Chapter IV: Link Analysis

Information Retrieval & Data Mining
Universität des Saarlandes, Saarbrücken
Winter Semester 2011/12

Chapter IV: Link Analysis*

*Mostly following **Manning/Raghavan/Schütze,** with additions from other sources

Chapter IV.1: Background and PageRank

Based on **Manning/Raghavan/Schütze,** Chapter 21

World Wide Web as a web

- WWW pages are interlinked via *hyperlinks*

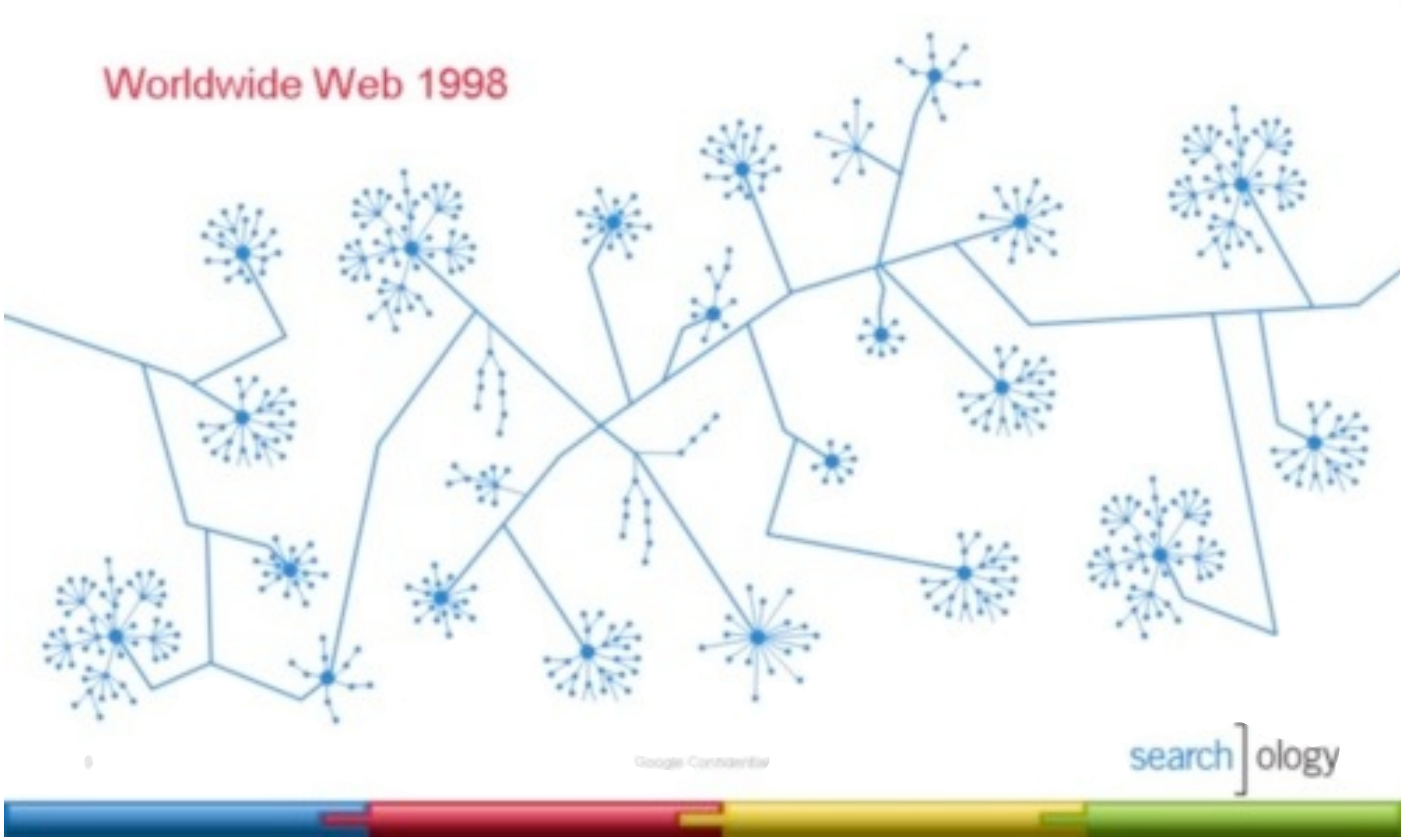


Image: Google Searchology 2007 <http://www.shareholder.com/Visitors/event/build2/mediapresentation.cfm?MediaID=25550&Player=1#>

World Wide Web as a web

- WWW pages are interlinked via *hyperlinks*

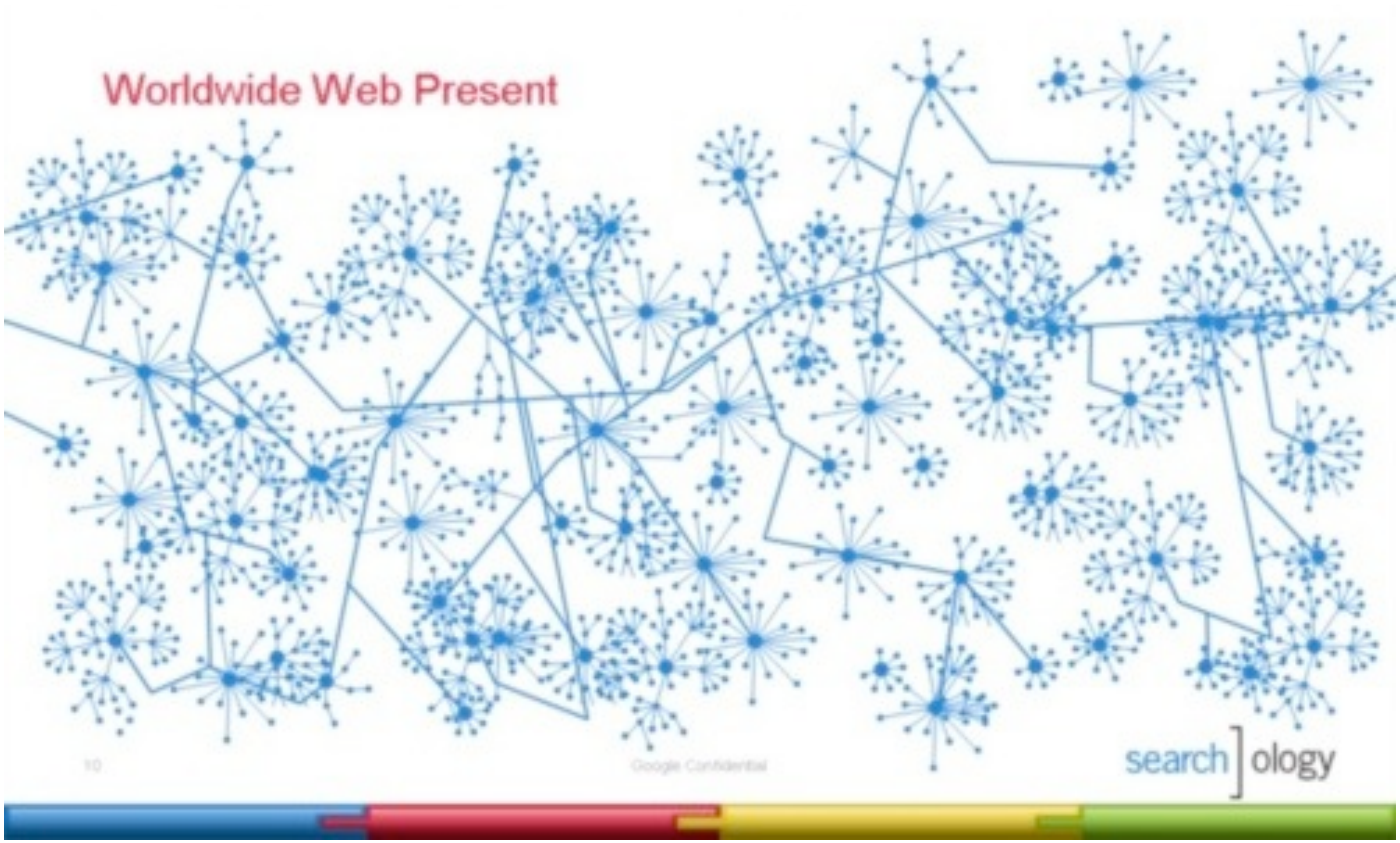


Image: Google Searchology 2007 <http://www.shareholder.com/Visitors/event/build2/mediapresentation.cfm?MediaID=25550&Player=1#>

Power-law distribution (Zipf's law)

