Link Analysis

Introduction to Information Retrieval
INF 141
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org

Key Observation

- A citation in scientific literature is like a link on the web

Link Analysis

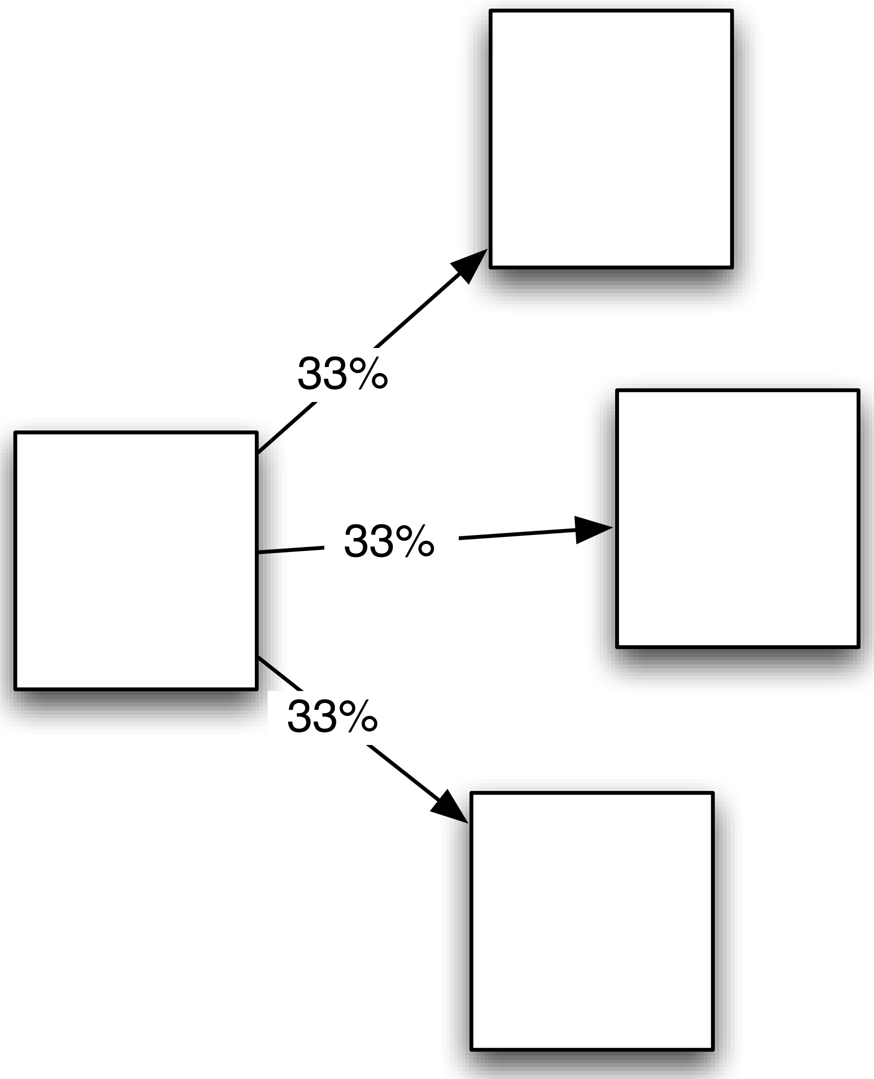
- A full search engine ranks based on many different scores
  - Cosine similarity
  - Term proximity
  - Zone scoring
  - Contextual relevance (implicit queries)
  - Link analysis

Link based query ranking

- Retrieve all pages meeting the query
  - First generation:
    - Then order them by their link popularity
      - citation frequency
    - Easy to spam. Why?
  - Second generation:
    - Order them by their weighted link popularity
      - PageRank

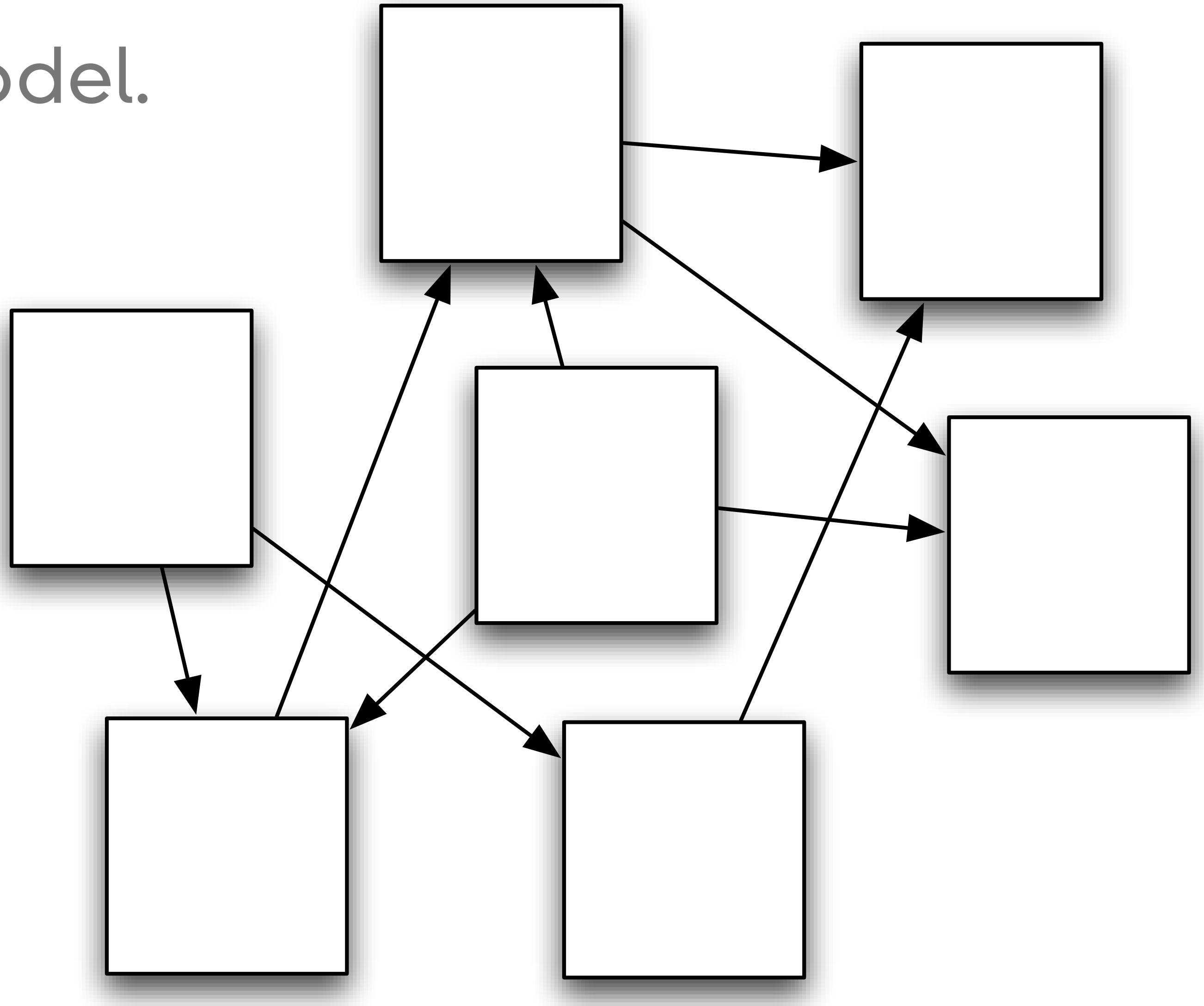
PageRank

- Every webpage gets a score
  - between 0 and 1
  - it's PageRank
- The random walk
  - Start at a random page
  - Follow an out edge with equal probability
- In the long run each page has a long-term visit rate.



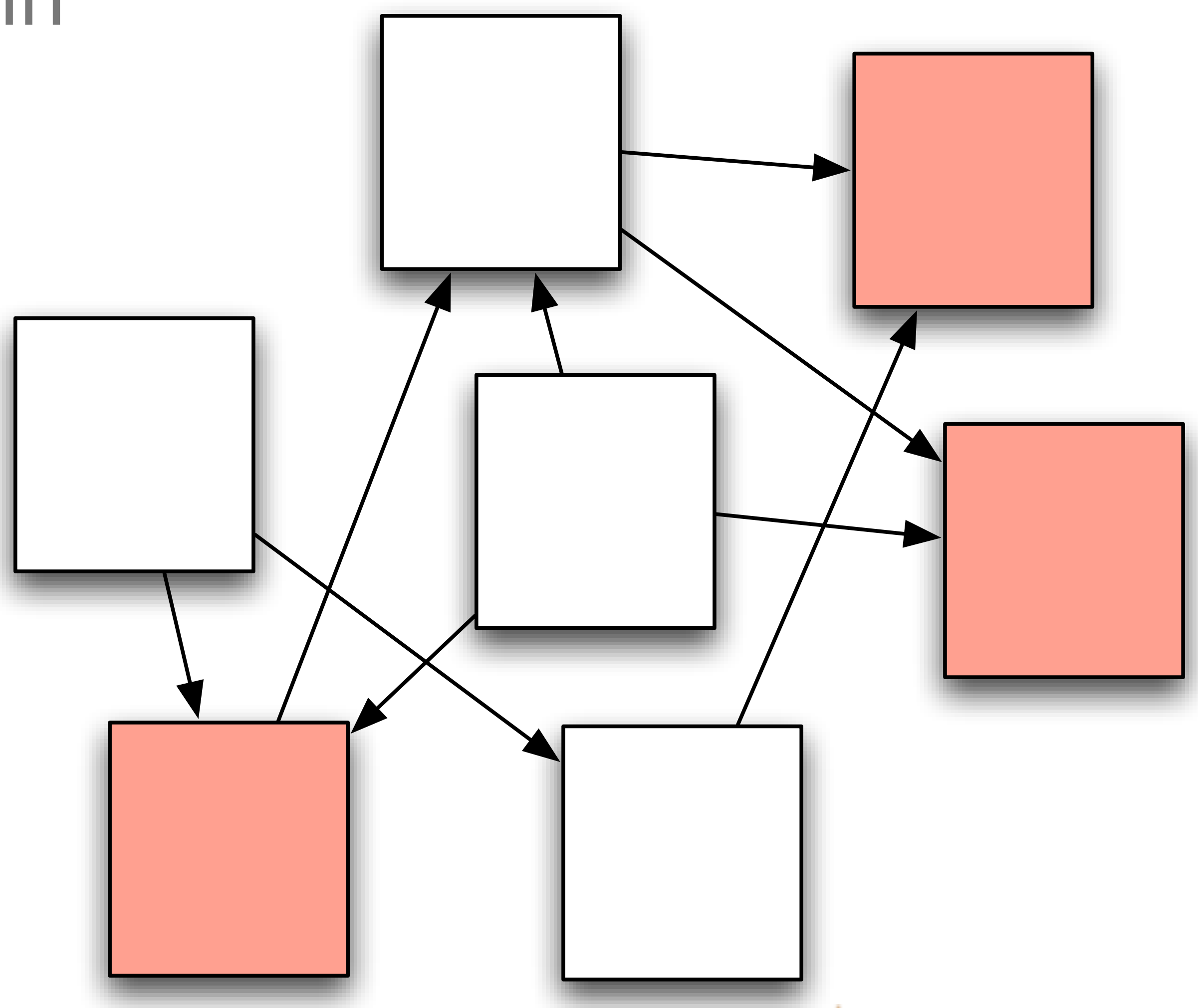
PageRank

- PageRank is a page's long-term steady state visit rate based on a random walk model.



