Link Analysis

Introduction to Information Retrieval
CS 221

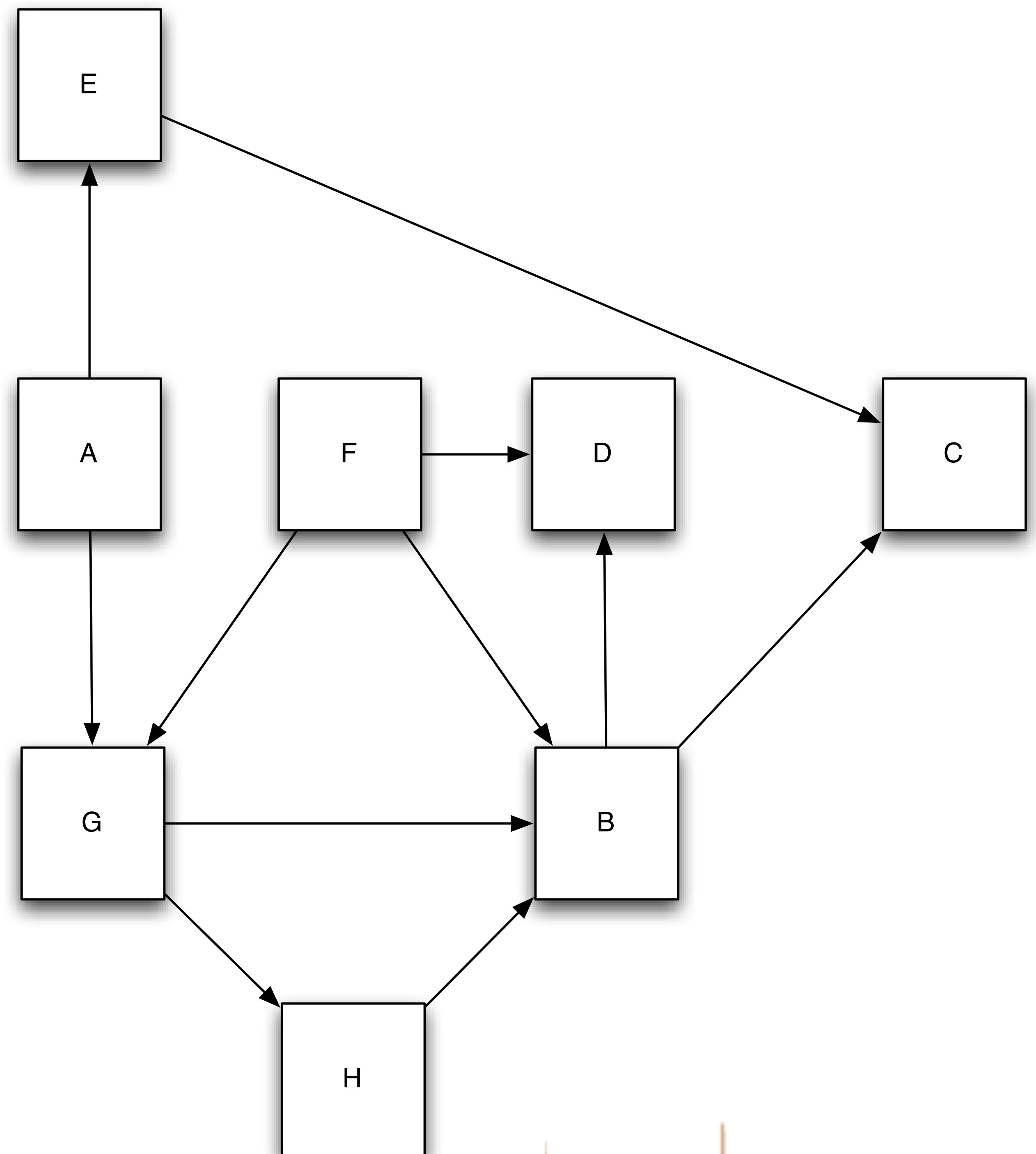
Donald J. Patterson

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

Markov Chains

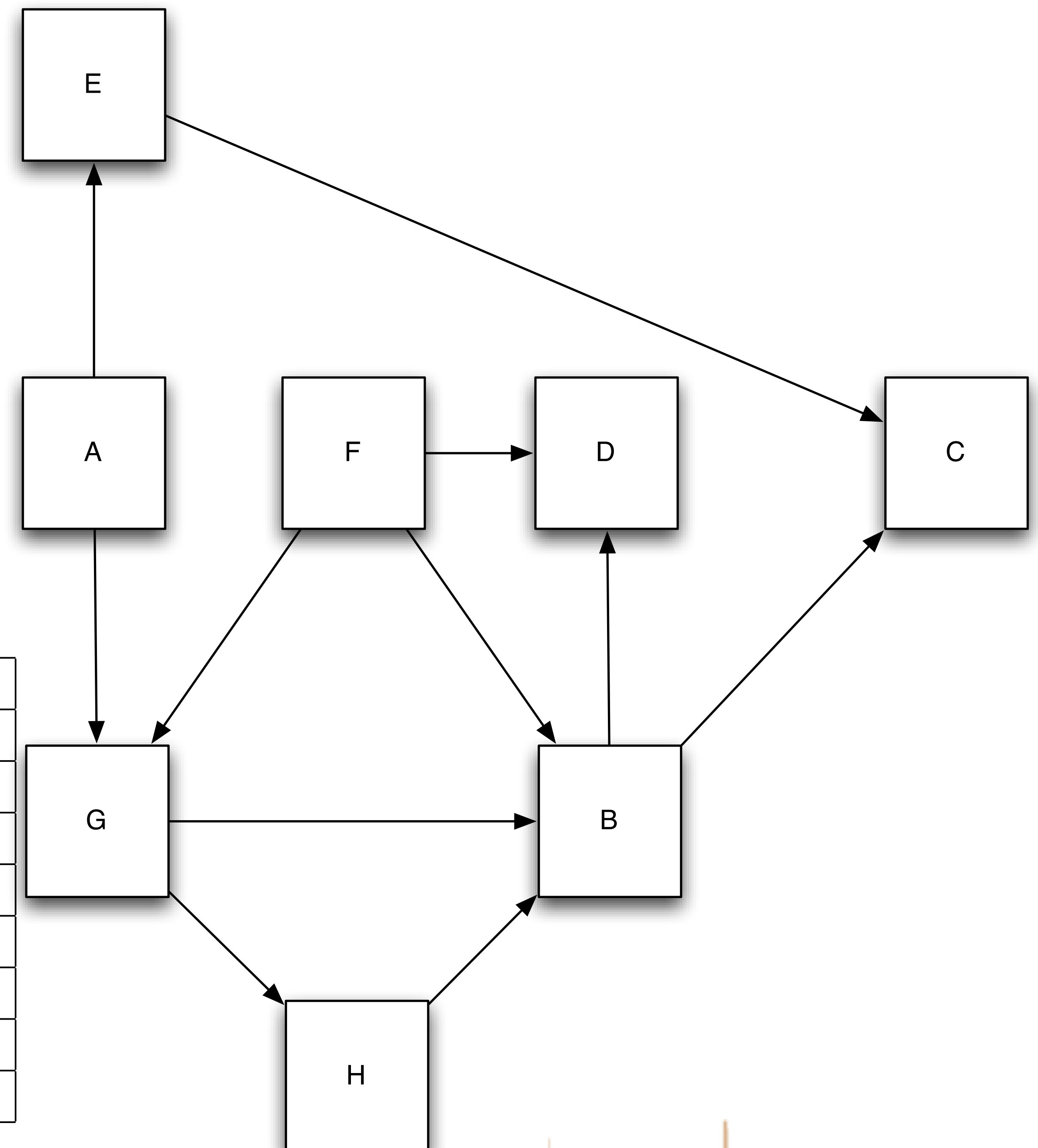
- Example:
  - 8 states
    - (web pages or whatever)
  - 8 by 8 transition prob. matrix