Probability mass function *f(k; s, N)*:

$$f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$$

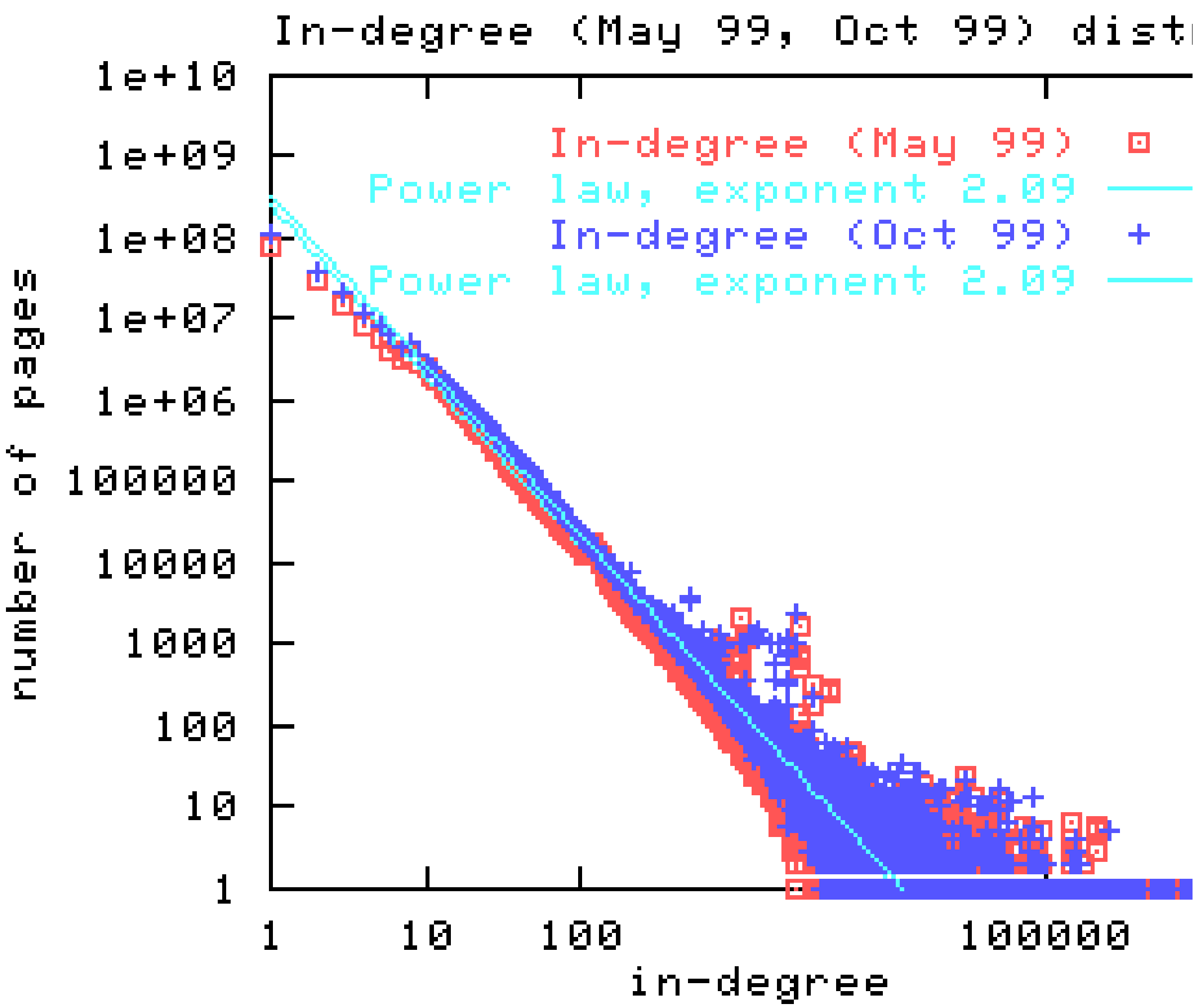
k = rank; s = parameter; N = total number of elements

Zipf's law models the frequency of *k*th most frequent element in

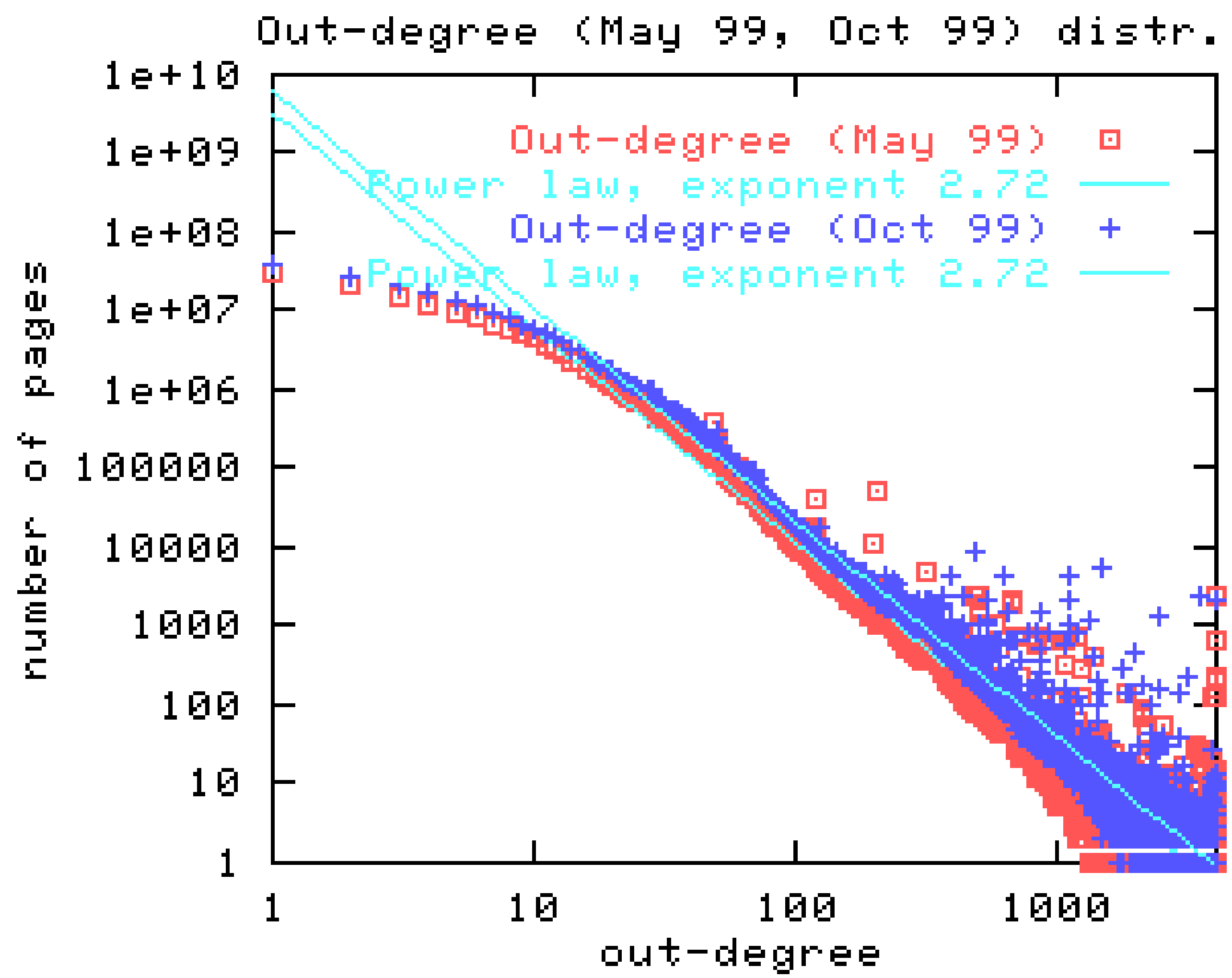
- word frequencies in corpora
- populations of cities in different countries
- income rankings
- ...

Link numbers follow power law

In-degree



Out-degree



Broder et al. *Graph structure in the web*. WWW'00

$s = 2.09$

$s = 2.72$

Using links to rank

- Linking to a page can be considered as an endorsement
  - This idea obviously pre-dates Facebook...
- This information could be used to find authoritative web pages
  - Rough idea: on two pages about the same topic if the first links to the second, the second is more authoritative
- Analogies in scientific citations
  - High citation count = prestigious article
  - But what if the citations/links say ”This work is rubbish”
    - Apparently not a big problem

The random surfer

- The model:
  - A random surfer goes to a random web page
  - Clicks a random link to move to other web page
  - Repeats ad infinitum
- Intuition: most visited pages will be the ones with most in-links from pages with most in-links etc.
  - I.e. the ones that should have the highest rank
- Can this be formalized?