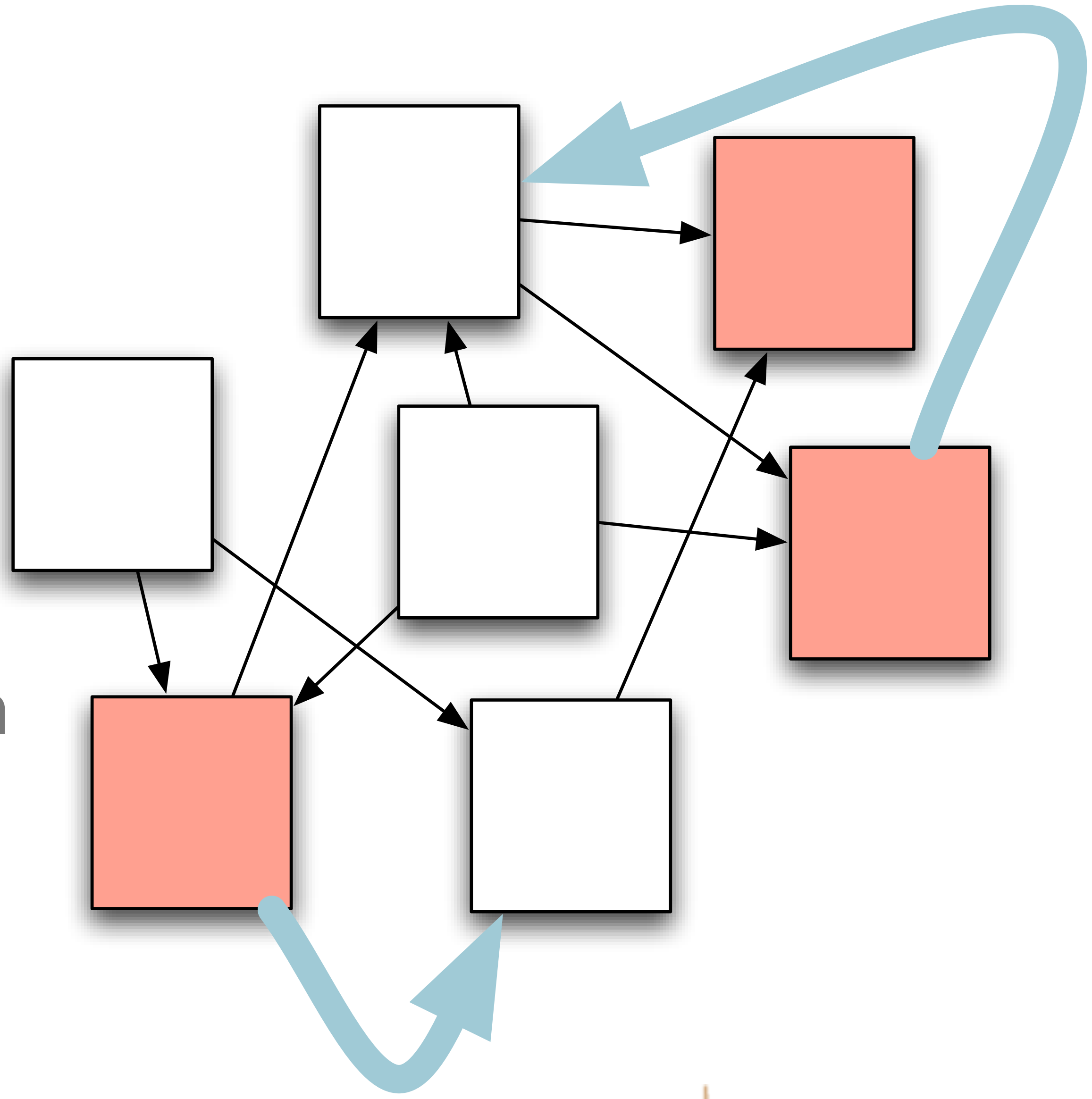
Visit Rate not quite enough

- The web is full of dead-ends
    - A random walk can get stuck in dead-ends
    - Makes no sense to talk about long-term visit rates



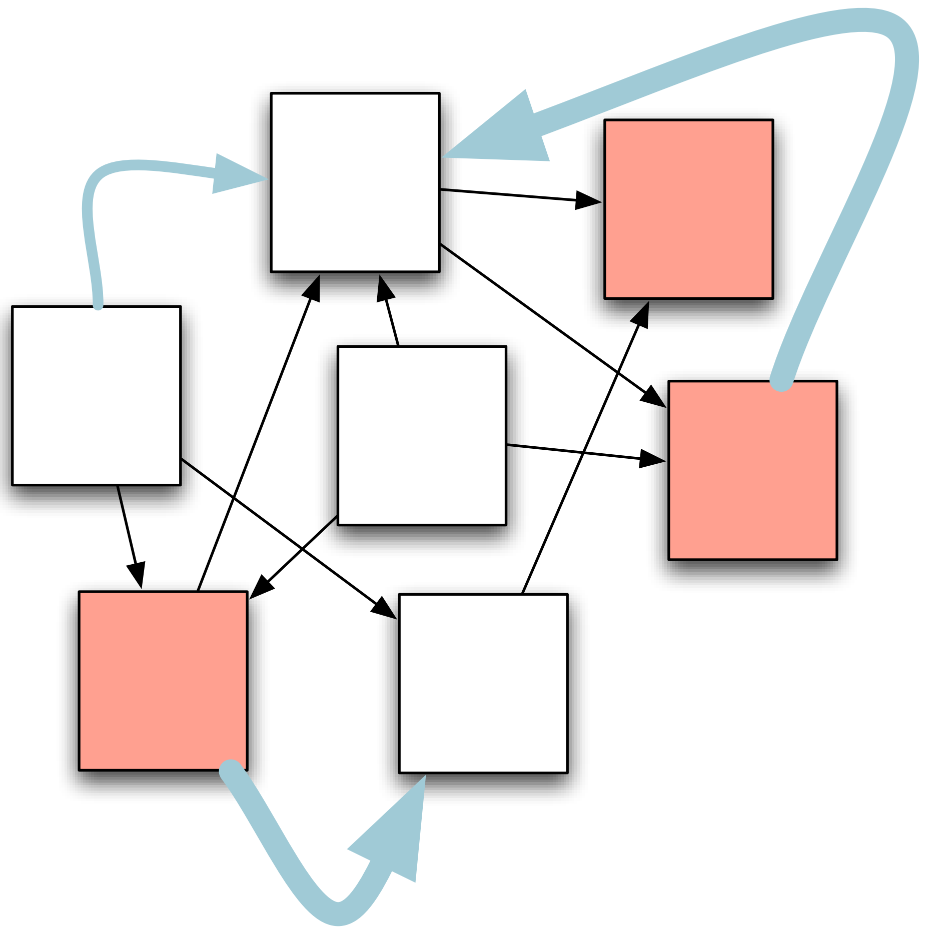
Teleporting

- At a dead end, jump to a random web page
  - at any non-dead end, with probability 10% jump to a random web page anyway
  - the other 90% choose a random out link
  - "10%" is a tunable parameter



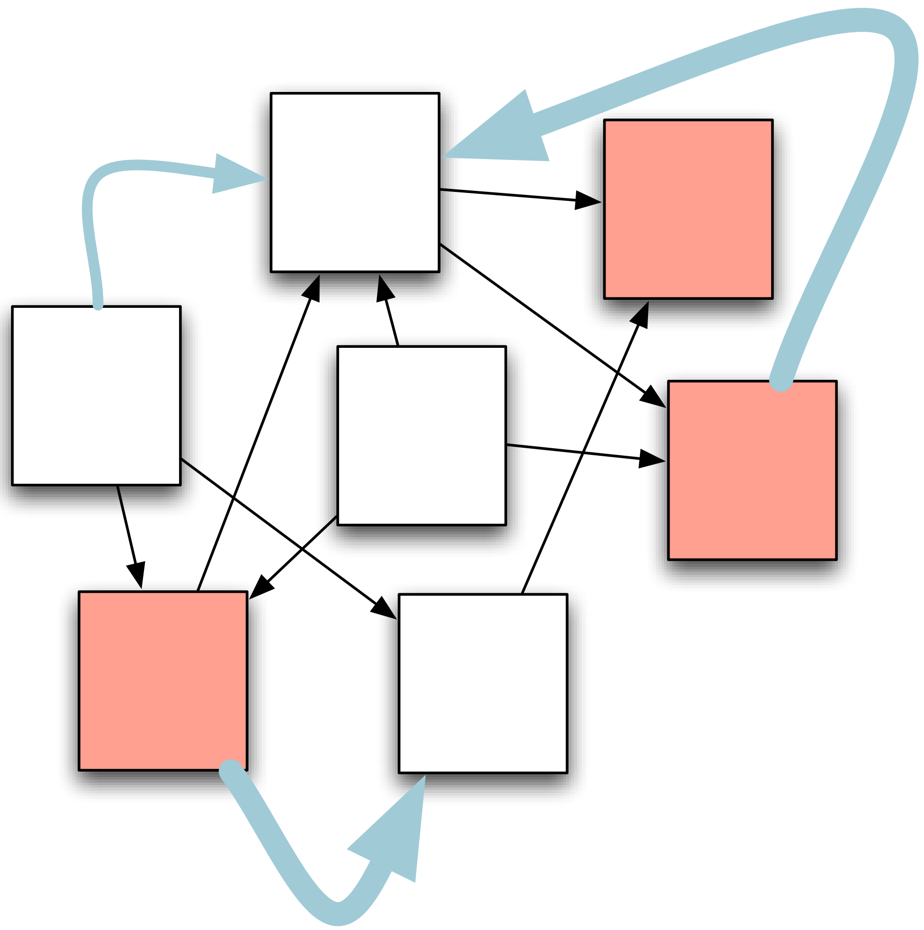
Teleporting

- Now we cannot get stuck locally
- There is a long-term visit rate at which any page is visited.
- How do we compute the visit rate?
    - How do we compute PageRank?
- (By the way this is a Markov Chain)