	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

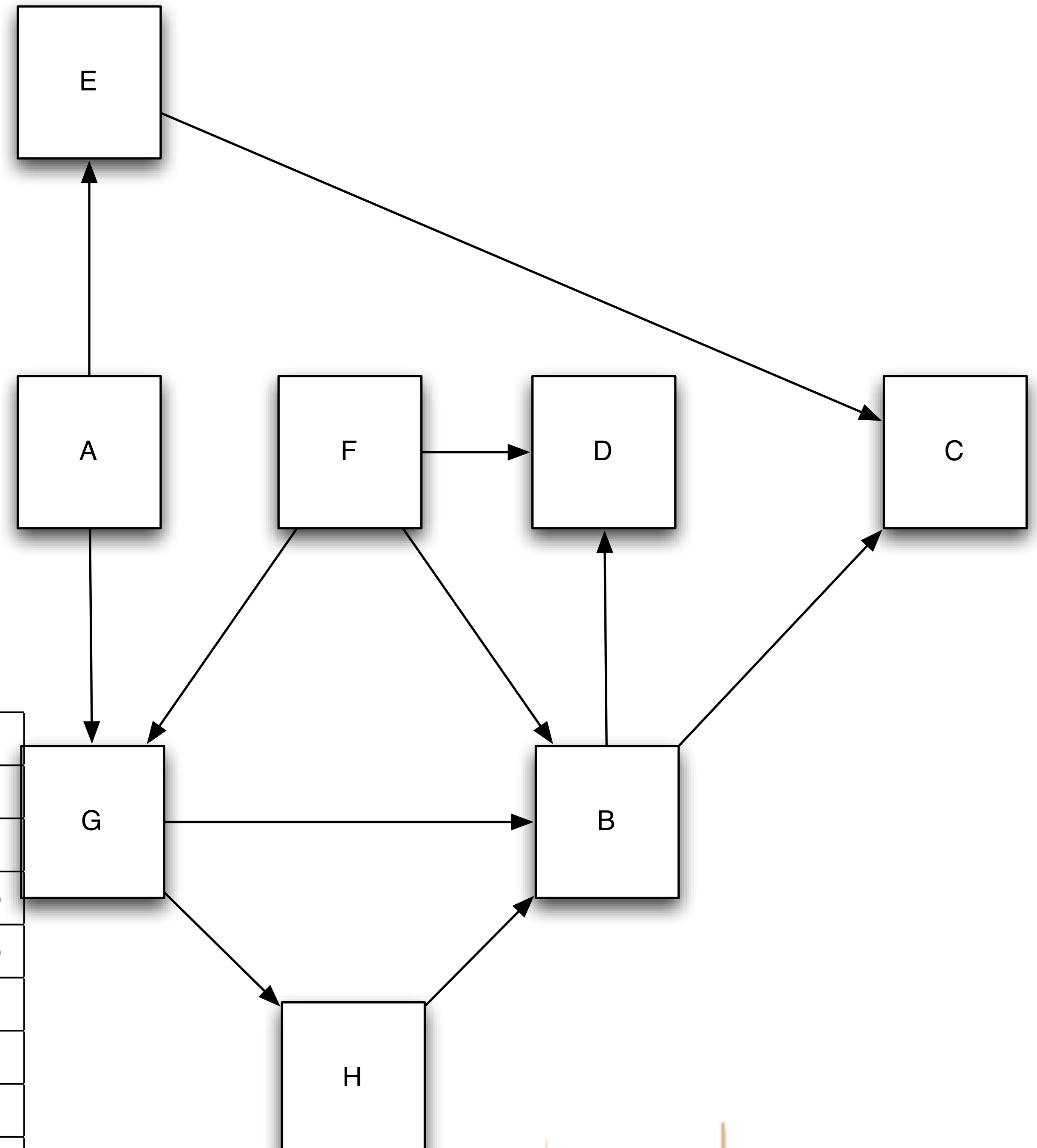


	A	B	C	D	E	F	G	H
A	0	0	0	0	0.5	0	0.5	0
B	0	0	0.5	0.5	0	0	0	0
C	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
D	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
E	0	0	1.0	0	0	0	0	0
F	0	0.33	0	0.33	0	0	0.33	0
G	0	0.5	0	0	0	0	0	0.5
H	0	1.0	0	0	0	0	0	0