Image: Wikipedia

Interlude: Markov chains

- A **stochastic process** is a family of random variables $\{X_t : t \in T\}$
  - Henceforth $T = \{0, 1, 2, \ldots\}$ and t is called *time*
    - This is *discrete stochastic process*
- Stochastic process $\{X_t\}$ is **Markov chain** if always

$$\Pr[X_t = x \mid X_{t-1} = a, X_{t-2} = b, \ldots, X_0 = z]$$
$$= \Pr[X_t = x \mid X_{t-1} = a]$$

  - Memory-less property
- A Markov chain is **time-homogenous** if for all t

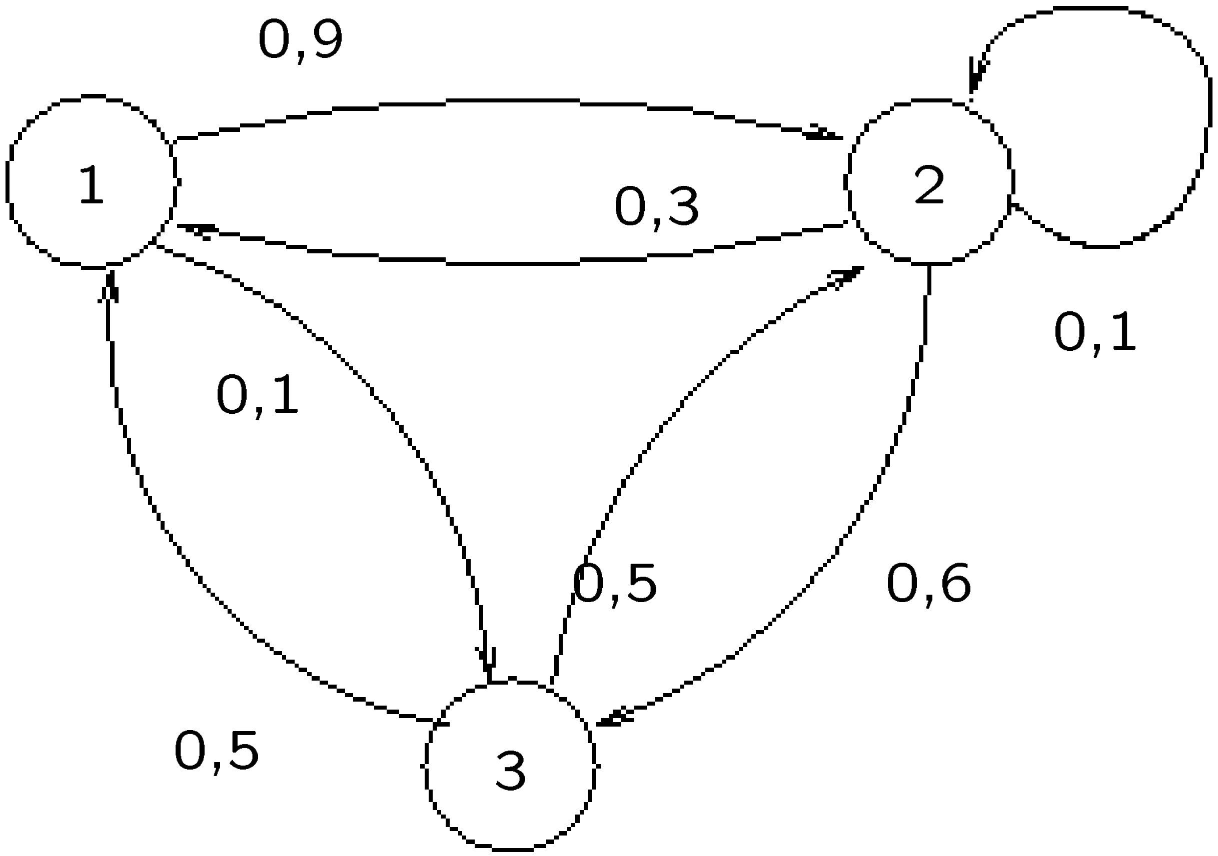
$$\Pr[X_{t+1} = x \mid X_t = y\,] = \Pr[X_t = x \mid X_{t-1} = y]$$

  - We only consider time-homogenous Markov chains

Transition matrix

- The **state space** of a Markov chain $\{X_t\}_{t \in T}$ is the countable set S of all values X_t can assume
  - $X_t: \Omega \to S$ for all $t \in T$
  - Markov chain is in state s at time t if $X_t = s$
  - A Markov chain $\{X_t\}_{t \in T}$ is *finite* if it has finite state space
- If Markov chain $\{X_t\}$ is finite and time-homogenous, its **transition probabilities** can be expressed with a matrix $\mathbf{P} = (p_{ij})$, $p_{ij} = \Pr[X_1 = j \mid X_0 = i]$
  - Matrix $\mathbf{P}$ is n-by-n if Markov chain has n states and it is *right stochastic*, i.e. $\sum_j p_{ij} = 1$ for all i (rows sum to 1)

Example Markov chain



$$P = \begin{pmatrix} 0 & 9/10 & 1/10 \\ 3/10 & 1/10 & 6/10 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

Classifying the states

- State i can be *reached* from state j if there exists $n \geq 0$ such that $(\mathbf{P}^n)_{ij} > 0$
  - $\mathbf{P}^n$ is the nth exponent of $\mathbf{P}$, $\mathbf{P}^n = \mathbf{P}\times\mathbf{P}\times\cdots\times\mathbf{P}$
- If i can be reached from j and vice versa, i and j *communicate*
  - If all states $i, j \in S$ communicate, Markov chain is **irreducible**
- If the probability that the process visits a state i infinitely many times is 1, then state i is **recurrent**
  - State is **positive recurrent** if the estimated return time to it is finite
  - Markov chain is recurrent if all of its states are

More classifying of the states

- State i has **period** k if any return to i must occur in time that is multiple of k:

$$k = \gcd\{n : \Pr[X_n = i \mid X_0 = i] > 0\}$$

  - State i is **aperiodic** if it has period $k = 1$; otherwise it is **periodic** with period k
  - Markov chain is aperiodic if all of its states are
- State i is **ergodic** if it is aperiodic and positive recurrent
  - Markov chain is ergodic if all of its states are

Two important results for finite MCs

Lemma IV.1: Every finite Markov chain has at least one recurrent state and all of its recurrent states are positive recurrent.

Corollary IV.2: Finite, irreducible, and aperiodic Markov chain is ergodic.

Stationary distributions

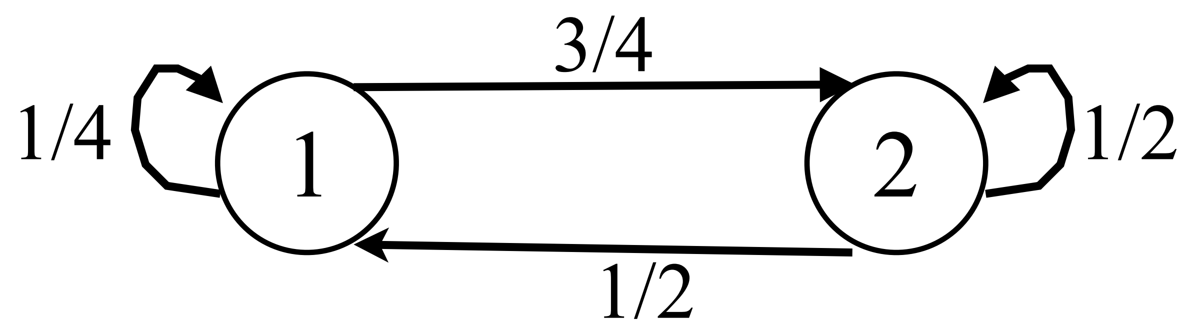
- If $\boldsymbol{\pi}$ is such that $\pi_i \geq 0$ for all i, $\sum_i \pi_i = 1$, and
$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$$
then $\boldsymbol{\pi}$ is the **stationary distribution** of the Markov chain
- Let $h_{ii} = \sum_{t\geq 1} t\Pr[X_t = i \text{ and } X_n \neq i \text{ for } n < t \mid X_0 = i]$ be the estimated return time to state i

Theorem IV.3: If Markov chain is finite, irreducible, and ergodic, then
1. it has an unique stationary distribution $\boldsymbol{\pi}$
2. for all i and j, $\lim_{t\to\infty} (\mathbf{P}^t)_{ji}$ exists and is the same for all j
3. $\pi_i = \lim_{t\to\infty} (\mathbf{P}^t)_{ji} = 1/h_{ii}$