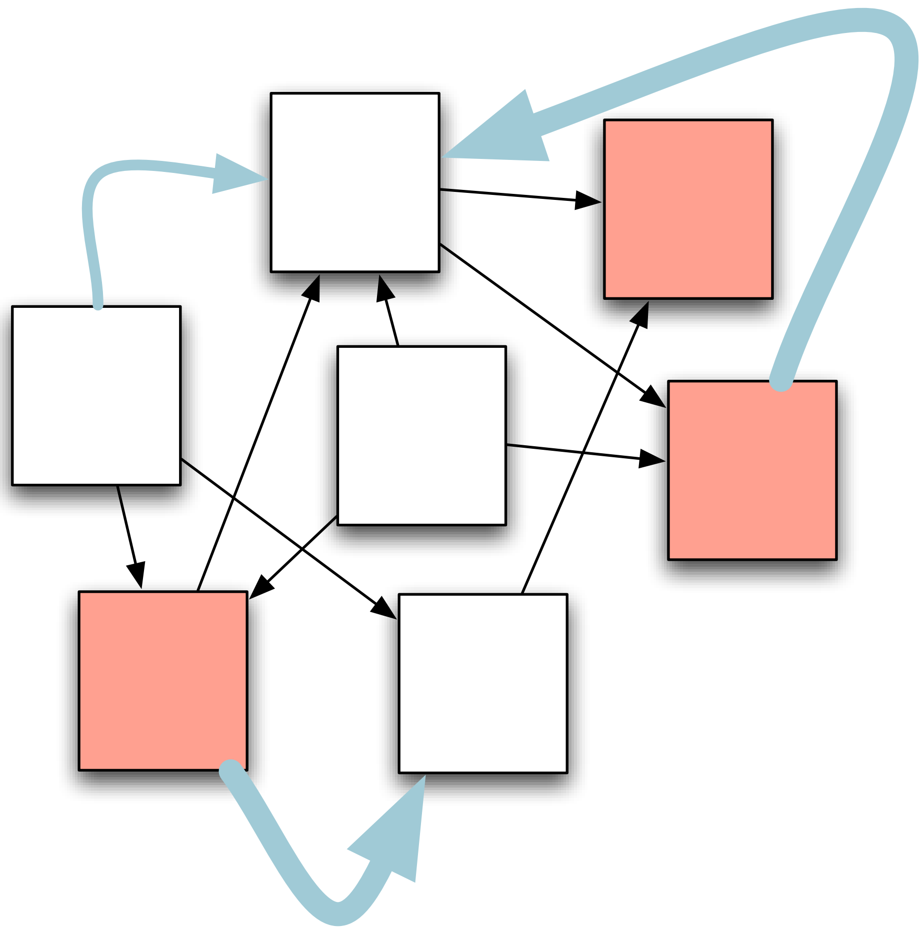
Markov Chains

- A Markov Chain is a mathematical "game"
- It consists of n states
  - corresponds to web pages
- And a transition probability matrix
  - corresponds to links
  - it is like an adjacency matrix



Markov Chains

- At any moment in the game we are in one of the states
- In the next step we move to a new state
- We use the transition matrix to decide which state to move into.
- If you are in state "i" then the probability of moving into state "j" is P(i -> j)

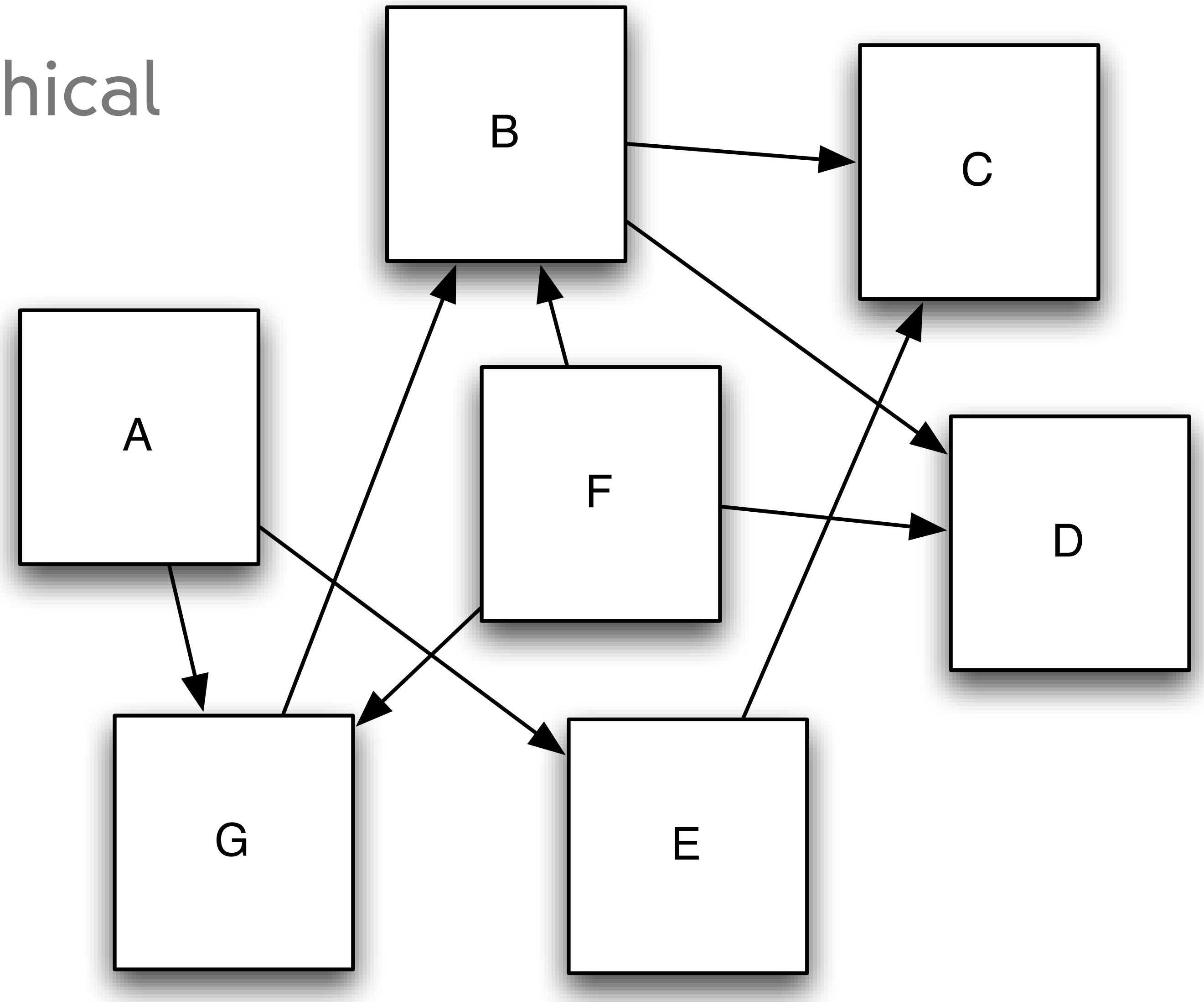


Markov Chains

- Markov Chains are described by two parameters:
  - A list of n states
  - An (n by n) transition probability table
  - It's like a graph, except that links aren't boolean, they are real numbers.
    - A link doesn't just exist or not exist
    - It exists with a probability also

Exercise

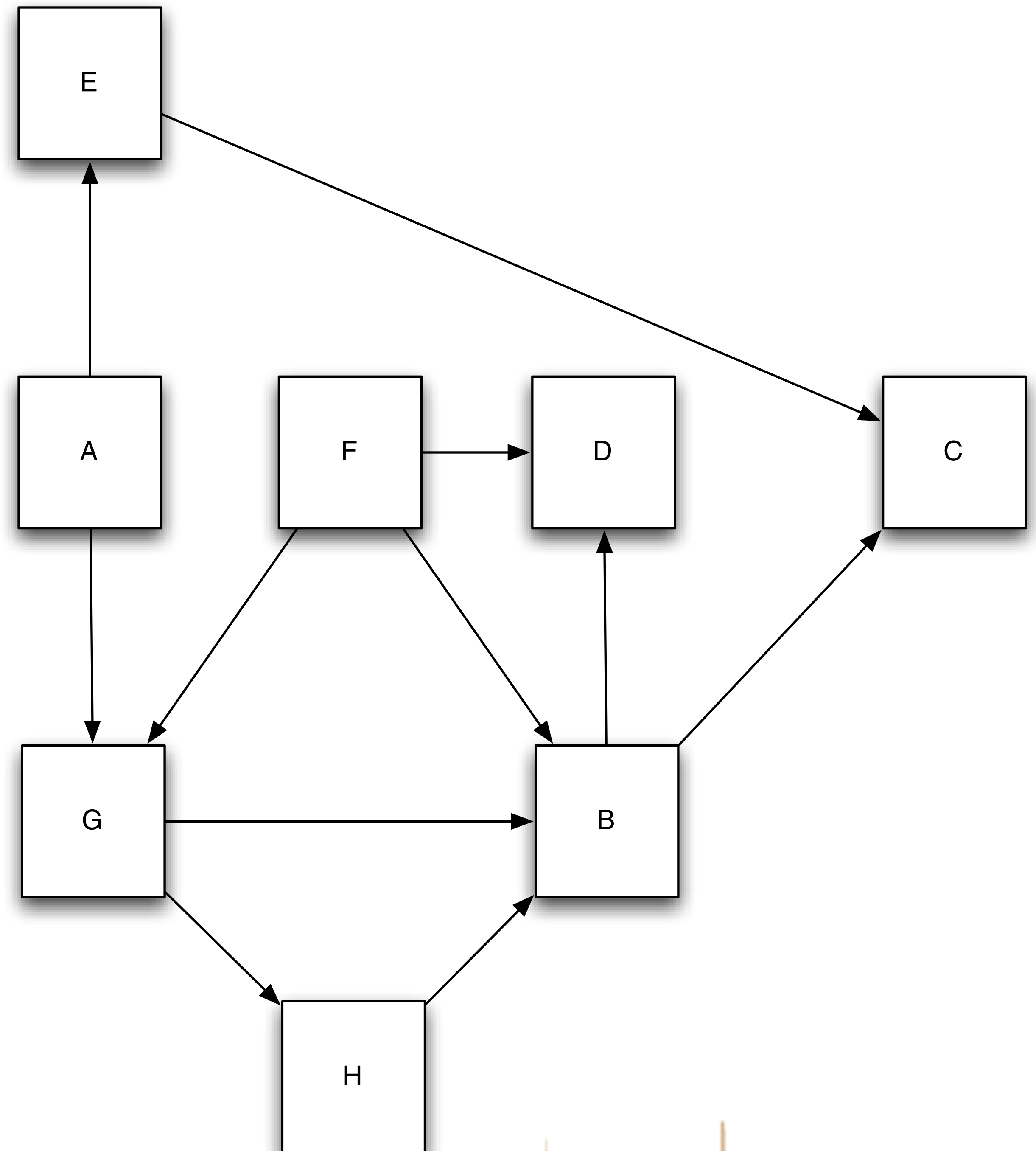
- Compute the parameters of the Markov Chain for this graphical model



Markov Chains

- Example:
  - 8 states
    - (web pages or whatever)
  - 8 by 8 transition prob. matrix