Markov Chains

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

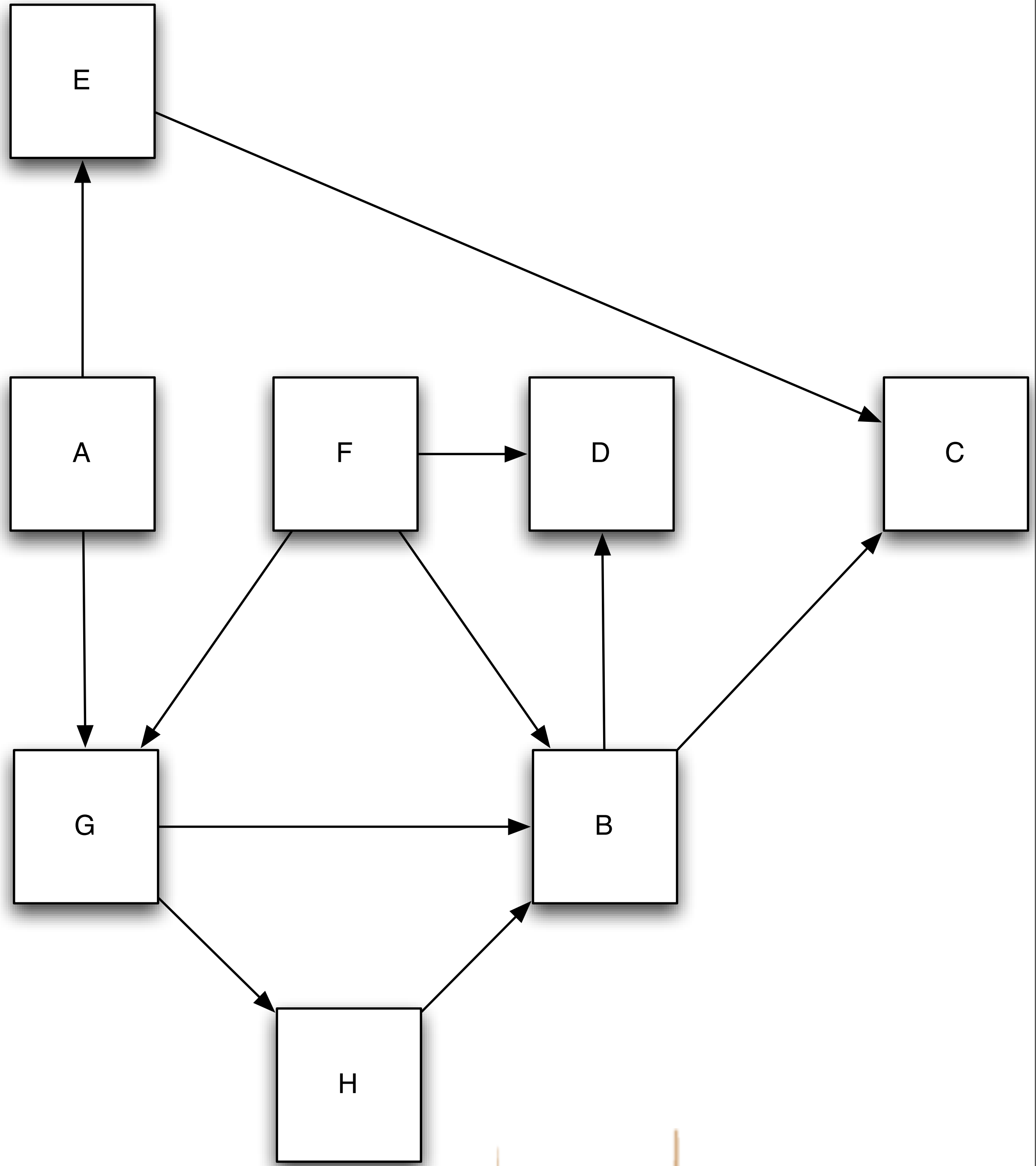
	A	B	C	D	E	F	G	H
A	0.01	0.01	0.01	0.01	0.47	0.01	0.47	0.01
B	0.01	0.01	0.47	0.47	0.01	0.01	0.01	0.01
C	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
D	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
E	0.01	0.01	0.93	0.01	0.01	0.01	0.01	0.01
F	0.01	0.32	0.01	0.32	0.01	0.01	0.32	0.01
G	0.01	0.47	0.01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

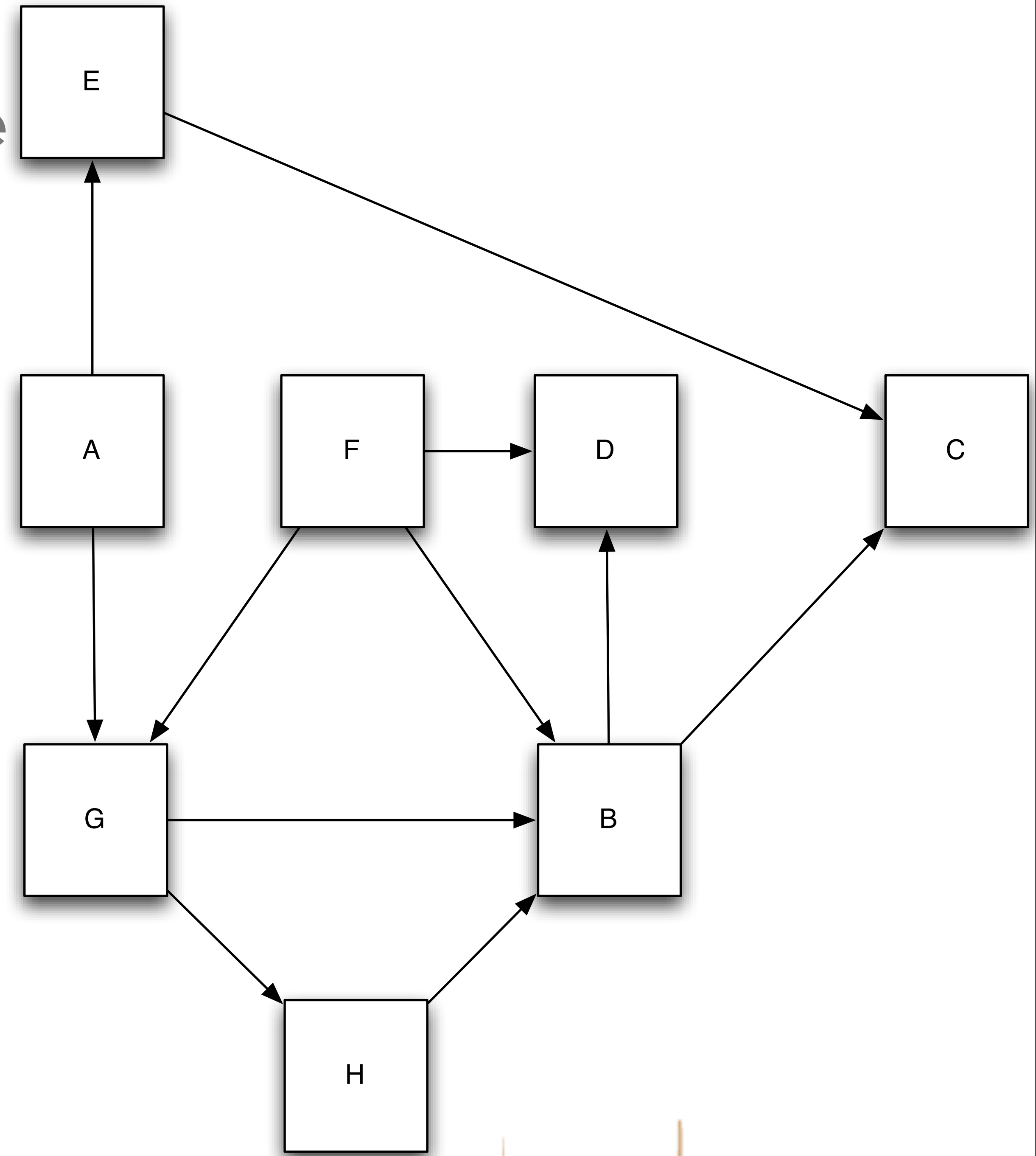
- You may be in one state at a time
- Every tick you move one step chosen randomly from the transition probability matrix

	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	0.125	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	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



