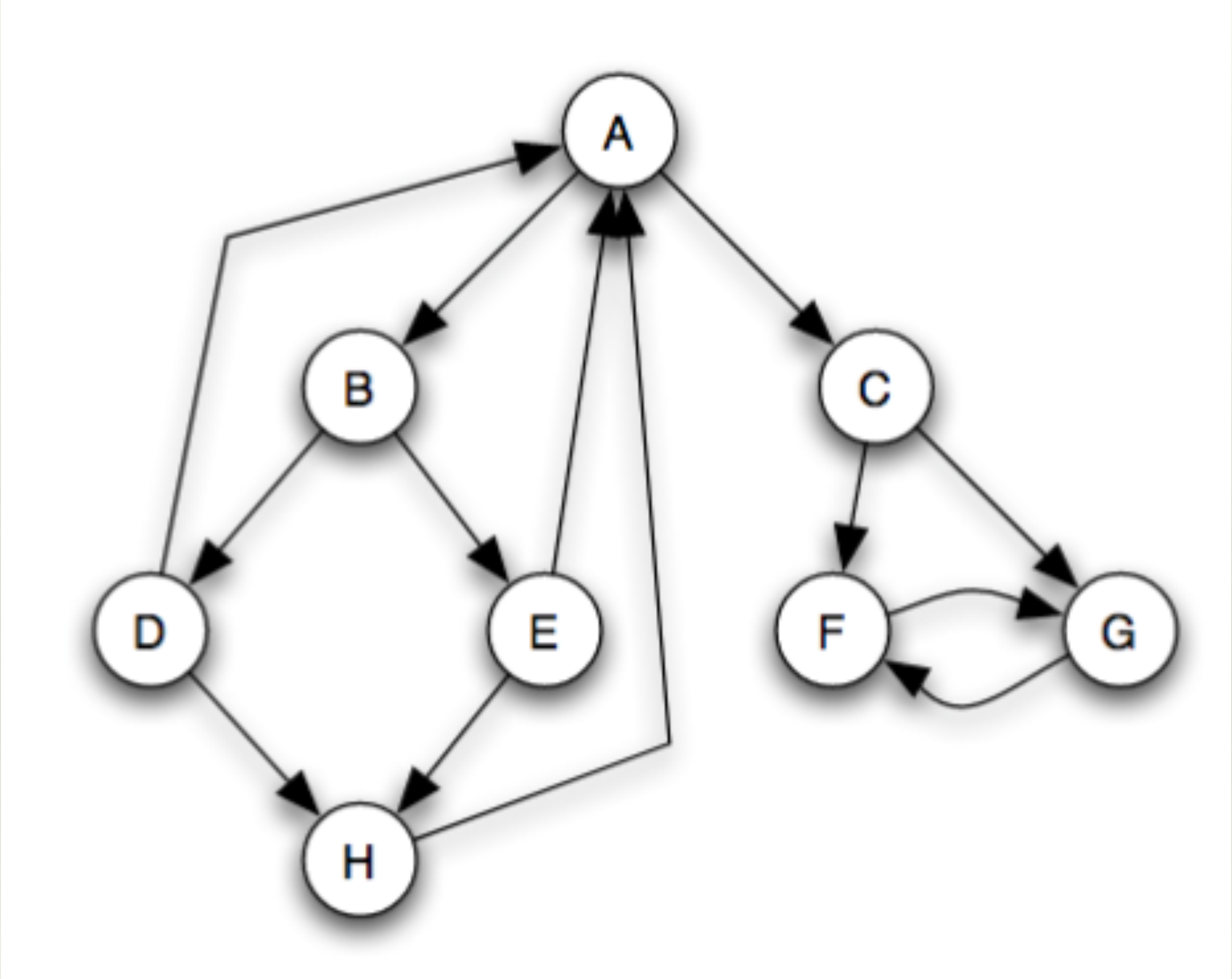
Note how at every step the total PageRank for all pages sums to 1

Equilibrium

Equilibrium is reached when the difference in the PageRank at every node changes very little between step k-1 and step k



What happens here?



All of the PageRank gets trapped in nodes F and G

We fix this by doing the PageRank update as usual, but afterwards, scale down the PageRank at each node by a factor s and then add (1-s)/n PageRank to every node

Teleporting some of the PageRank. In practice we usually have s=0.8 or 0.9

Probabilistic interpretation

1/(number of outlinks from page i) if link i -> j exists, 0 otherwise

$$P(X_k = j) = \sum_i P(X_k = j | X_{k-1} = i) P(X_{k-1} = i)$$

Amount of PageRank in page j after k steps

Amount of PageRank in page i after k-1 steps

So if we define

$$\pi = \left[\tfrac{1}{n} \quad \tfrac{1}{n} \quad \dots \quad \tfrac{1}{n}\right]$$

$$p_{ij} = \begin{cases} 1/\mathrm{deg}(i) & (i,j) \in G \\ 0 & \text{otherwise} \end{cases}$$

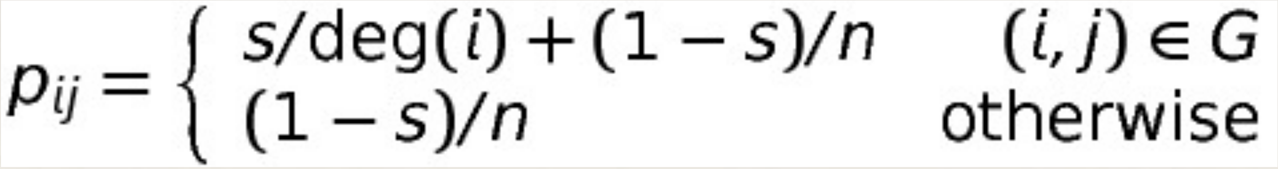
We get $\mathbf{p}^{(k)} = \pi P^k$

Probabilistic interpretation

This justifies our interpretation of PageRank vector as being the proportion of time a random web surfer would spend at each page

Teleportation in this interpretation captures the fact that you can always type in a new address in the address bar

With teleportation we have



$$p_{ij} = \begin{cases} s/\deg(i) + (1-s)/n & (i,j) \in G \\ (1-s)/n & \text{otherwise} \end{cases}$$

Probabilistic interpretation

The way we have defined things gives us a Markov chain

$$\pi = \begin{bmatrix} \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n} \end{bmatrix} \qquad p_{ij} = \begin{cases} s/\deg(i) + (1-s)/n & (i,j) \in G \\ (1-s)/n & \text{otherwise} \end{cases}$$

Furthermore it gives an irreducible Markov chain since there is a path from every state to every other state. This means the Markov chain has a stationary distribution **v**

$$\lim_{n\to\infty} \pi P^{(n)} = \mathbf{v}$$

And entry *i* of this column vector gives the PageRank of web page *i*