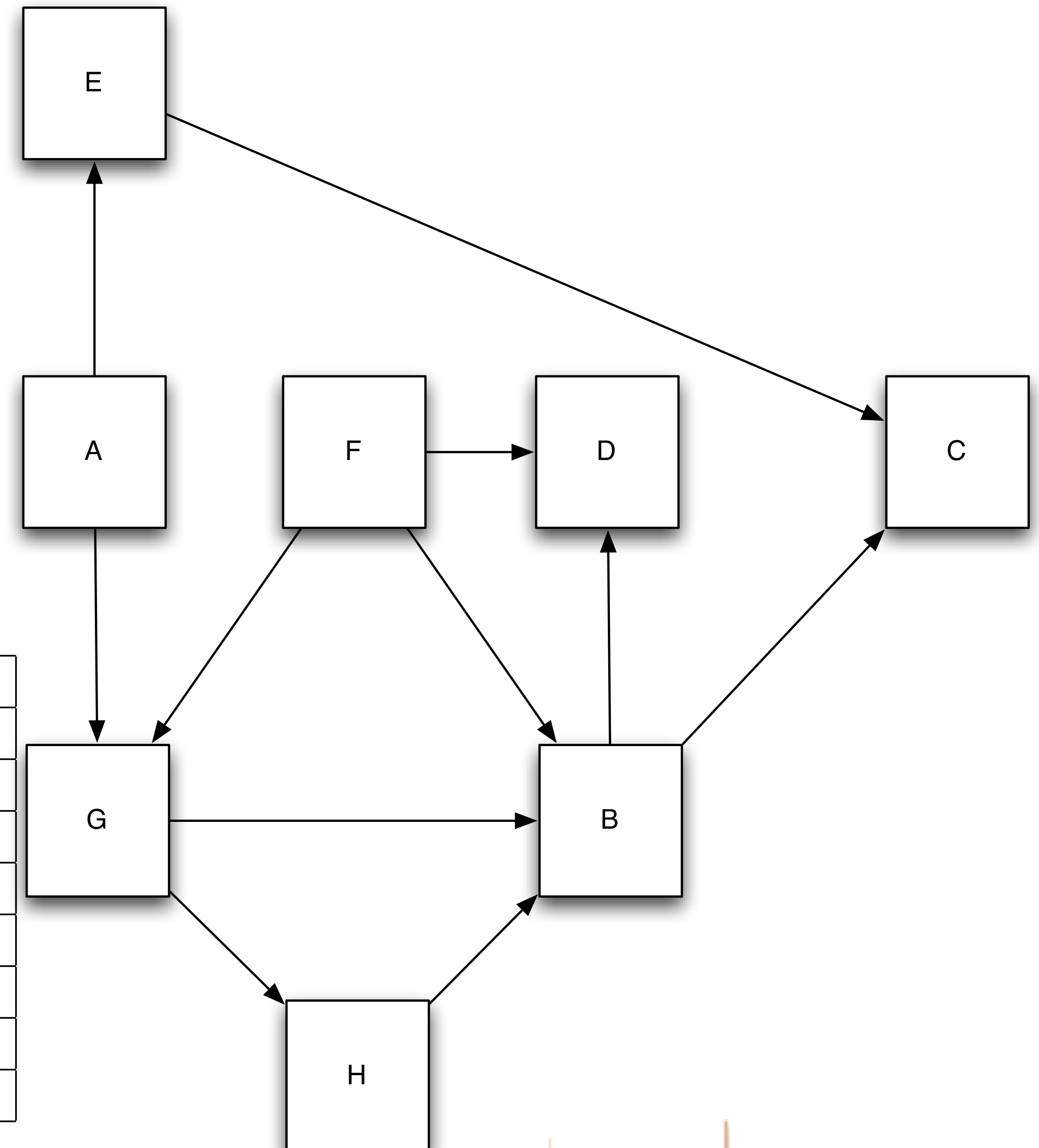
	A	B	C	D	E	F	G	H
A	0	0	0	0	0.5	0	0.5	0
B	0	0	0.5	0.5	0	0	0	0
C	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0
E	0	0	1.0	0	0	0	0	0
F	0	0.33	0	0.33	0	0	0.33	0
G	0	0.5	0	0	0	0	0	0.5
H	0	1.0	0	0	0	0	0	0



Markov Chains

- Example:
  - 8 states
  - 8 by 8 transition prob. matrix
  - Handle Dead-Ends also

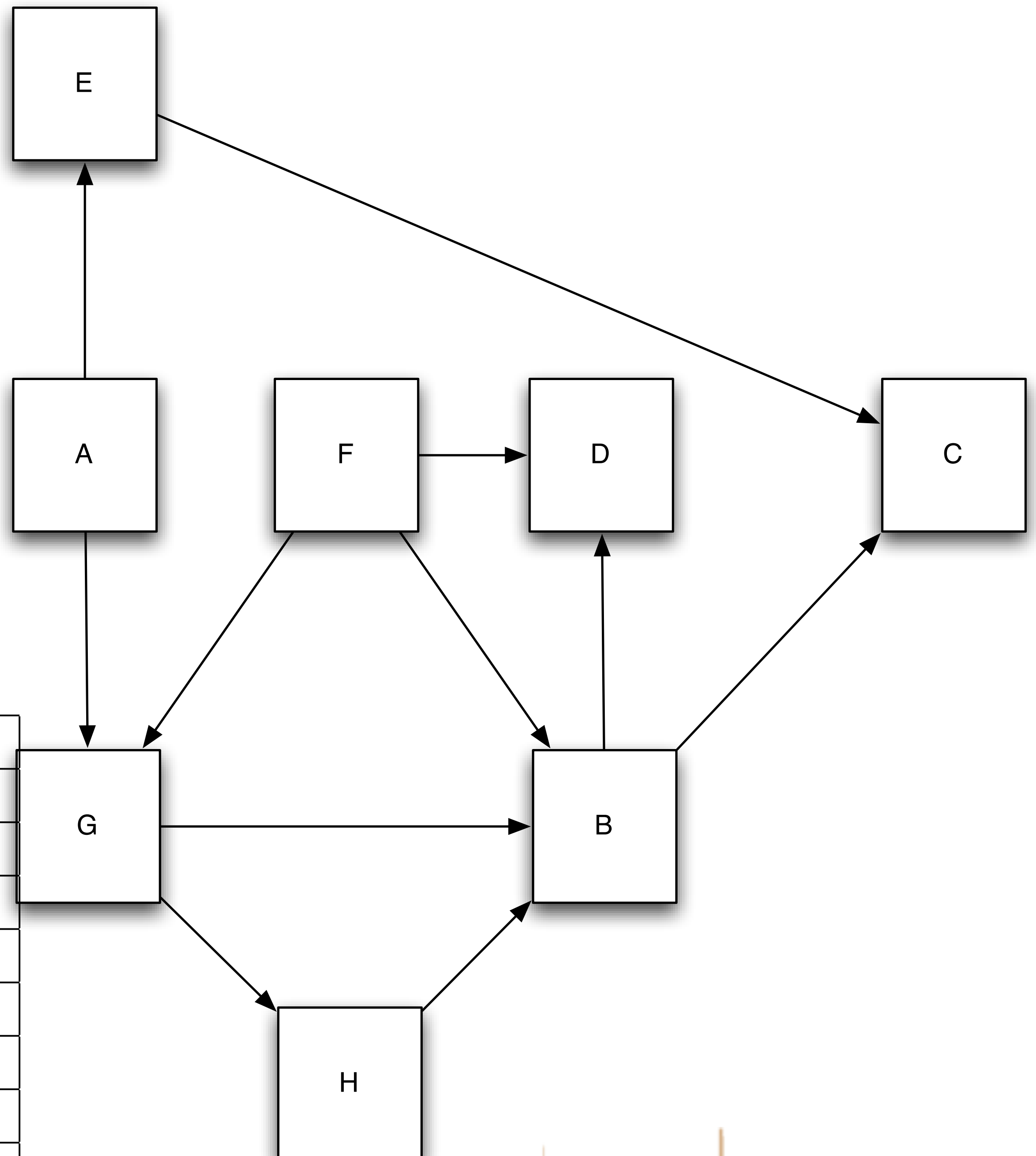
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0



Markov Chains

- Example:
  - 8 states
  - 8 by 8 transition prob. matrix
  - Handle Dead-Ends also
  - Handle teleports

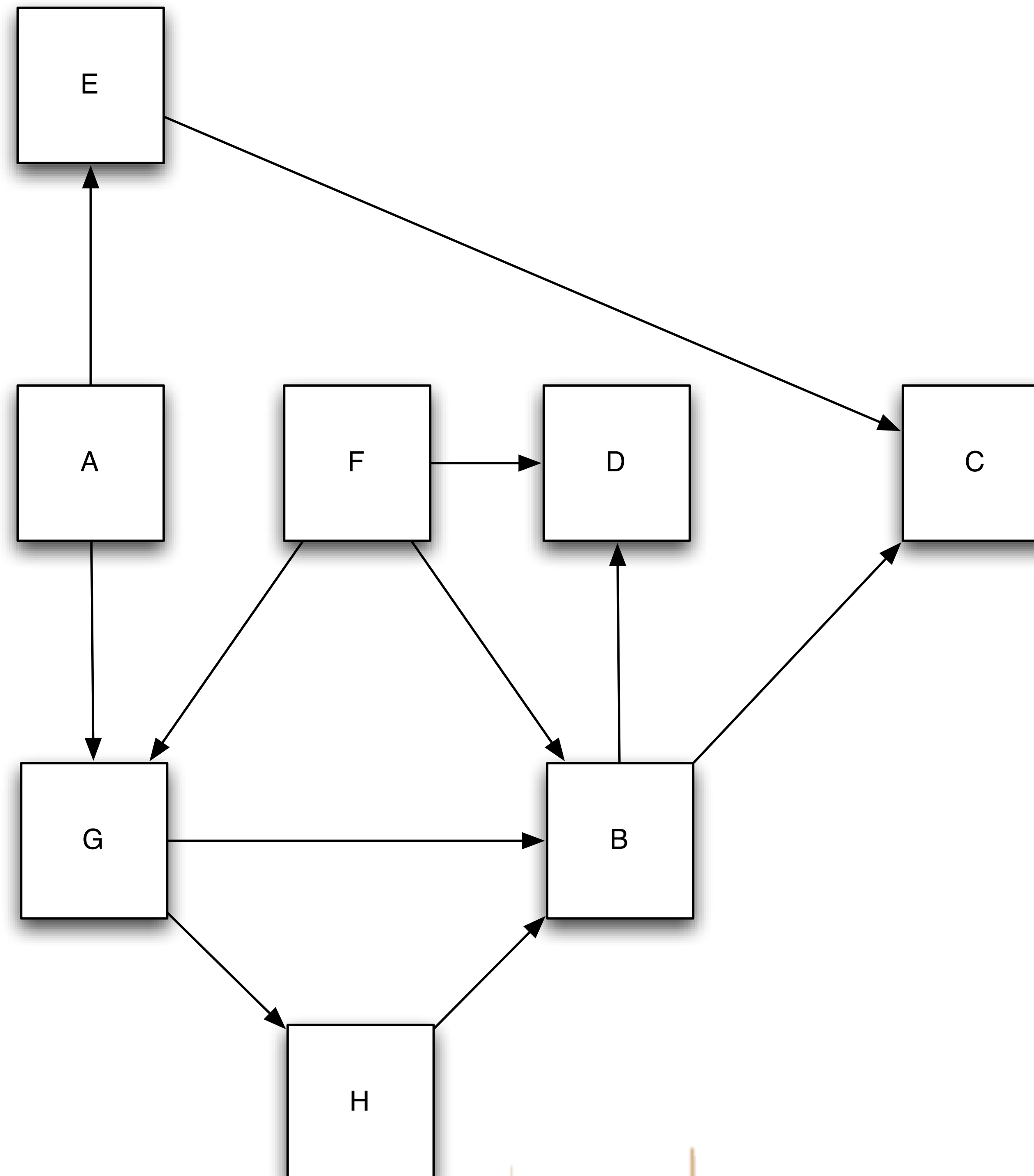
	A	B	C	D	E	F	G	H
A	0.01	0.01	0.01	0.01	0.47	0.01	0.47	0.01
B	0.01	0.01	0.47	0.47	0.01	0.01	0.01	0.01
C	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
D	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
E	0.01	0.01	0.93	0.01	0.01	0.01	0.01	0.01
F	0.01	0.32	0.01	0.32	0.01	0.01	0.32	0.01
G	0.01	0.47	0.01	0.01	0.01	0.01	0.01	0.47
H	0.01	0.93	0.01	0.01	0.01	0.01	0.01	0.01