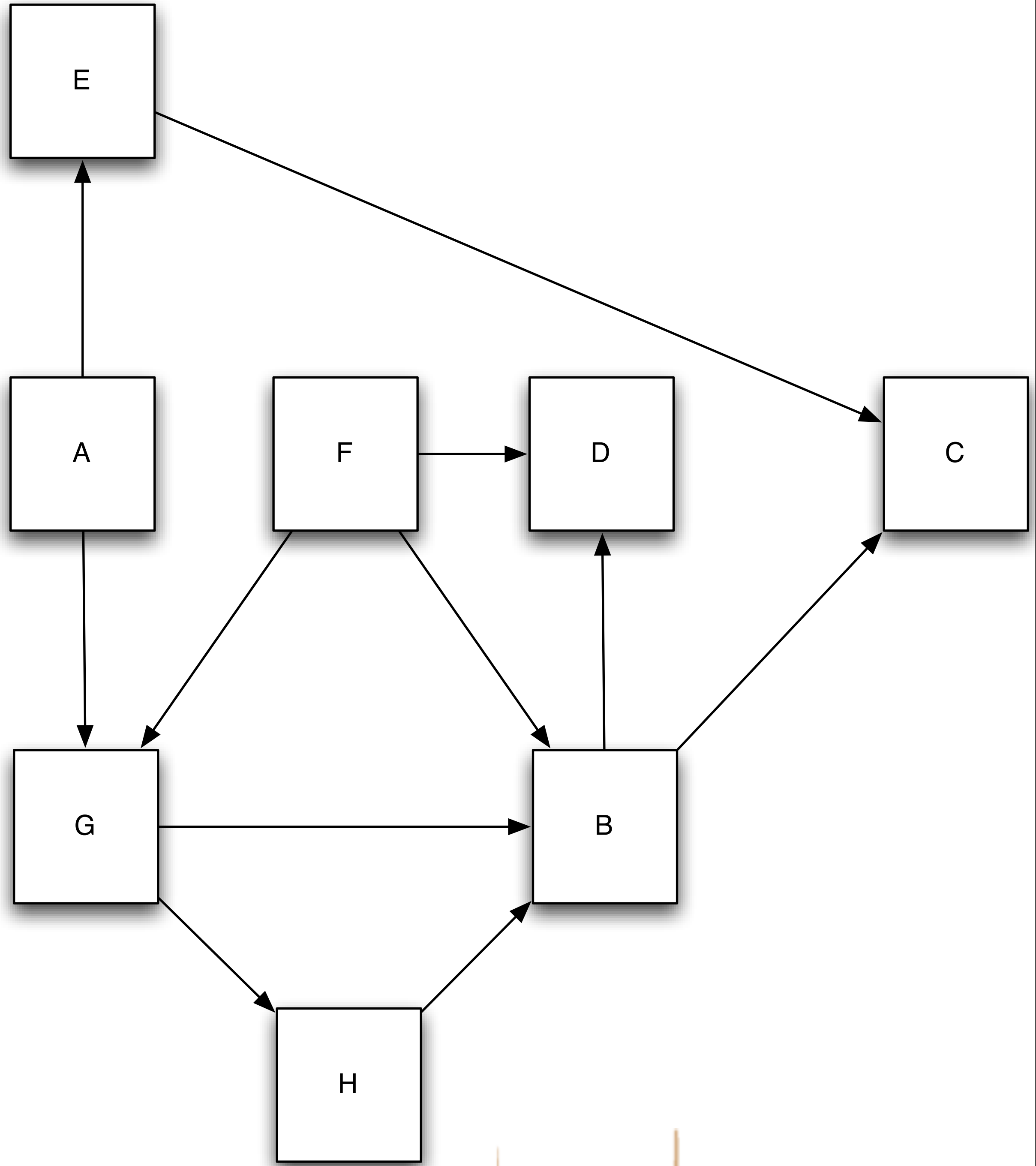
The Markov Property

- 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

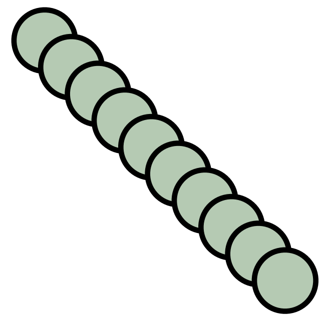


PageRank

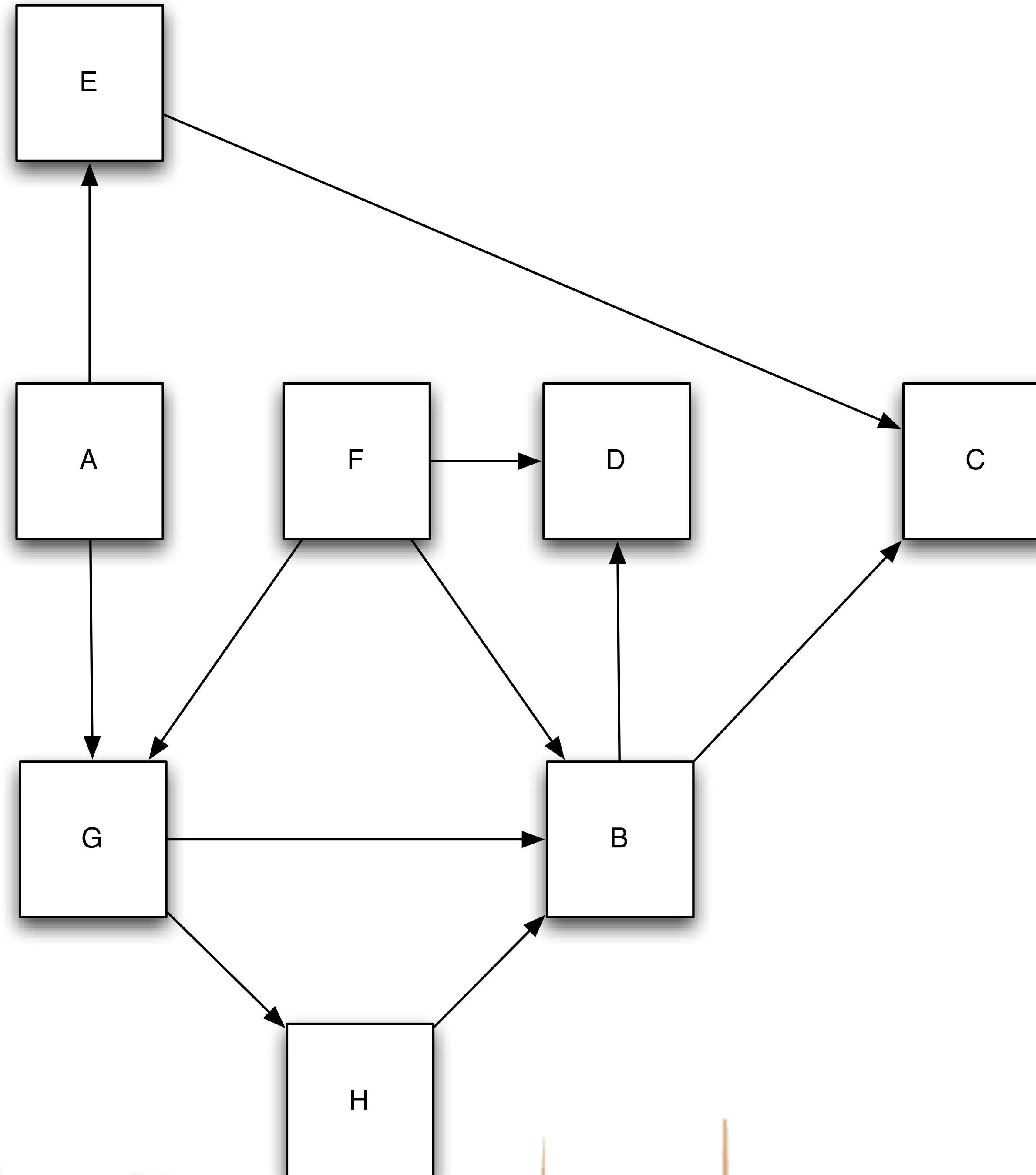
- PageRank is the long term visit rate of a random walk on the graph.
- With teleports