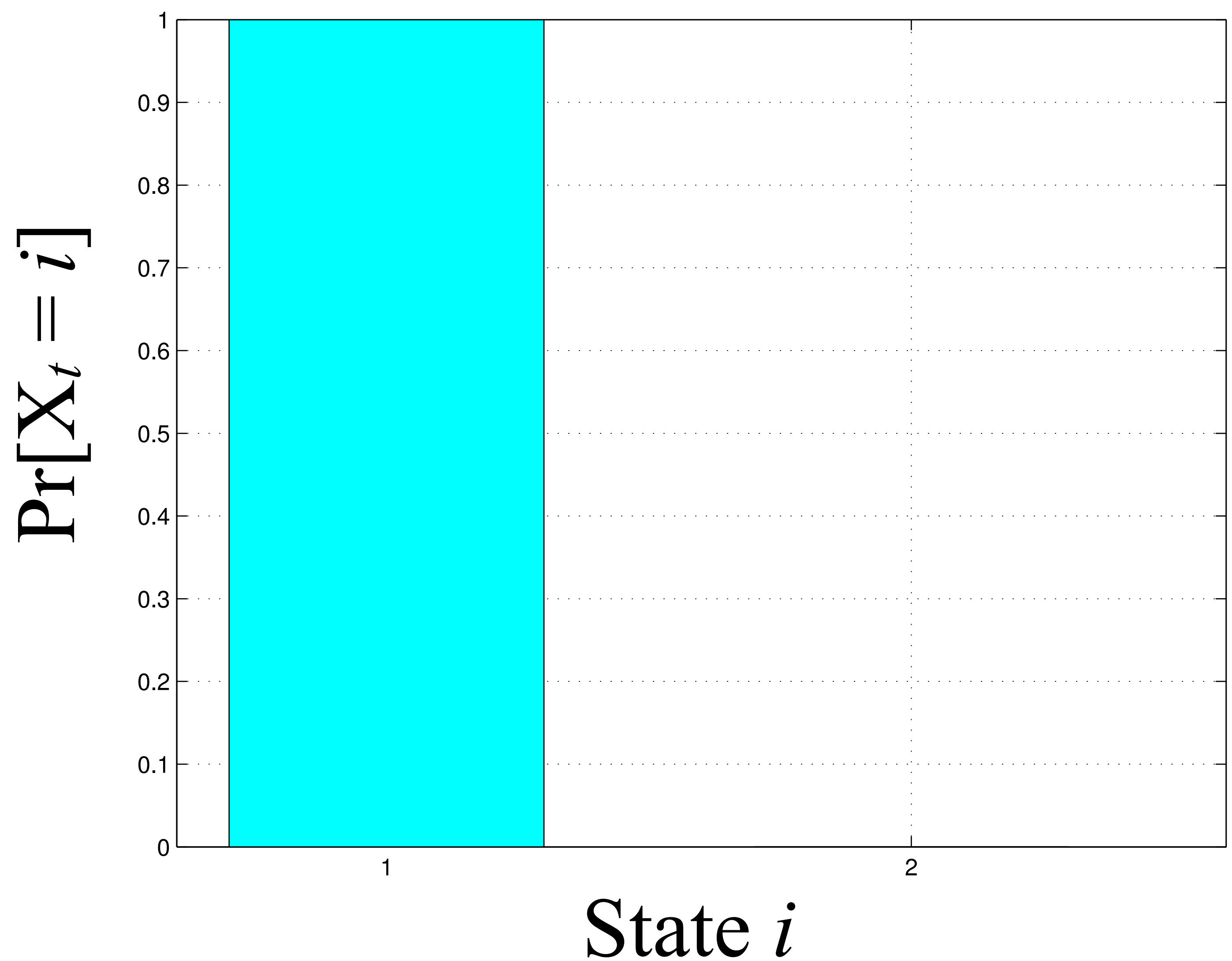
More on stationary distributions

- If Markov chain has a stationary distribution, then the probability that the chain is in state *i* after long-enough time is independent of the starting time but depends only on the stationary distribution
- Aperiodicity is not necessary condition for stationary distribution to exist, but then the stationary distribution will not be the limit of transition probabilities
  - Two-state chain that always switches the state has stationary distribution (1/2, 1/2), but the transitions look either (1, 2, 1, 2, ...) or (2, 1, 2, 1, ...) depending on the starting state