Markov Chain : The Game

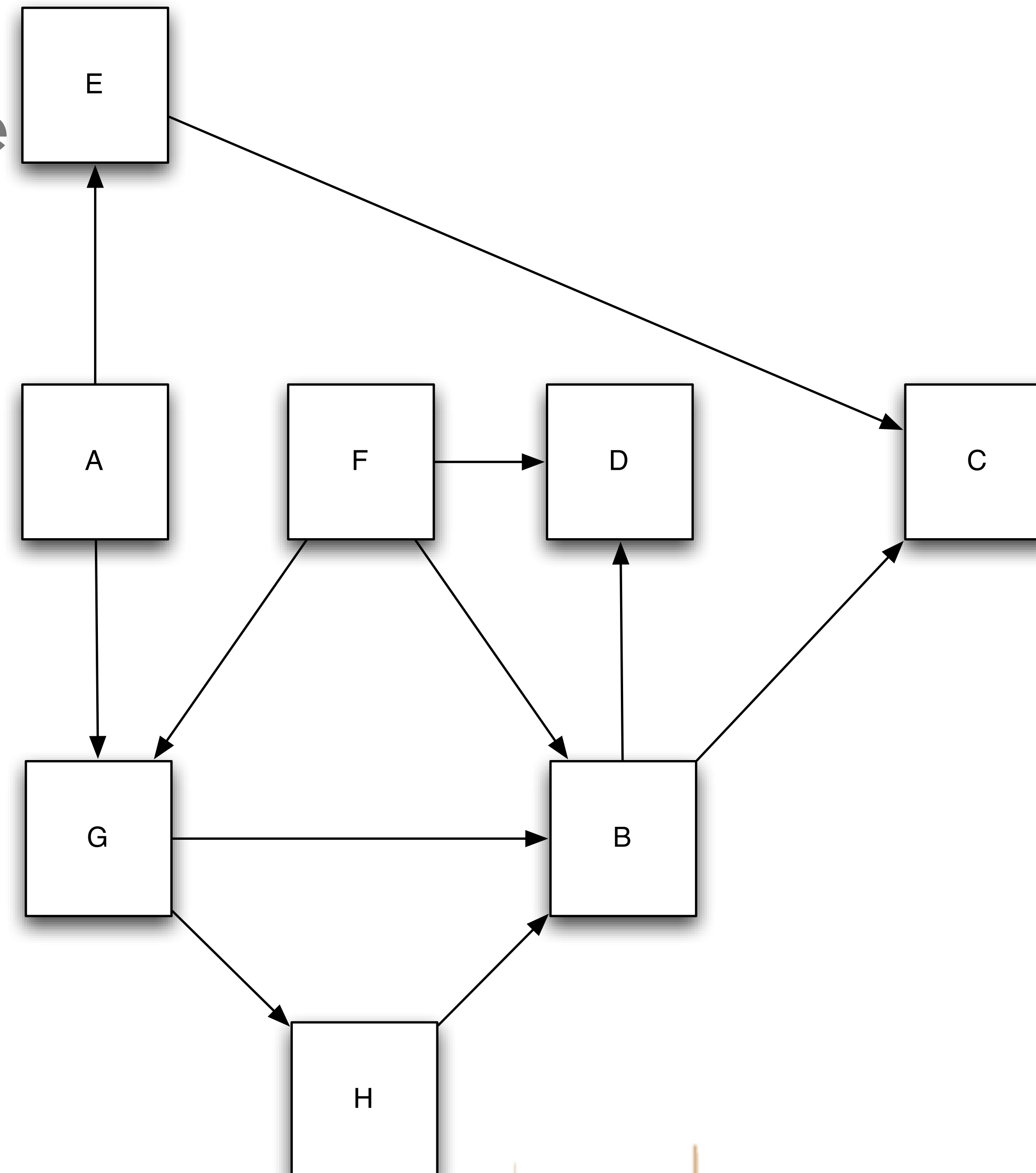
- You may be in one state at a time
- Every tick you move one step chosen randomly from the the transition probability matrix

	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0



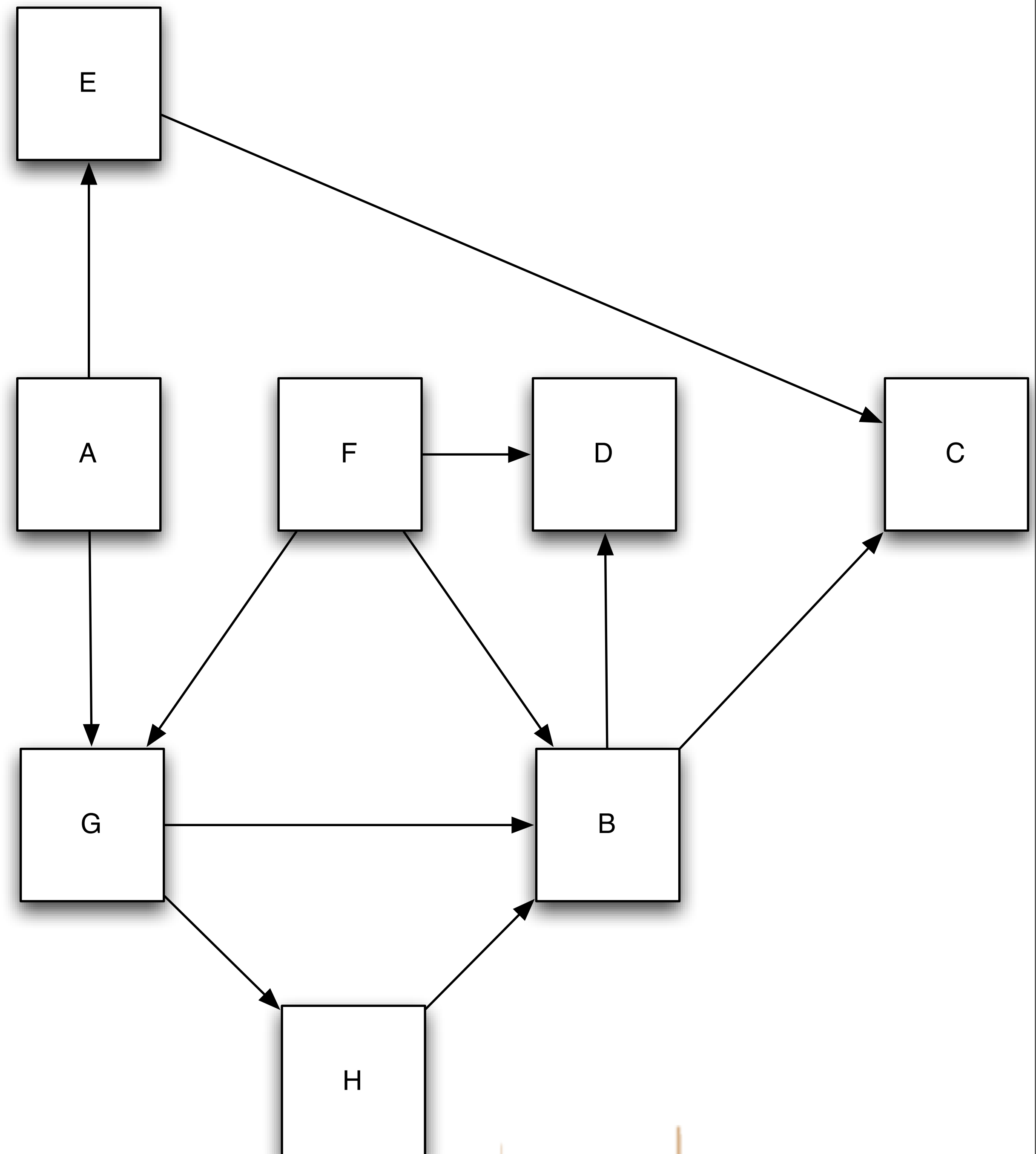
The Markov Property

- It doesn't matter where you came from.
- All information that you need to take the next step comes from your current state and the transition probability matrix
- History is irrelevant given your current state