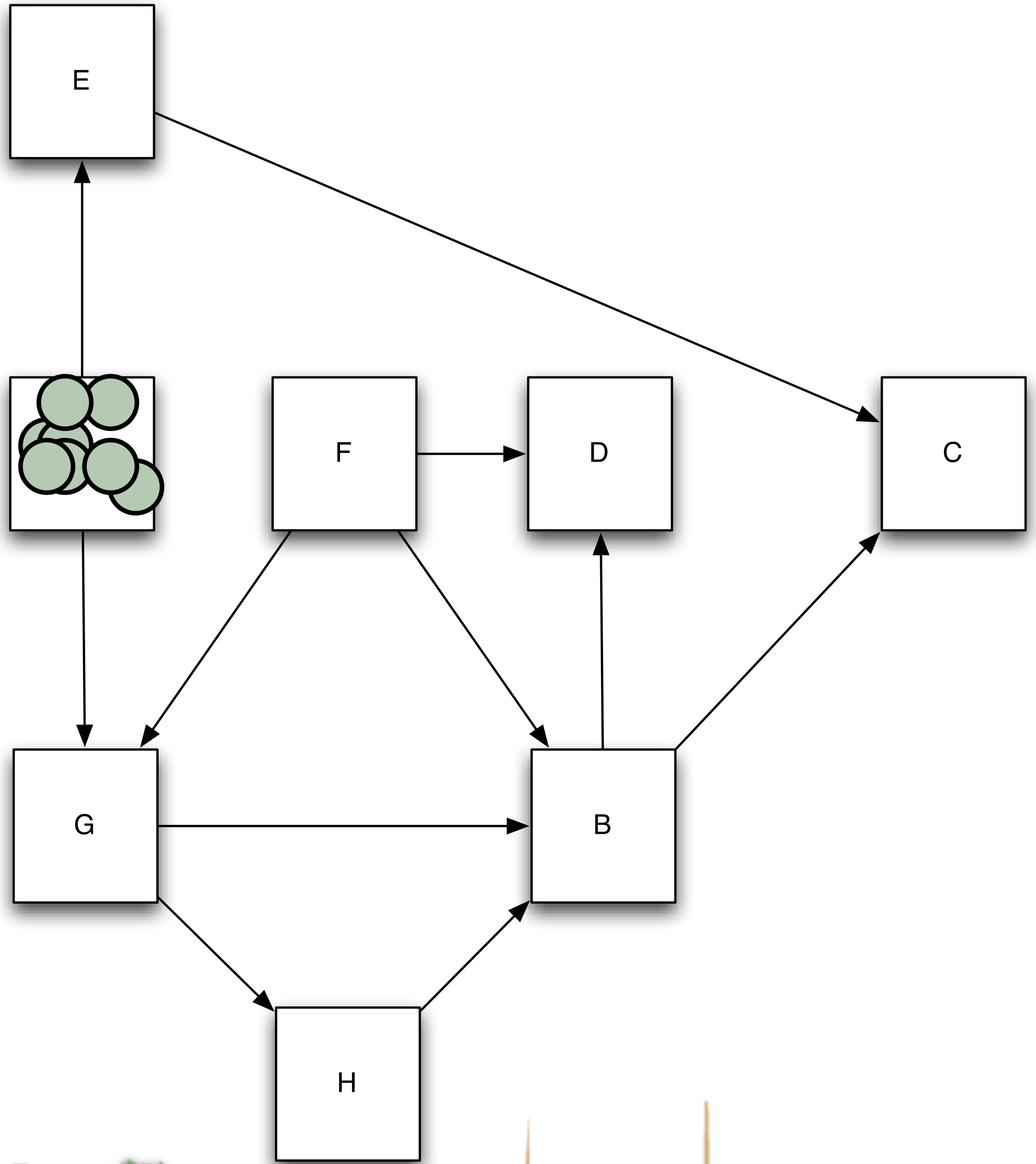
Long-Term visit rate



	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	0.125	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	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