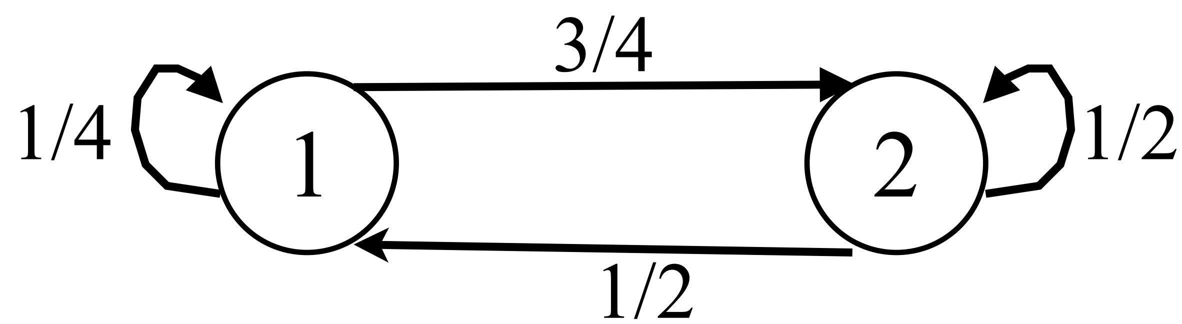
Example of stationary distribution

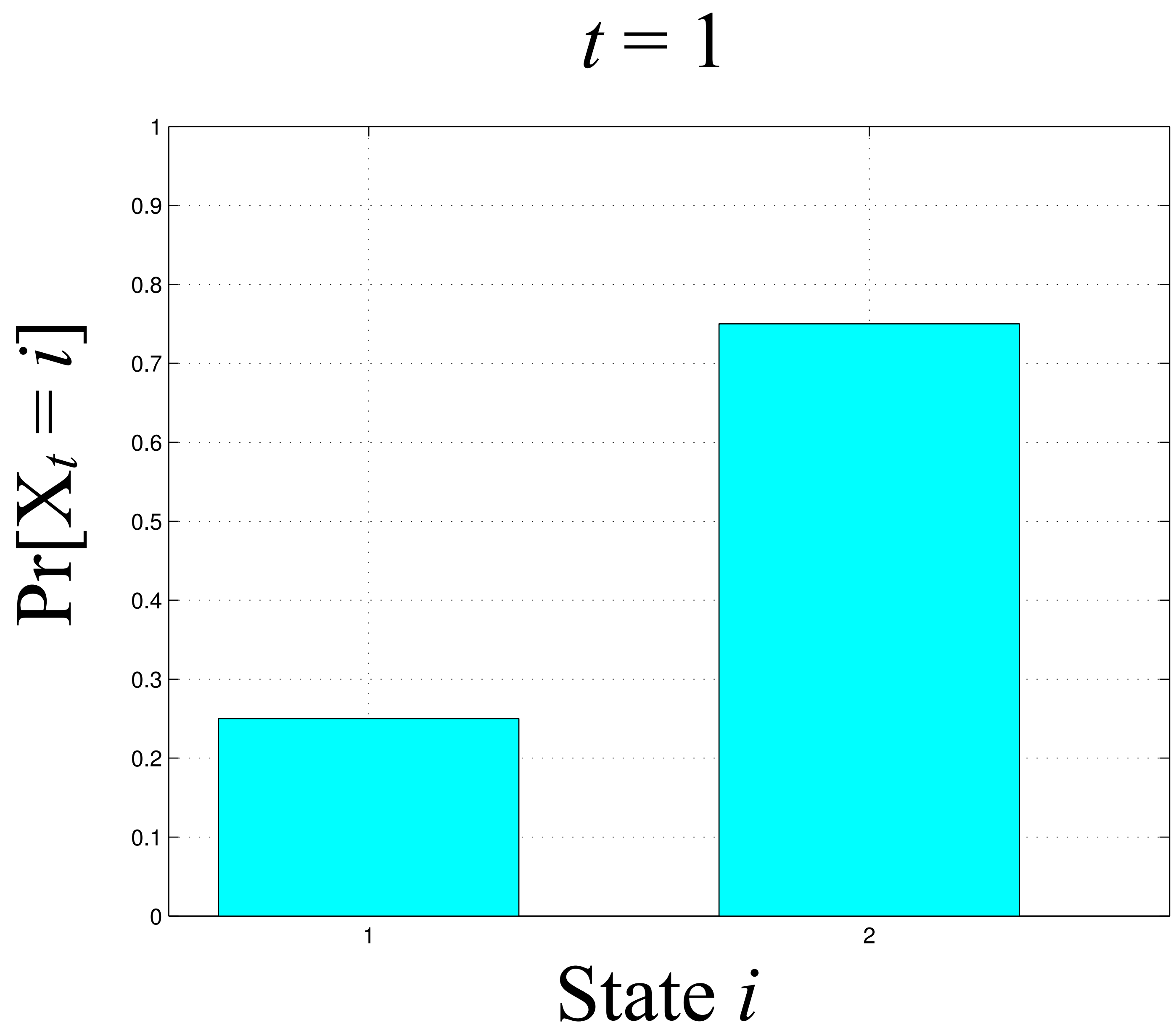


$t = 0$

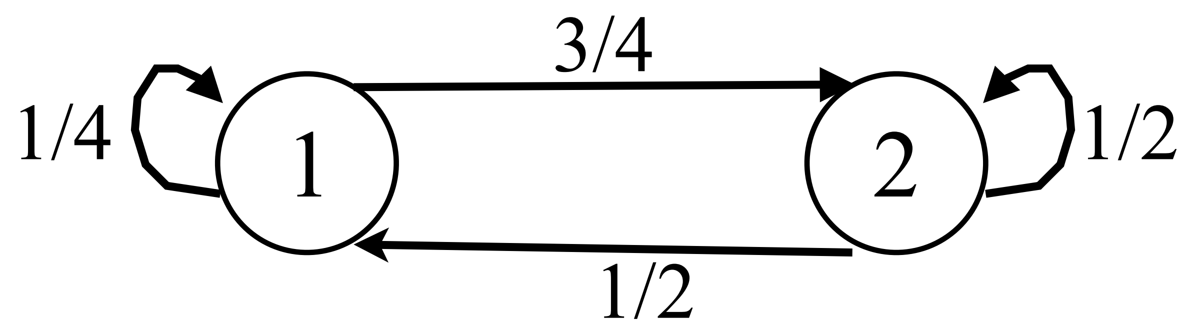


Example of stationary distribution

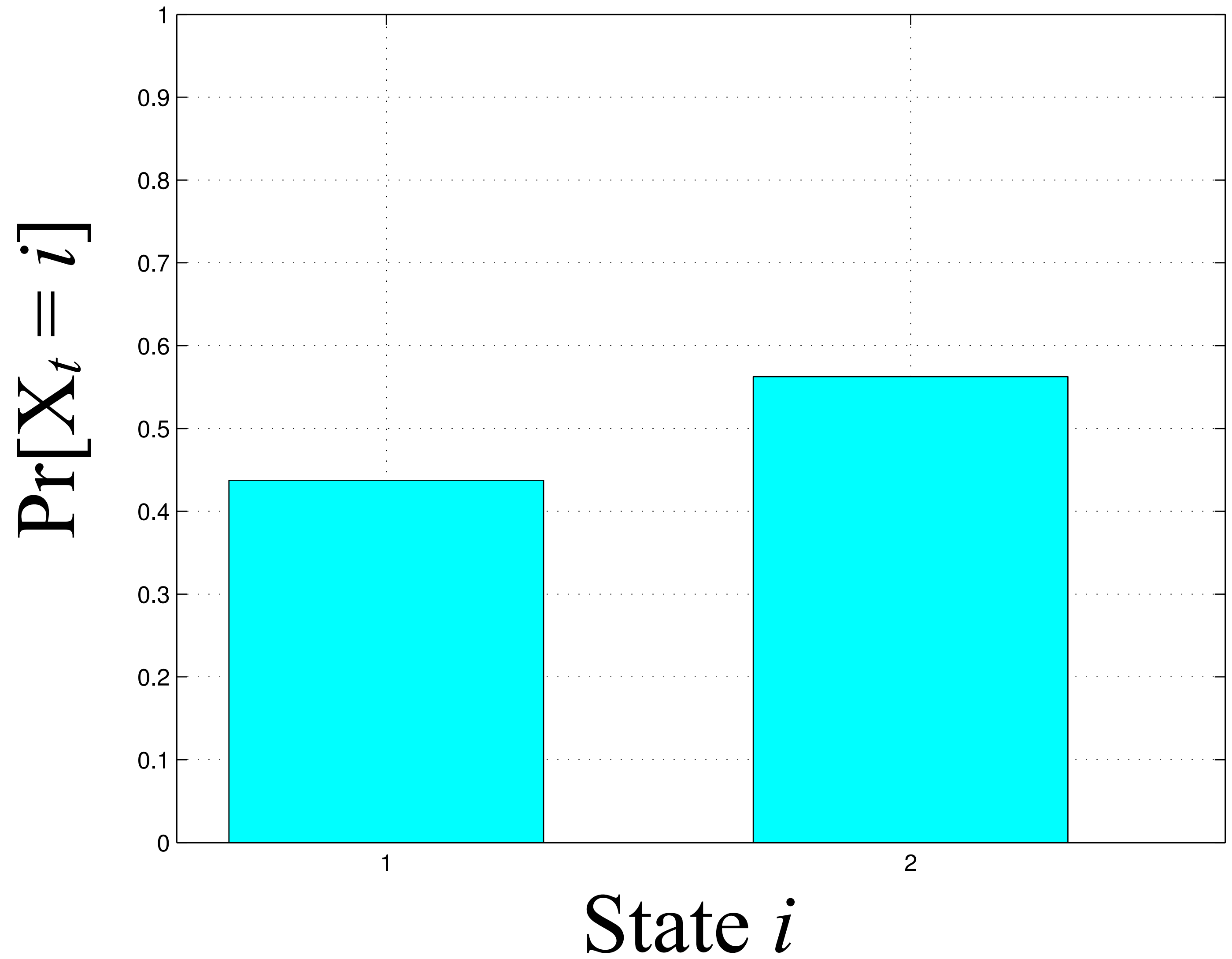