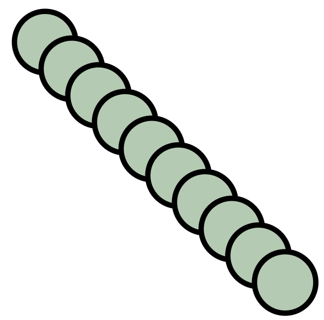


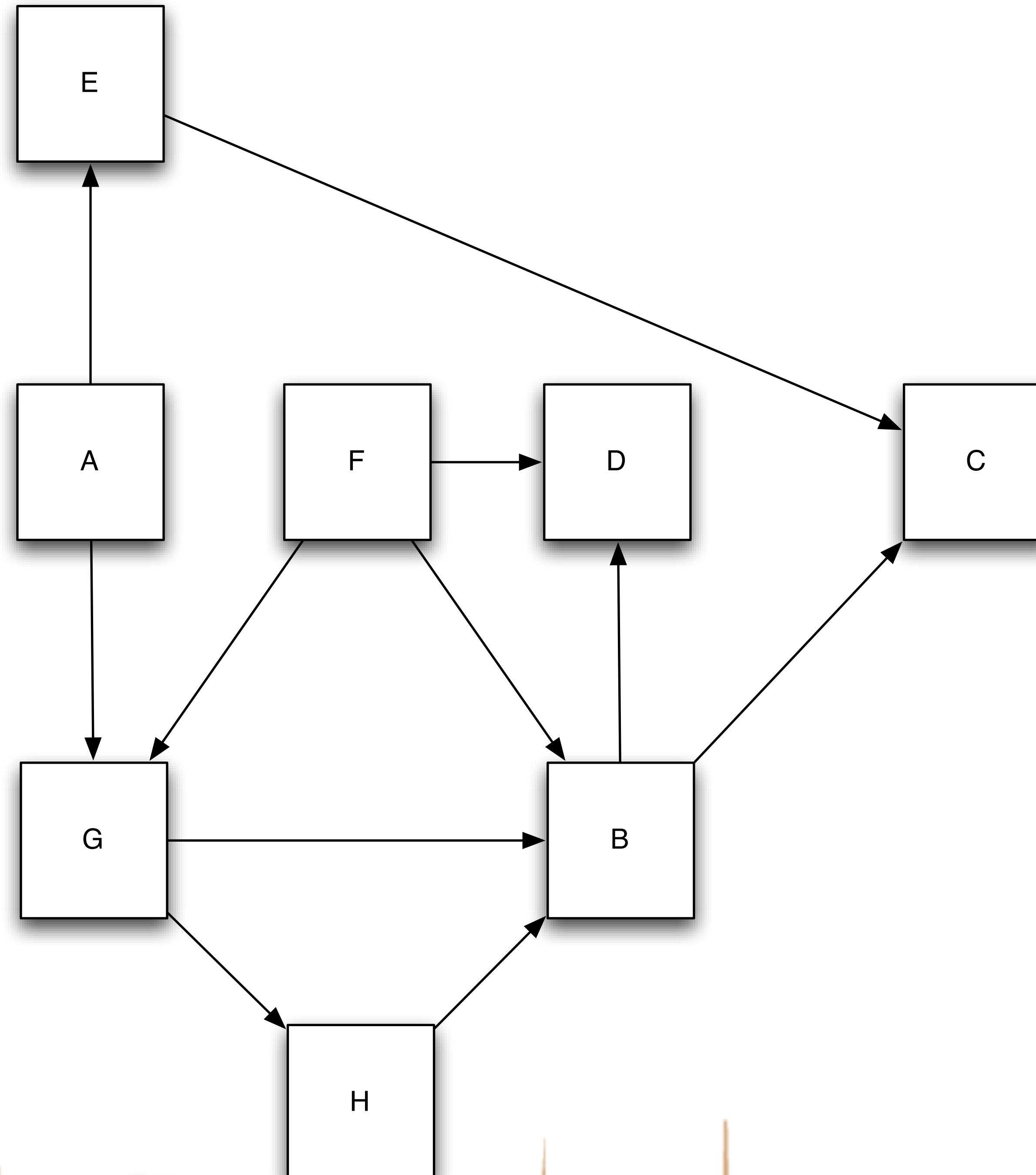
PageRank

- PageRank is the long term visit rate of a random walk on the graph.
- With teleports



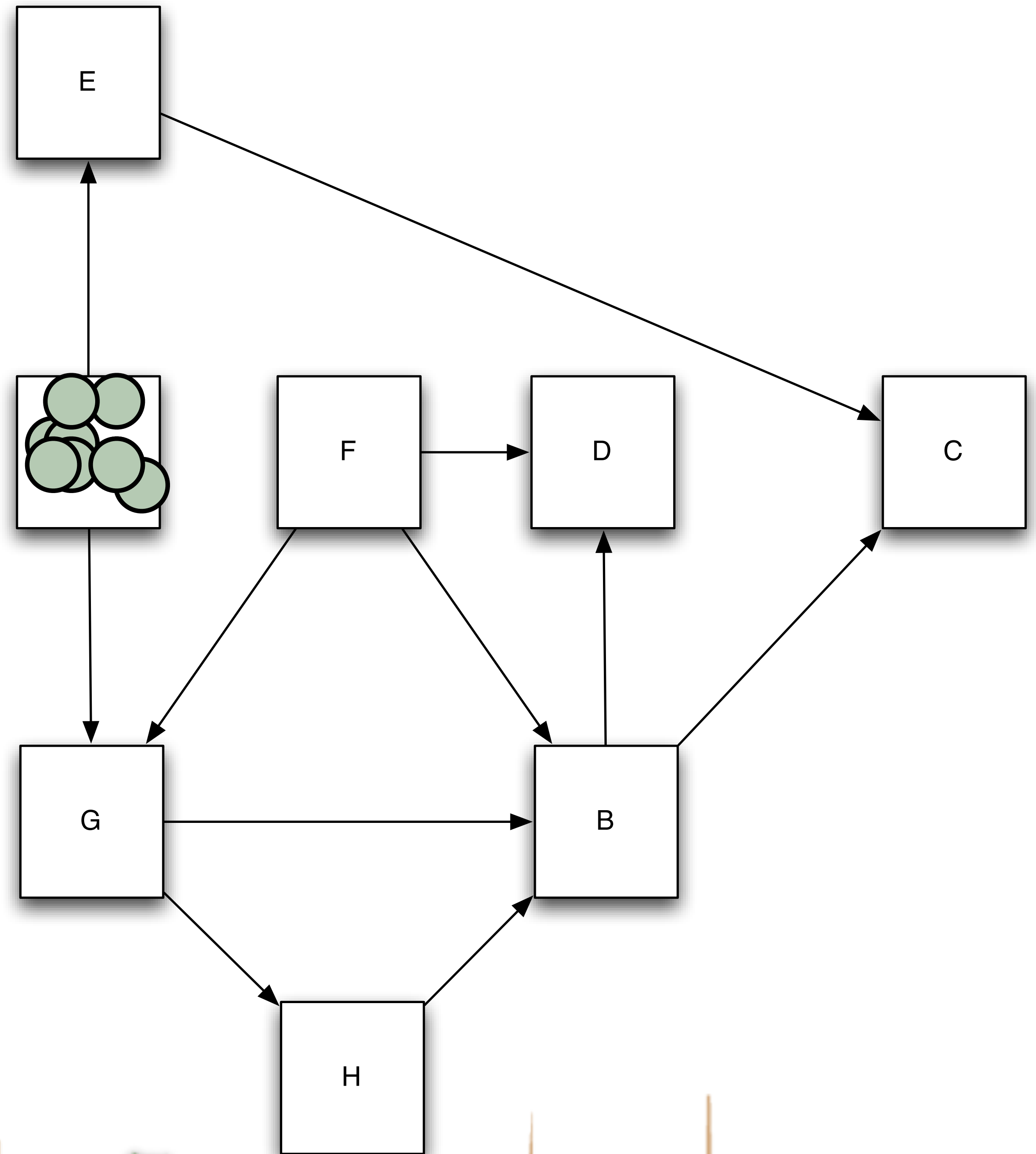
Long-Term visit rate





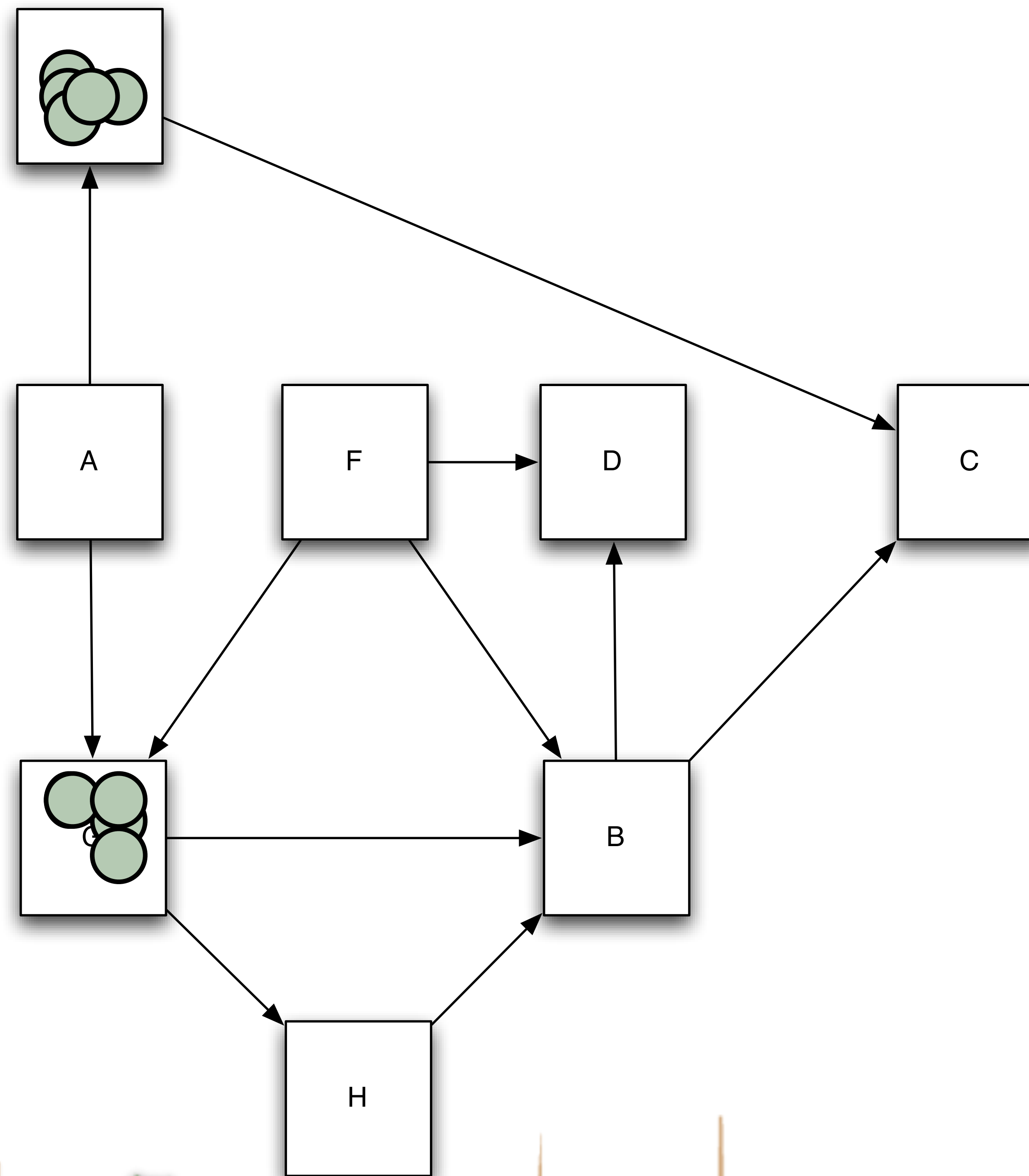
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