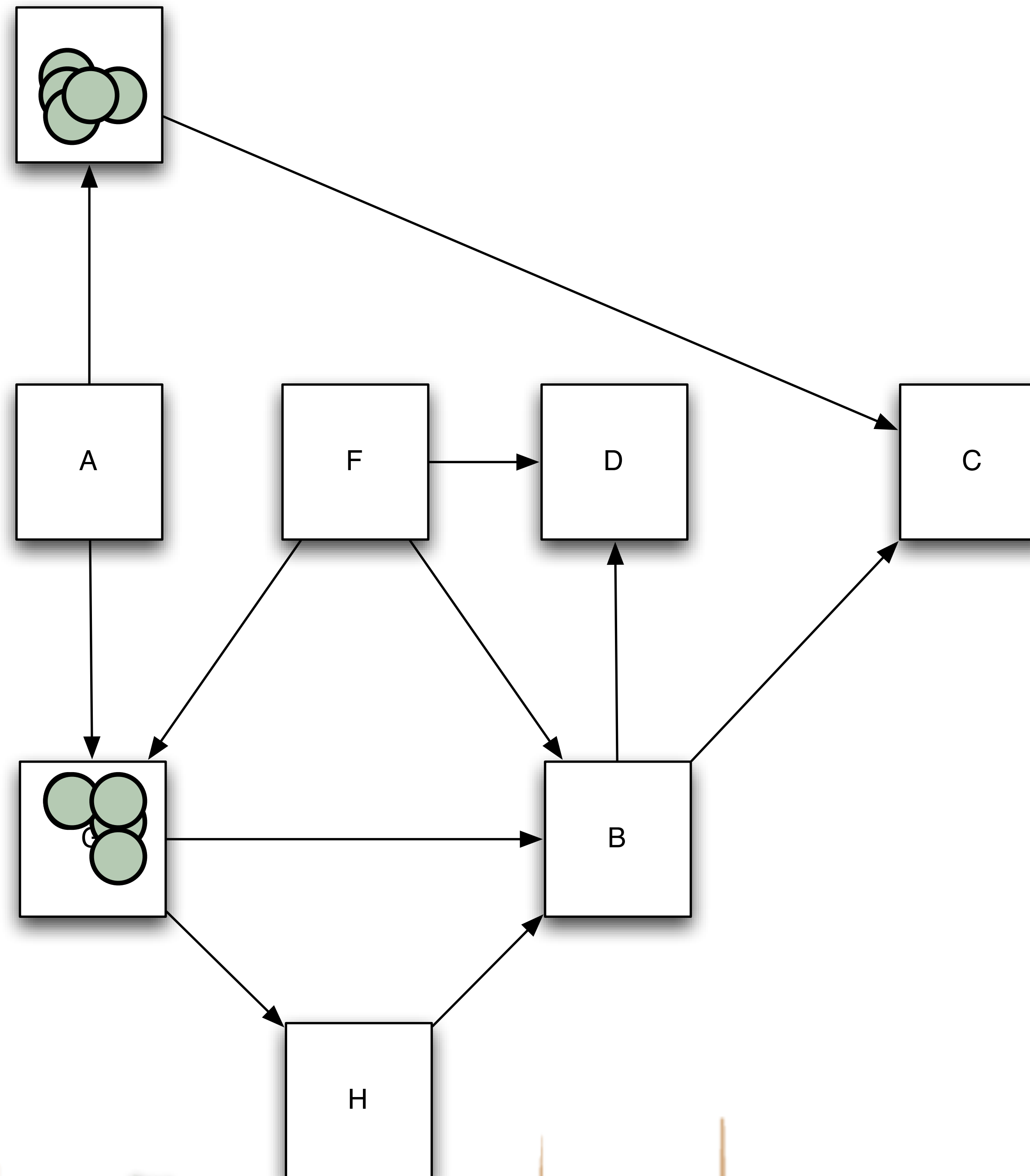


Long-Term visit rate



	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	0.125	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	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

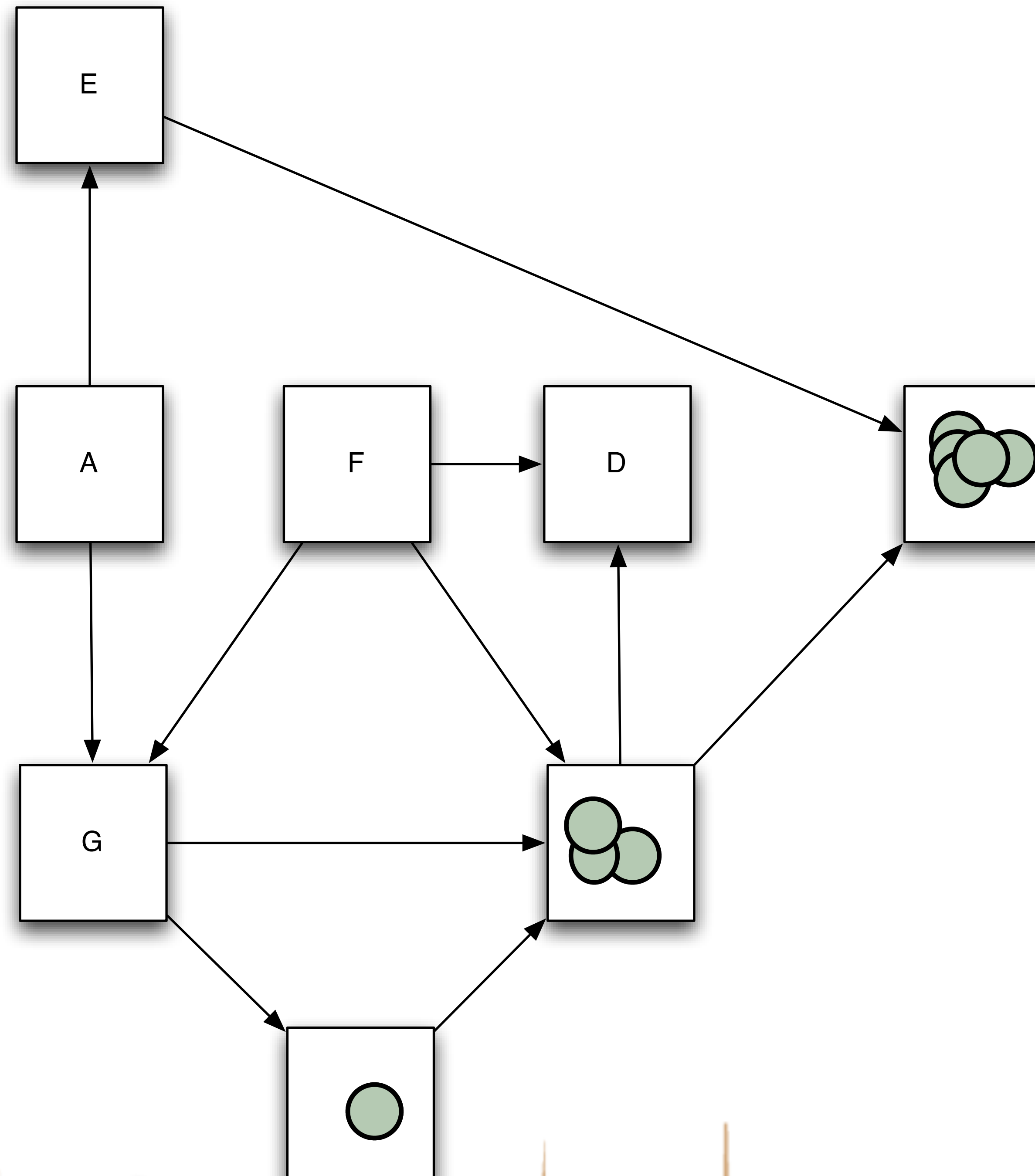
Long-Term visit rate



	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	0.125	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	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