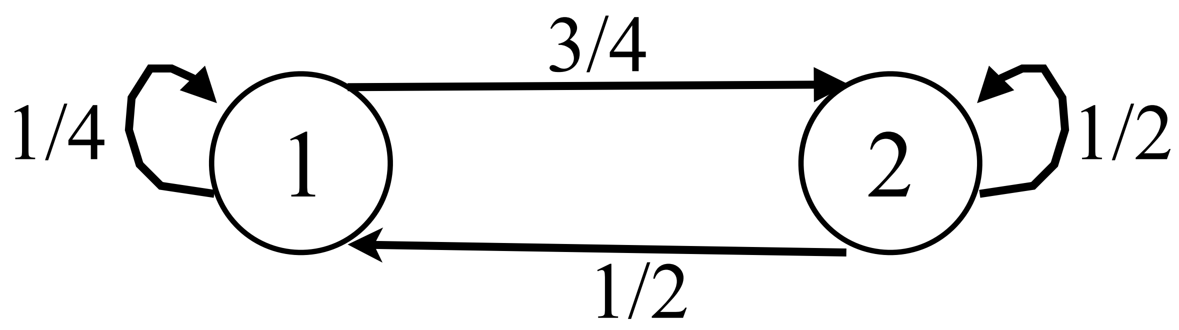
Example of stationary distribution



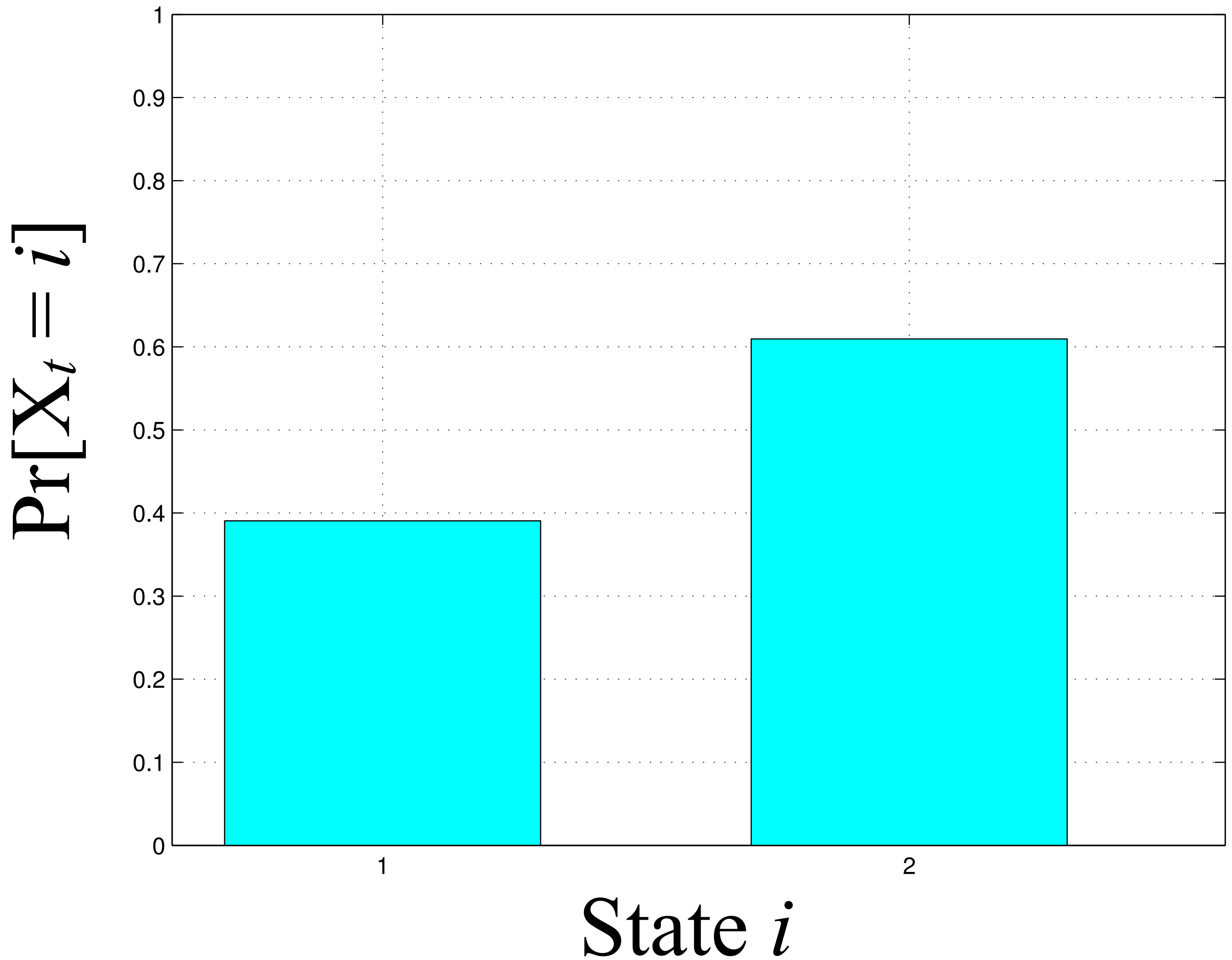
$t = 2$



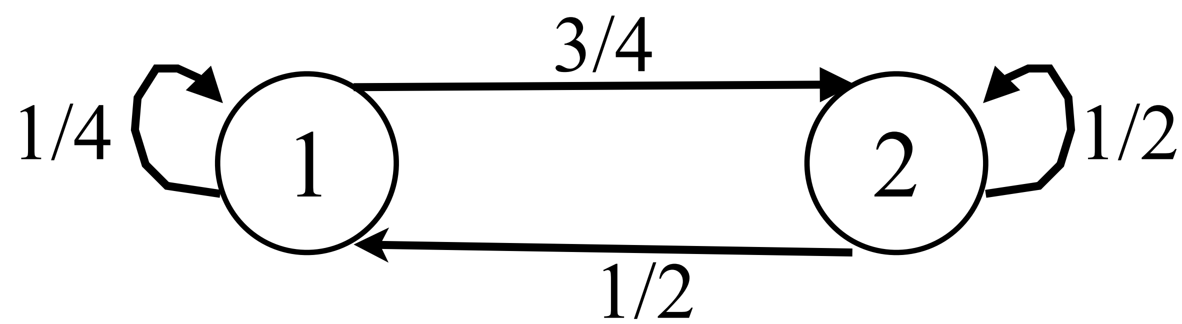
Example of stationary distribution

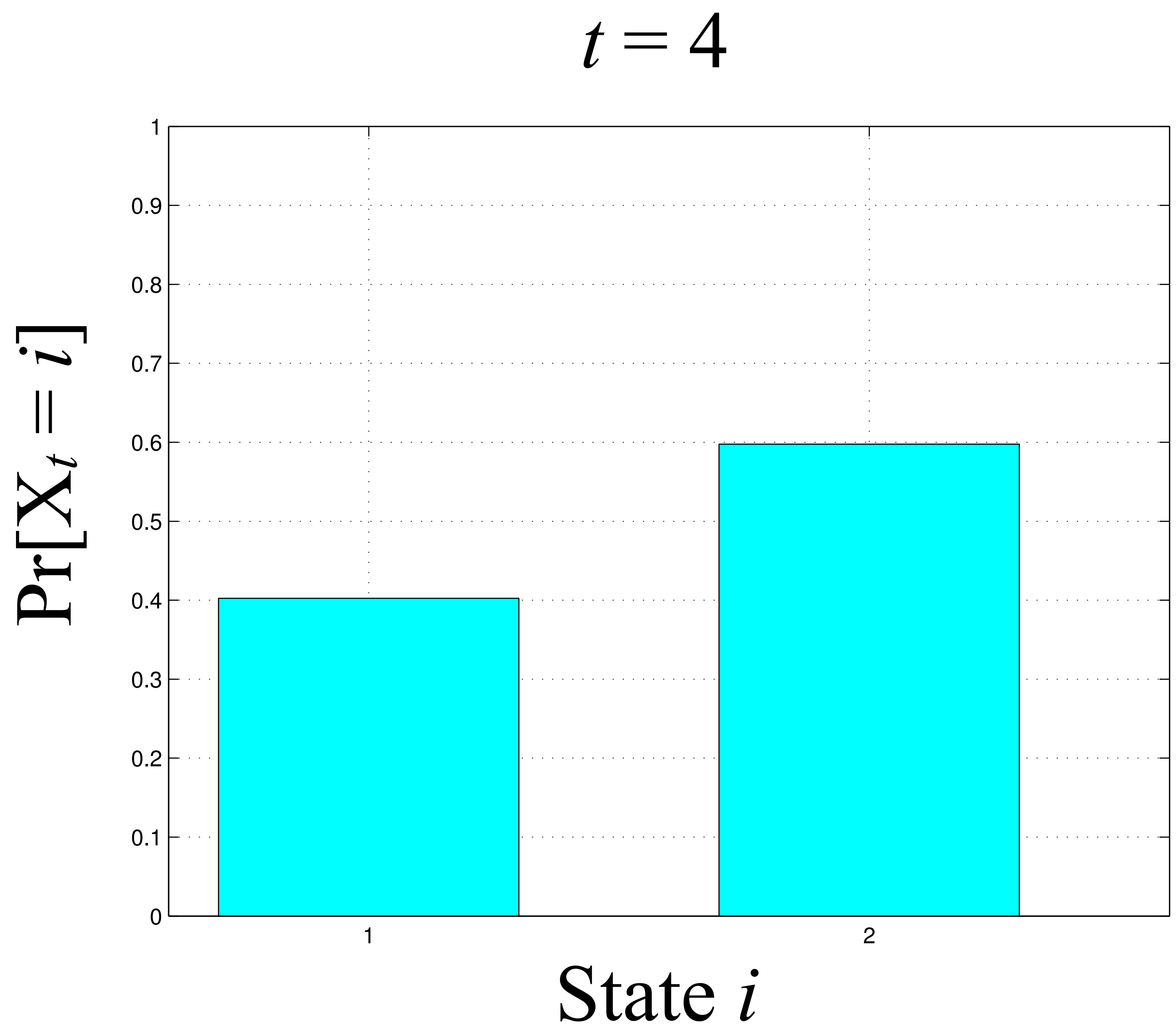


$t = 3$

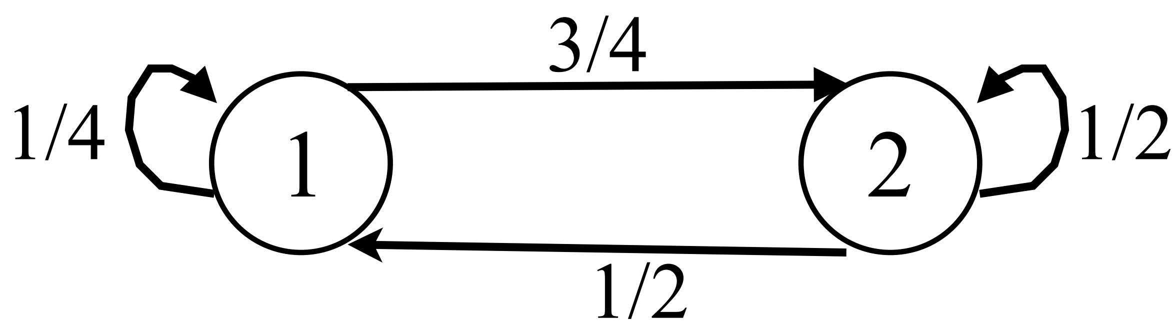