Long-Term visit rate



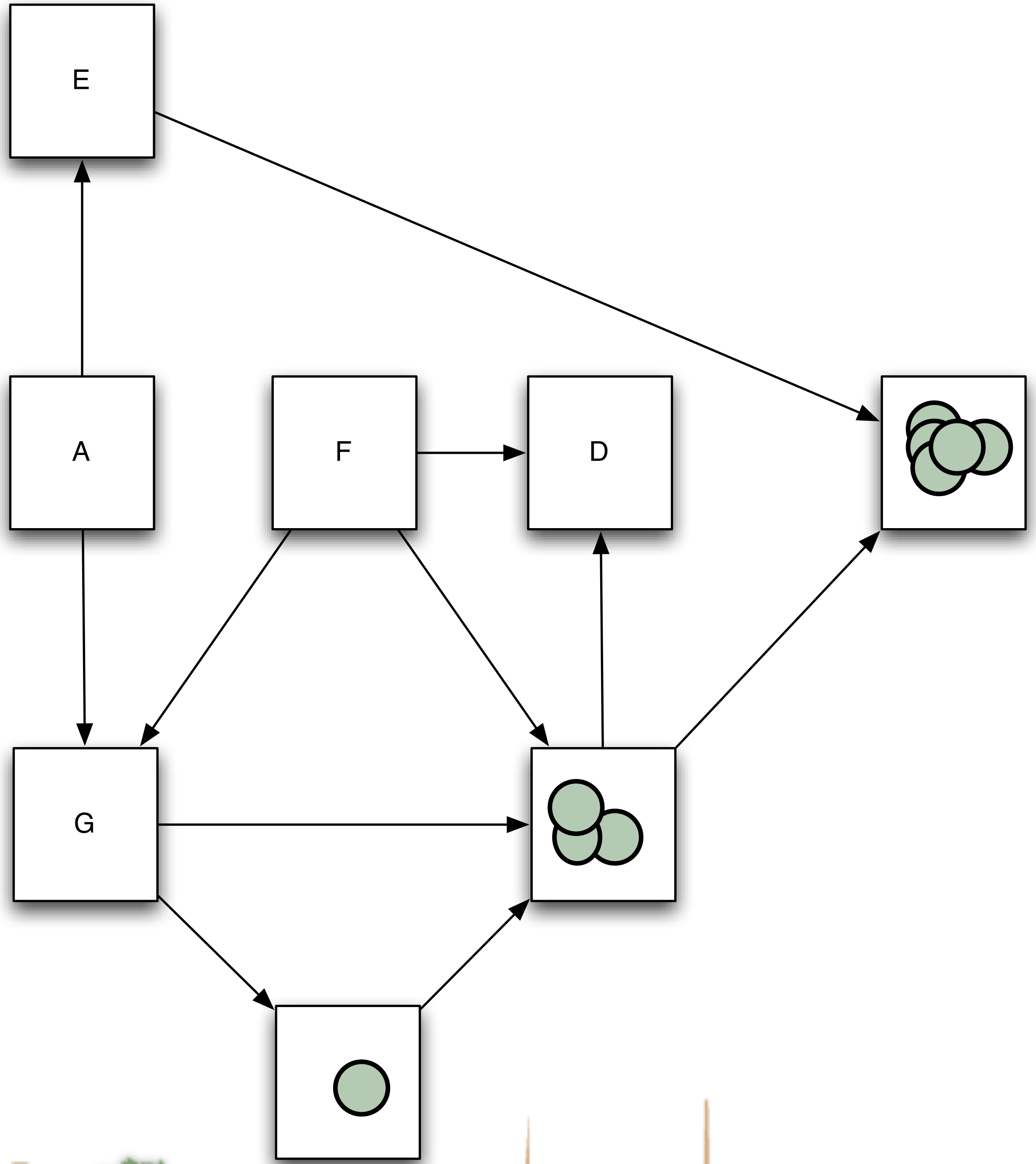
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