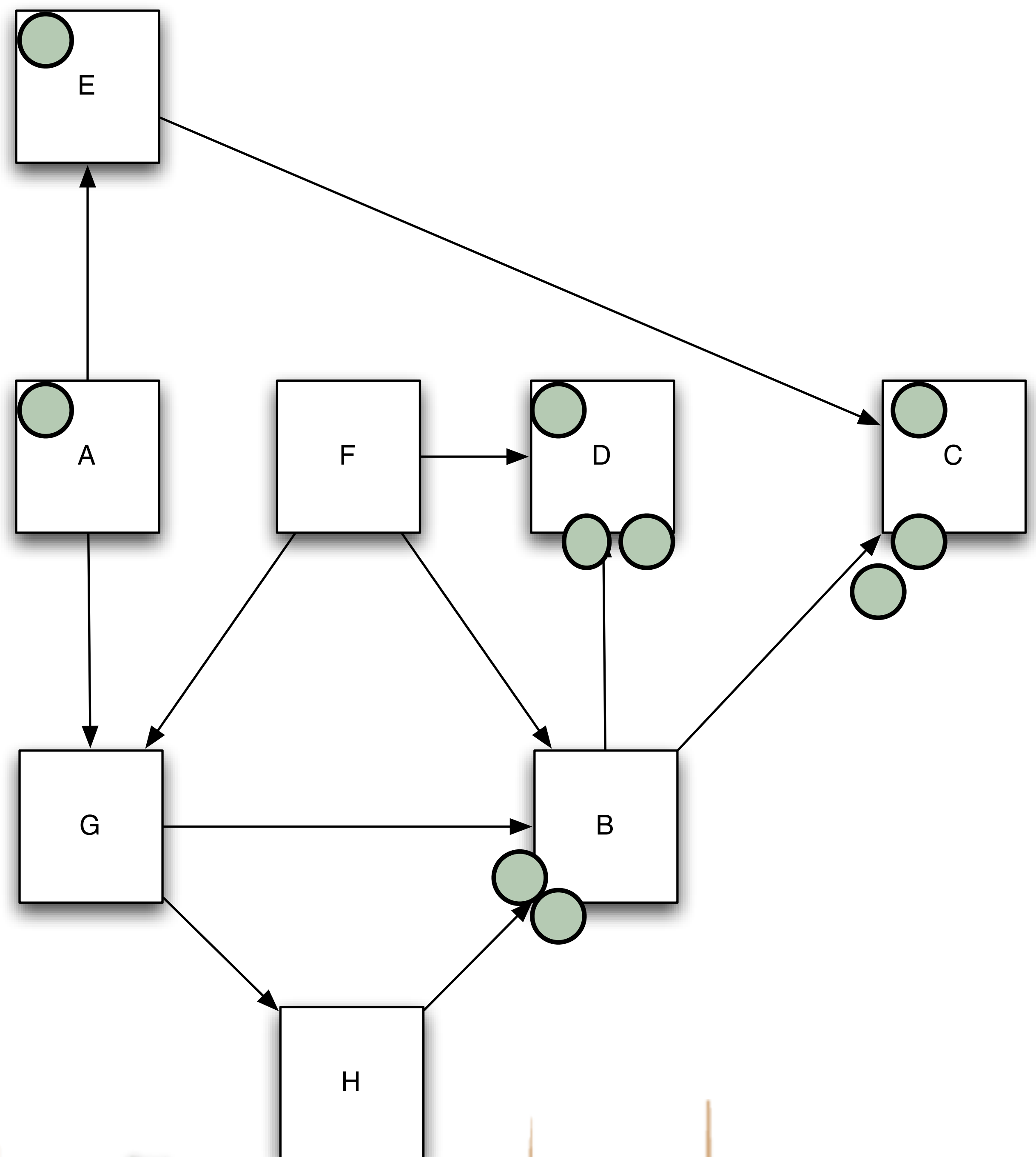
Long-Term visit rate

	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	0.125	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	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

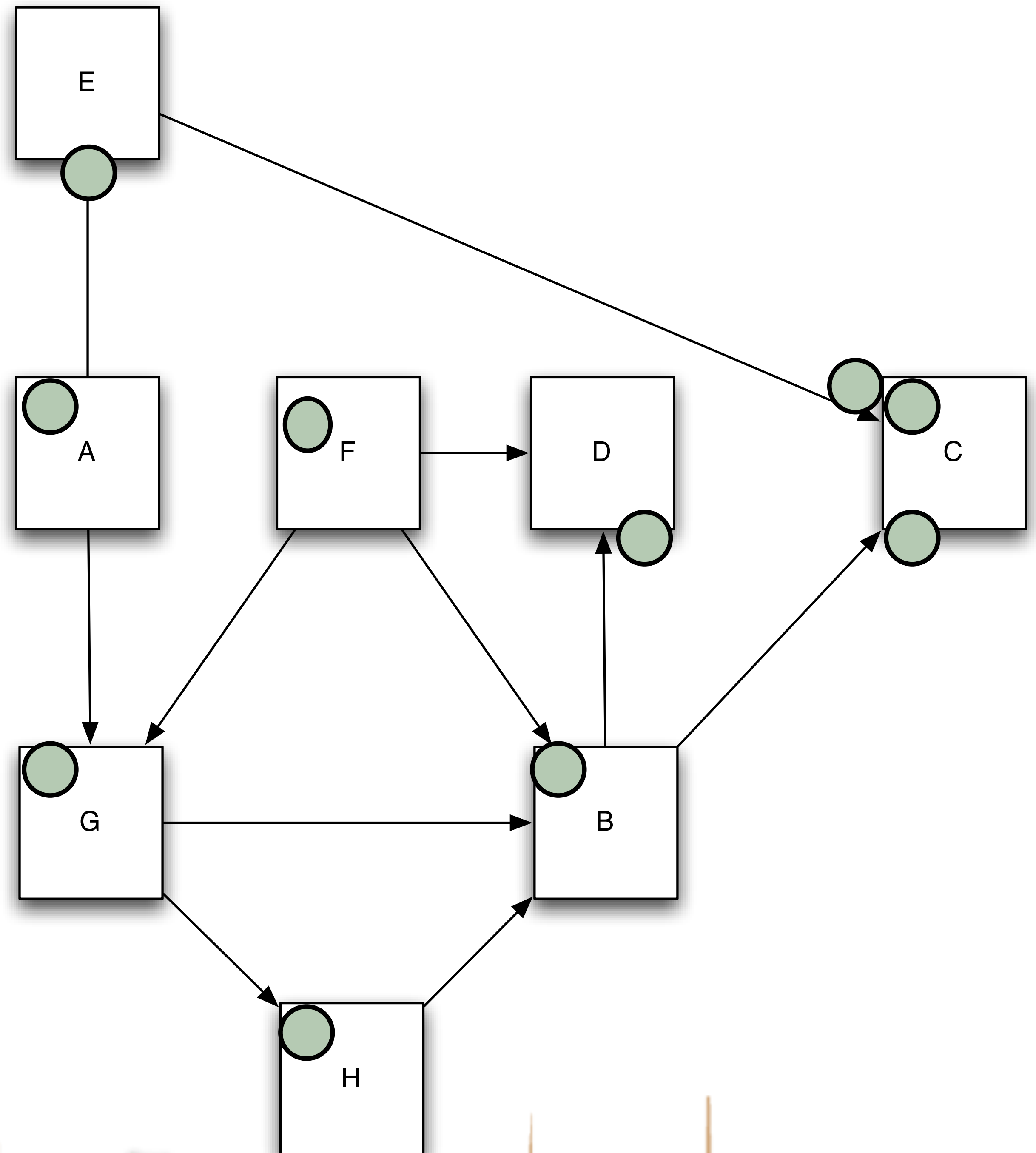


Long-Term visit rate

	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	0.125	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	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