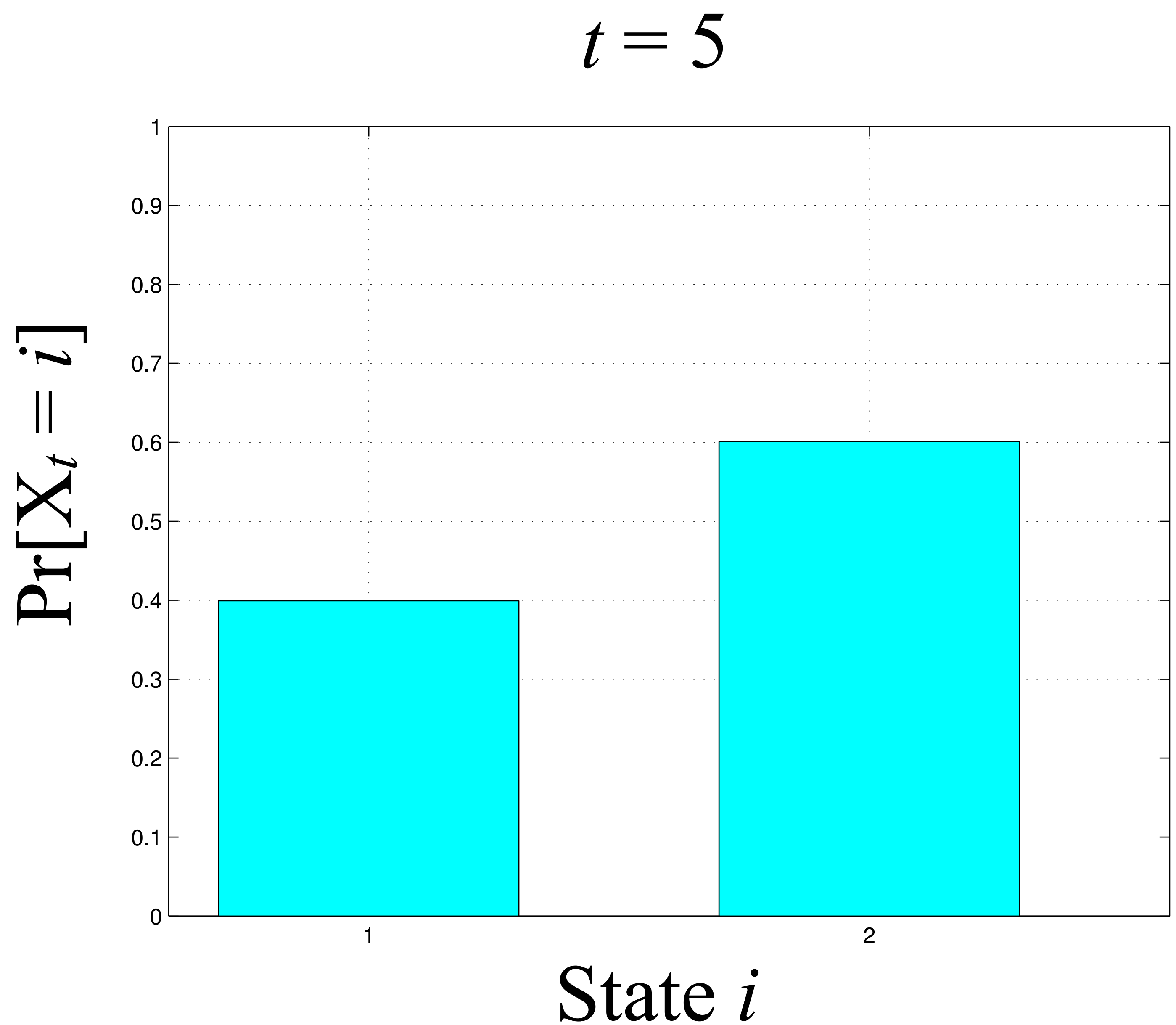
Example of stationary distribution



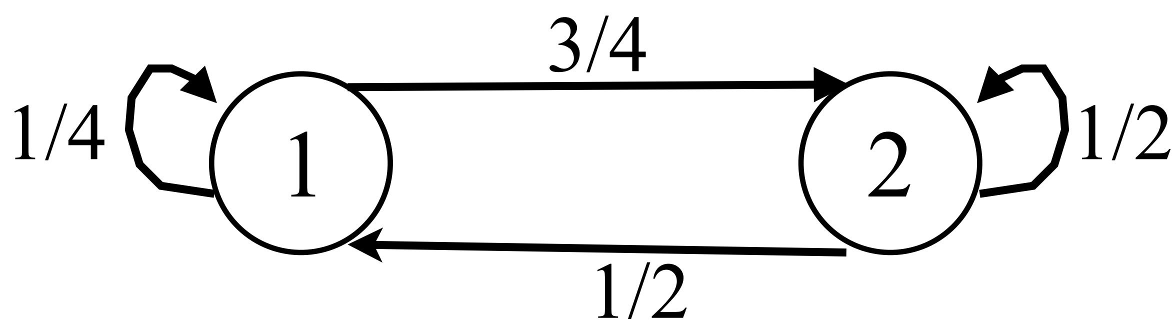


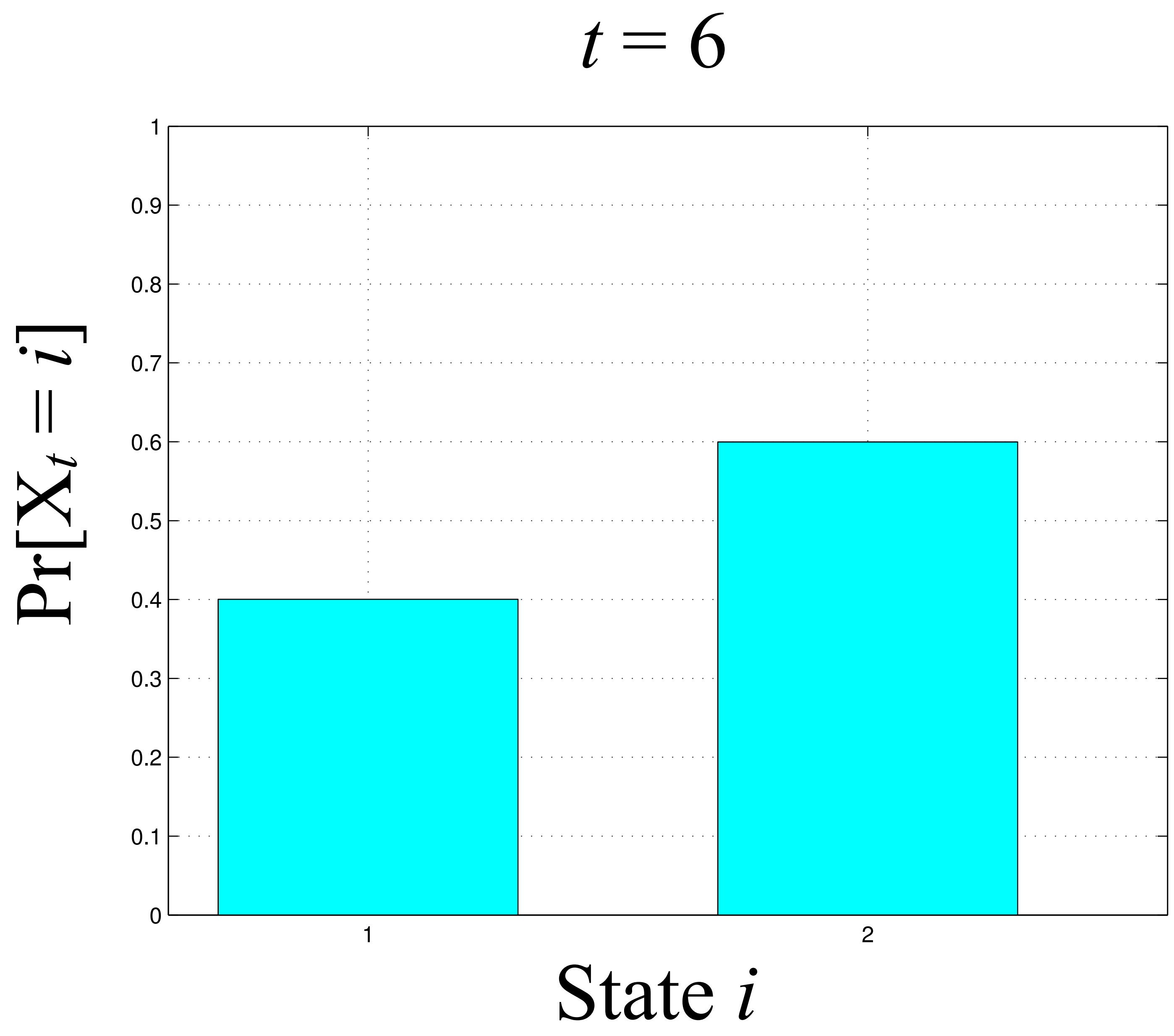
Example of stationary distribution



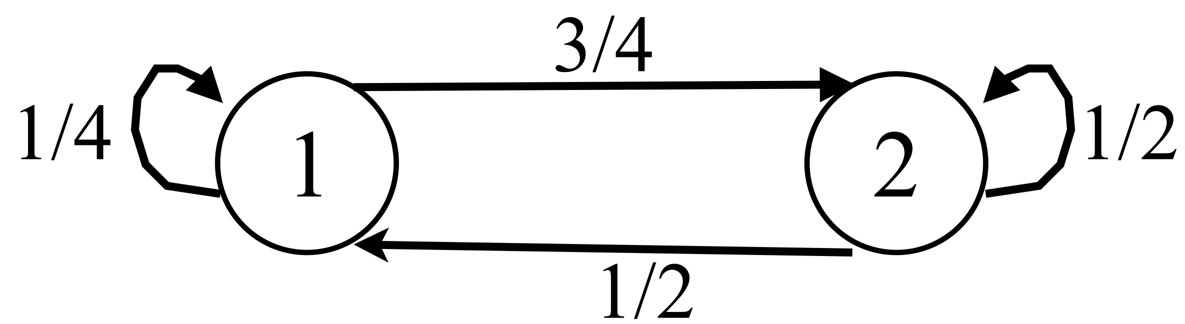