Long-Term visit rate



	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

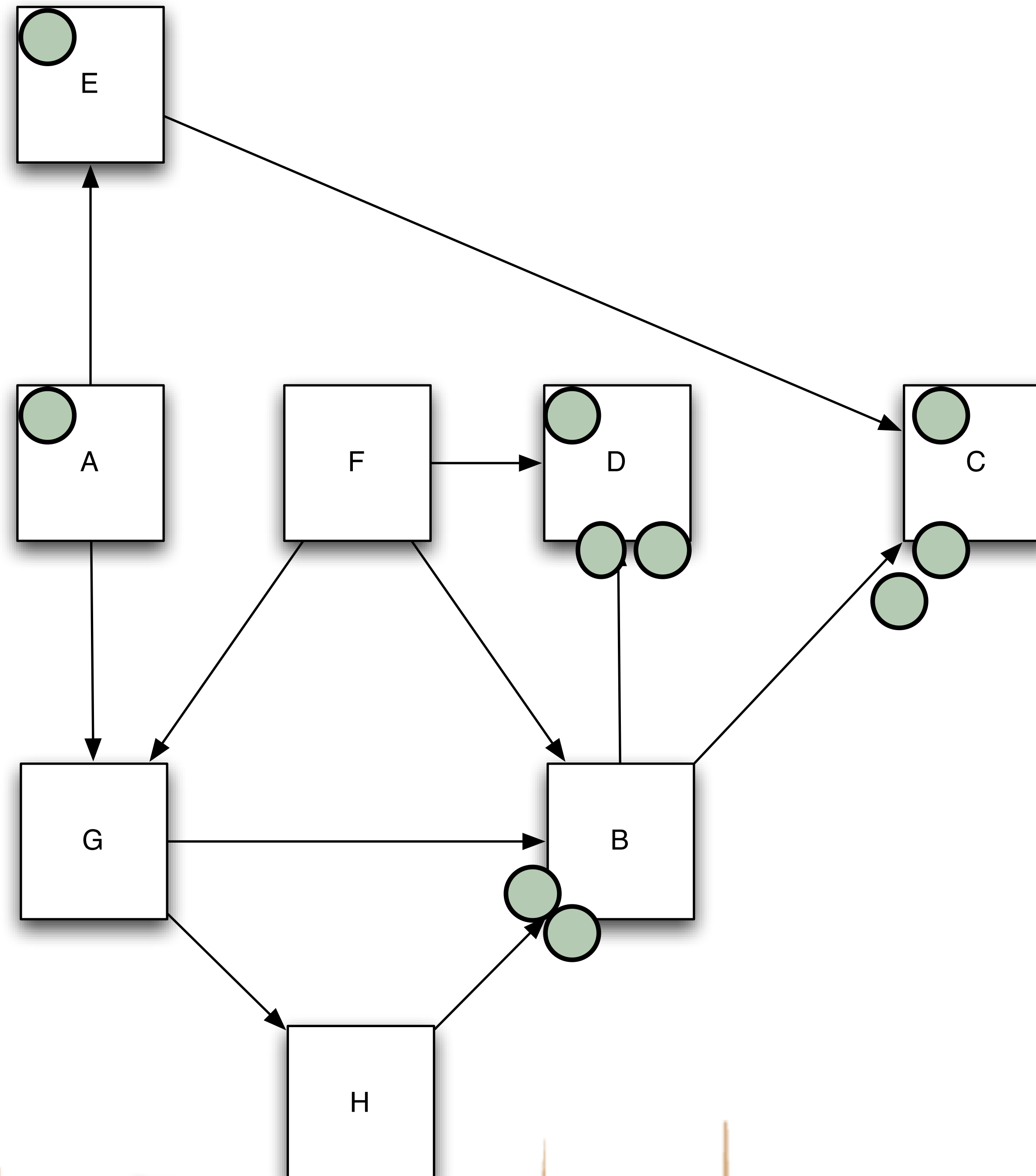
Long-Term visit rate



	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

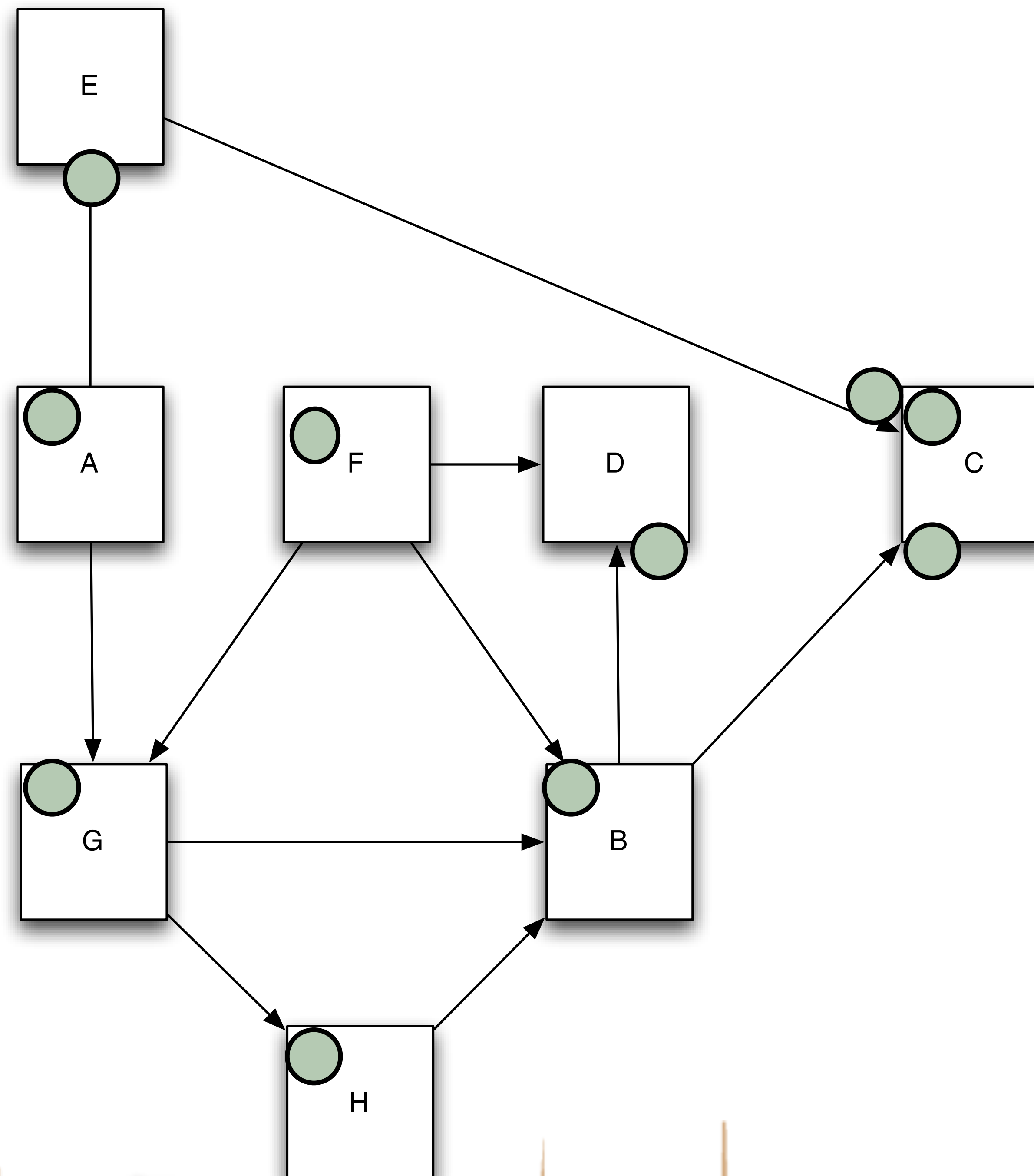
Long-Term visit rate

	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0



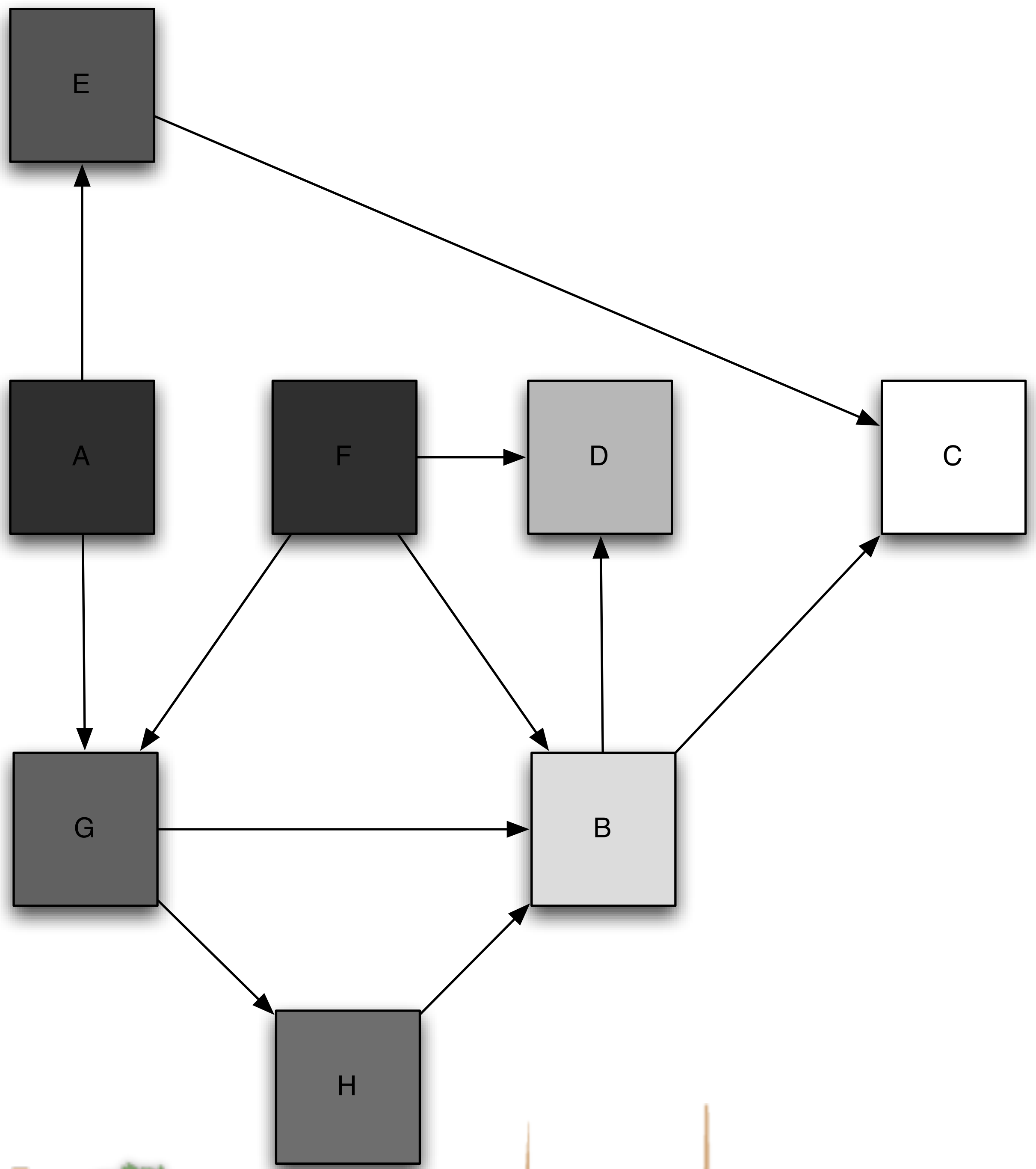
Long-Term visit rate

	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0



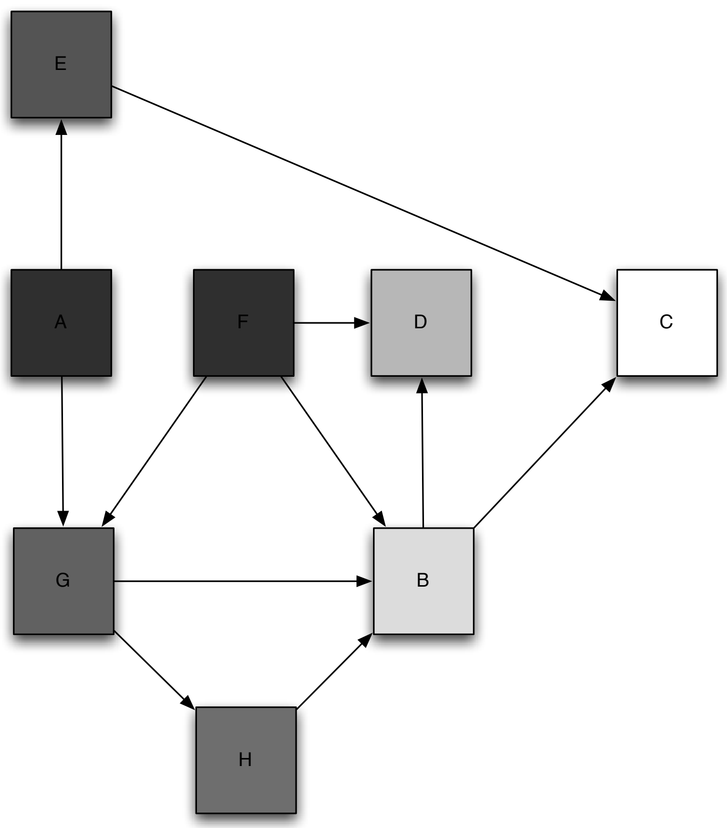
Long-Term visit rate

- A: 5%
- B: 21%
- C: 23%
- D: 18%
- E: 8%
- F: 5%
- G: 9%
- H: 10%



Some properties of Markov chains

- Ergodic:
  - All states can reach all states
  - What did we have to do to enable this for a web graph?
- Steady State Theorom:
  - Every ergodic markov chain has a steady state -> has a PageRank

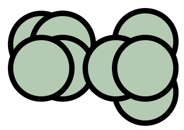


Calculating PageRank

- Visual representation to math representation

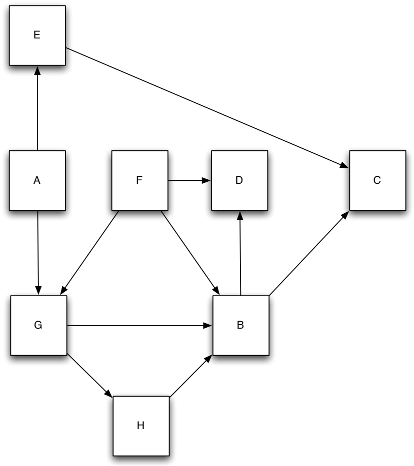
$\vec{x_0}$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0



P

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
[illegible]	0	0	1 0	0	0	0	0	0
[illegible]	0	0 33	0	0 33	0	0	0 33	0
[illegible]	0	0 5	0	0	0	0	0	0 5
[illegible]	0	1 0	0	0	0	0	0	0



Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0}P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

0	0	0	0	0 5	0	0 5	0
0	0	0 5	0 5	0	0	0	0
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0	0	1 0	0	0	0	0	0
0	0 33	0	0 33	0	0	0 33	0
0	0 5	0	0	0	0	0	0 5
0	1 0	0	0	0	0	0	0

0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0}P = \vec{x_1}$$

0	0	0	0	0 5	0	0 5	0
0	0	0 5	0 5	0	0	0	0
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0	0	1 0	0	0	0	0	0
0	0 33	0	0 33	0	0	0 33	0
0	0 5	0	0	0	0	0	0 5
0	1 0	0	0	0	0	0	0

0	0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

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

	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

0	0	0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

0	0	0	0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

0	0	0	0	0 5	0	0 5	0
0	0	0 5	0 5	0	0	0	0
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0	0	1 0	0	0	0	0	0
0	0 33	0	0 33	0	0	0 33	0
0	0 5	0	0	0	0	0	0 5
0	1 0	0	0	0	0	0	0

0	0	0	0	0.5

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

0	0	0	0	0.5	0.0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

[illegible]								
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

0	0	0	0	0.5	0.0	0.5

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0} P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

0	0	0	0	0 5	0	0 5	0
0	0	0 5	0 5	0	0	0	0
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0	0	1 0	0	0	0	0	0
0	0 33	0	0 33	0	0	0 33	0
0	0 5	0	0	0	0	0	0 5
0	1 0	0	0	0	0	0	0

0	0	0	0	0.5	0.0	0.5	0.0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_0}P = \vec{x_1}$$

A	B	C	D	E	F	G	H
1.0	0	0	0	0	0	0	0

0	0	0	0	0 5	0	0 5	0
0	0	0 5	0 5	0	0	0	0
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
0	0	1 0	0	0	0	0	0
0	0 33	0	0 33	0	0	0 33	0
0	0 5	0	0	0	0	0	0 5
0	1 0	0	0	0	0	0	0

$\vec{x_1} =$

0	0	0	0	0.5	0	0.5	0

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_1} P = \vec{x_2}$$

0	0	0	0	0.5	0	0.5	0

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

$\vec{x_2} =$

0	0.25	0.5	0	0.0	0	0.0	0.25

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_1} P = \vec{x_2}$$

0	0.25	0.5	0	0.0	0	0.0	0.25

	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	0	0	0	0	0 5	0	0 5	0
	0	0	0 5	0 5	0	0	0	0
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0 125	0 125	0 125	0 125	0 125	0 125	0 125	0 125
	0	0	1 0	0	0	0	0	0
	0	0 33	0	0 33	0	0	0 33	0
	0	0 5	0	0	0	0	0	0 5
	0	1 0	0	0	0	0	0	0

$\vec{x_3} =$

0.0625	0.3125	0.1875	0.1875	0.0625	0.0625	0.0625	0.0625

Calculating PageRank

- Take one step is multiplying state vector times transition probability matrix

$$\vec{x_1}P = \vec{x_2}$$

$$lim_{(n \rightarrow \infty)} x_n = PageRank$$

Long-Term visit rate

- A: 5%
- B: 21%
- C: 23%
- D: 18%
- E: 8%
- F: 5%
- G: 9%
- H: 10%