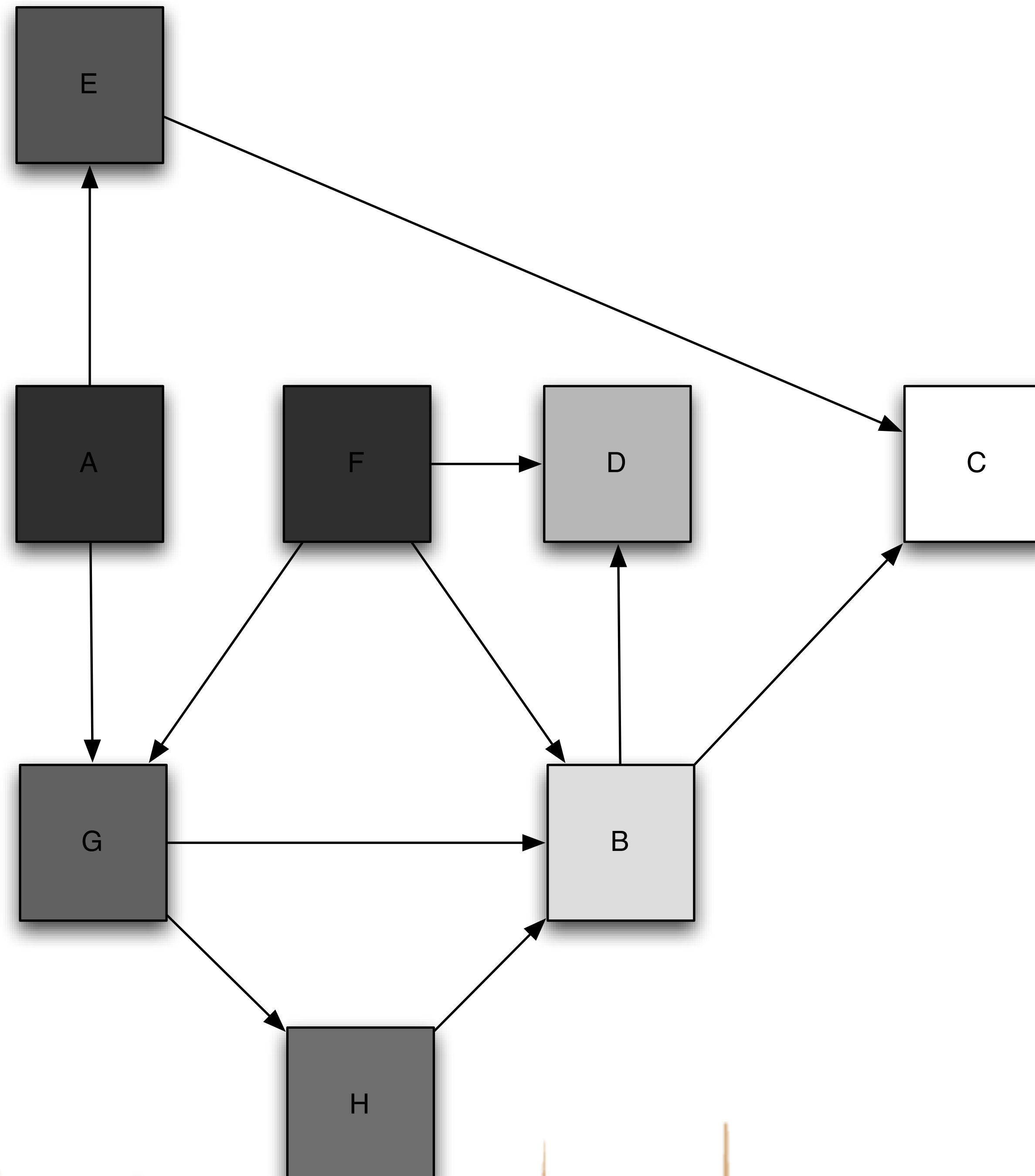
Long-Term visit rate



	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	0.125	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	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

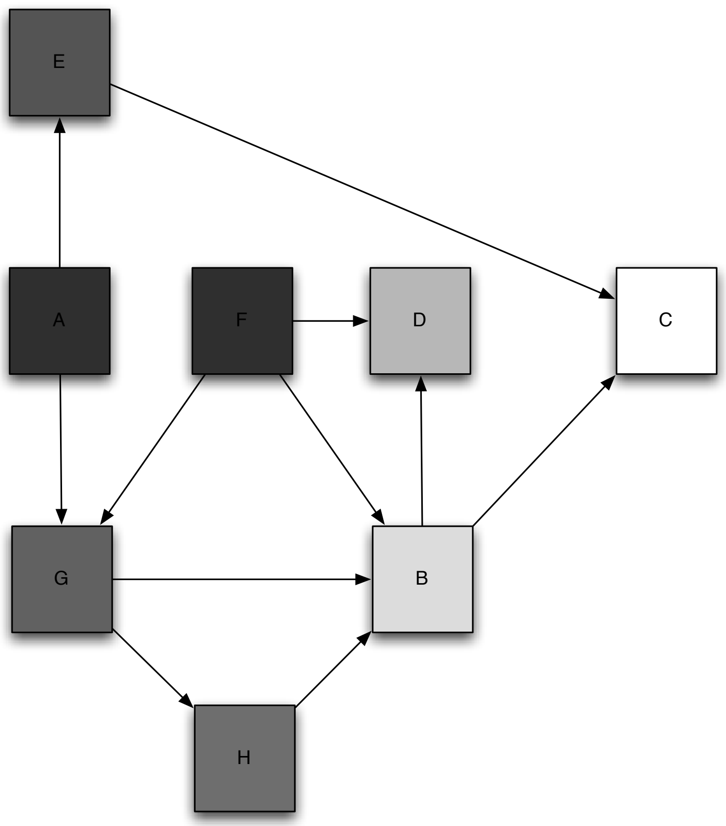
Long-Term visit rate

- 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 Theorem:
  - Every ergodic markov chain has a steady state -> has a PageRank



Calculating PageRank

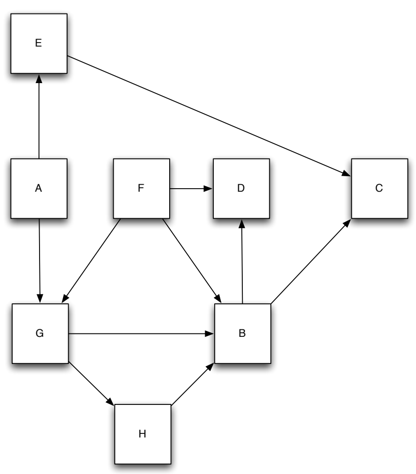
- Visual representation to math representation

$\vec{x_0}$

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

P

	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	0.125	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	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



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

	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	0.125	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	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

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

	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	0.125	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	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

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
A	0	0	0	0	0.5	0	0.5	0
B	0	0	0.5	0.5	0	0	0	0
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	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

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

	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	0.125	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	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

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

	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	0.125	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	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

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

	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	0.125	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	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

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

	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	0.125	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	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

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

	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	0.125	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	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

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

	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	0.125	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	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

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

	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	0.125	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	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

$\vec{x_1} =$

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

	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	0.125	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	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

$\vec{x_2} =$

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

	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	0.125	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	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

$\vec{x_3} =$

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

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

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

Long-Term visit rate

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