Example of stationary distribution

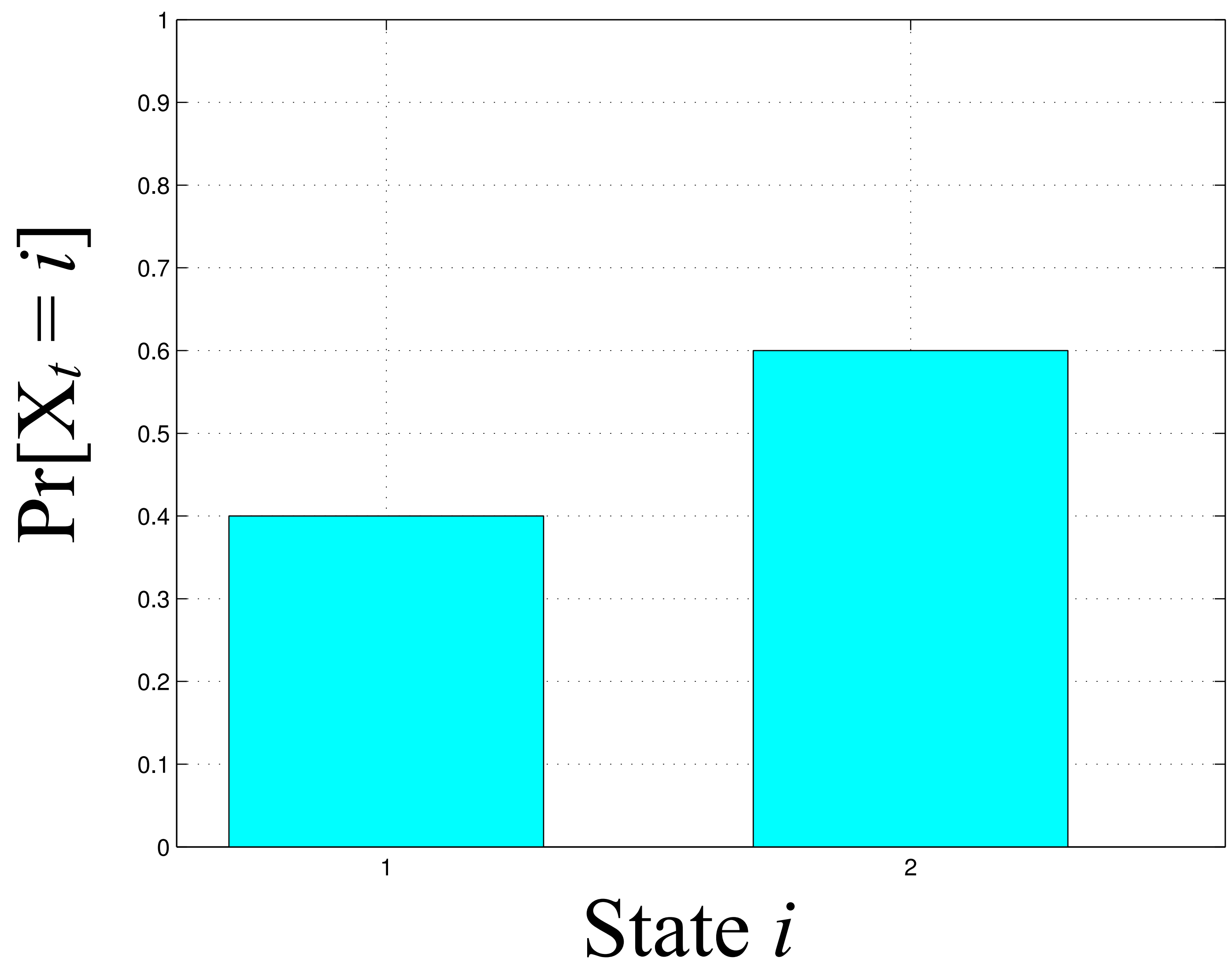




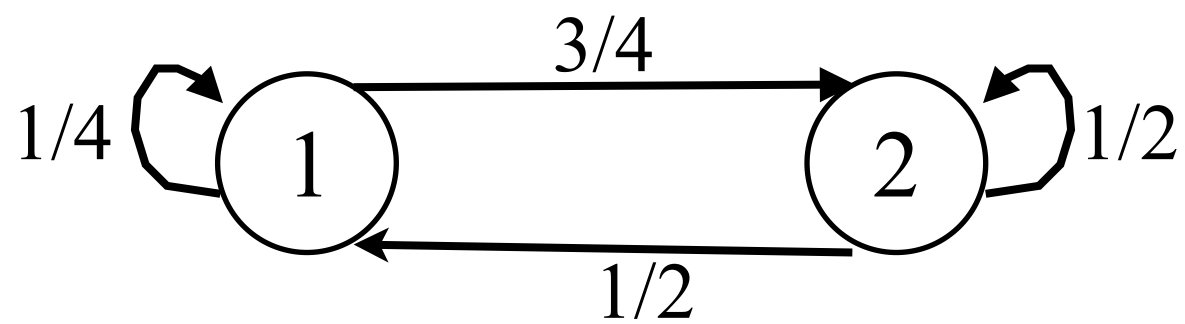
Example of stationary distribution

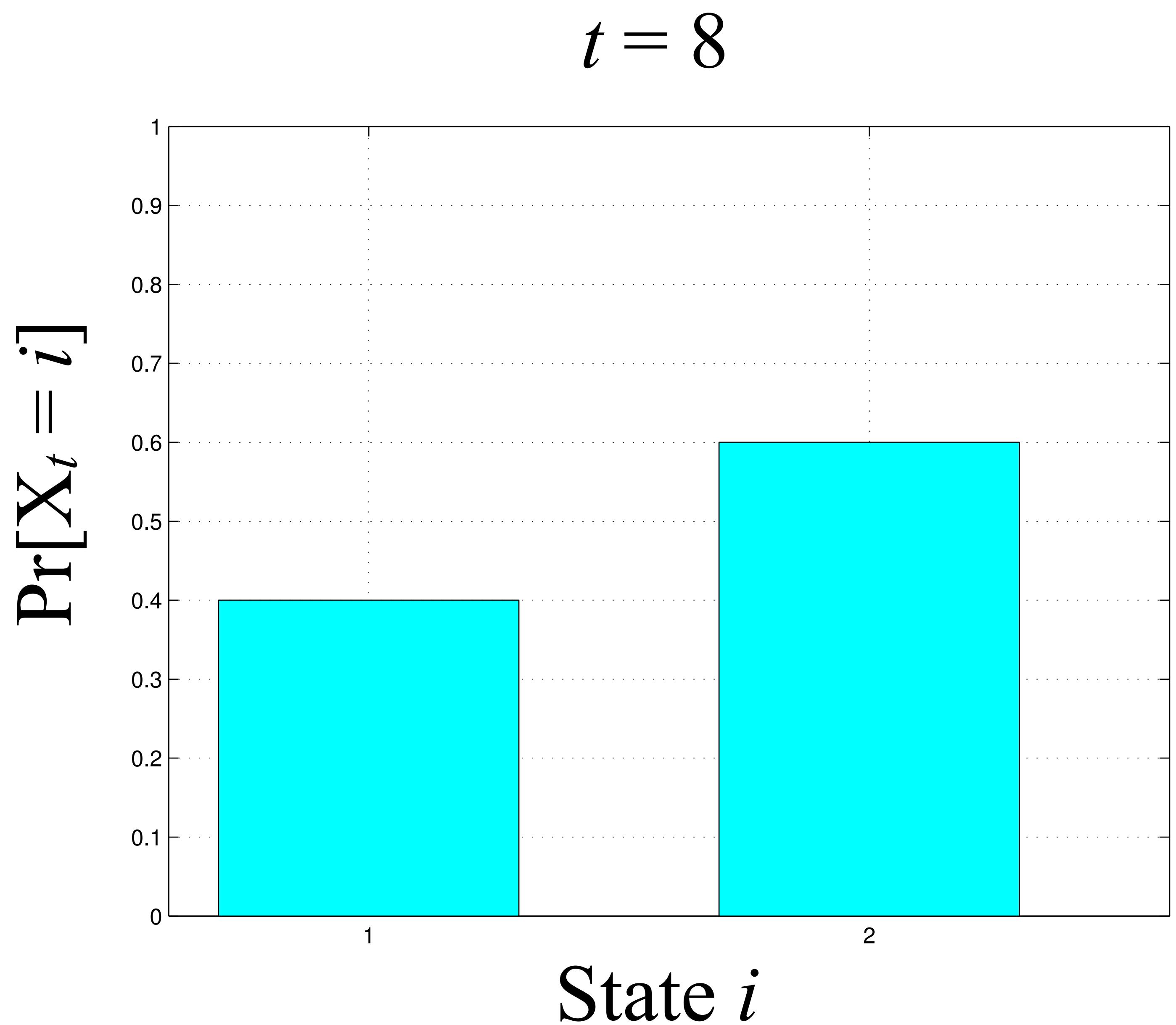


$t = 7$

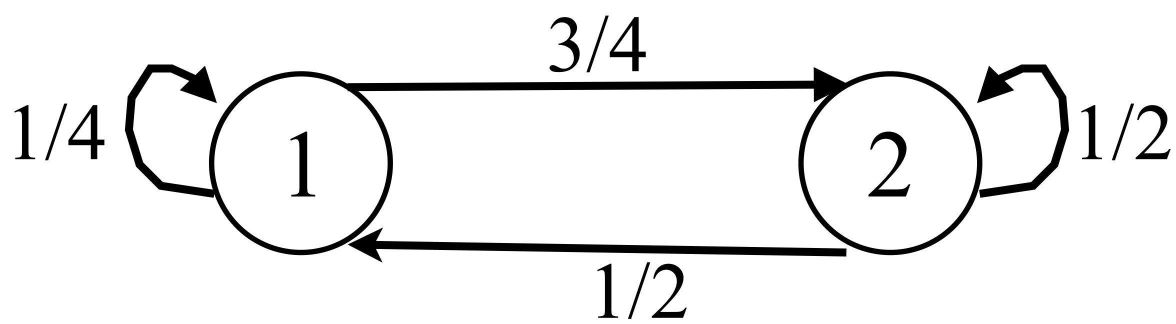


Example of stationary distribution

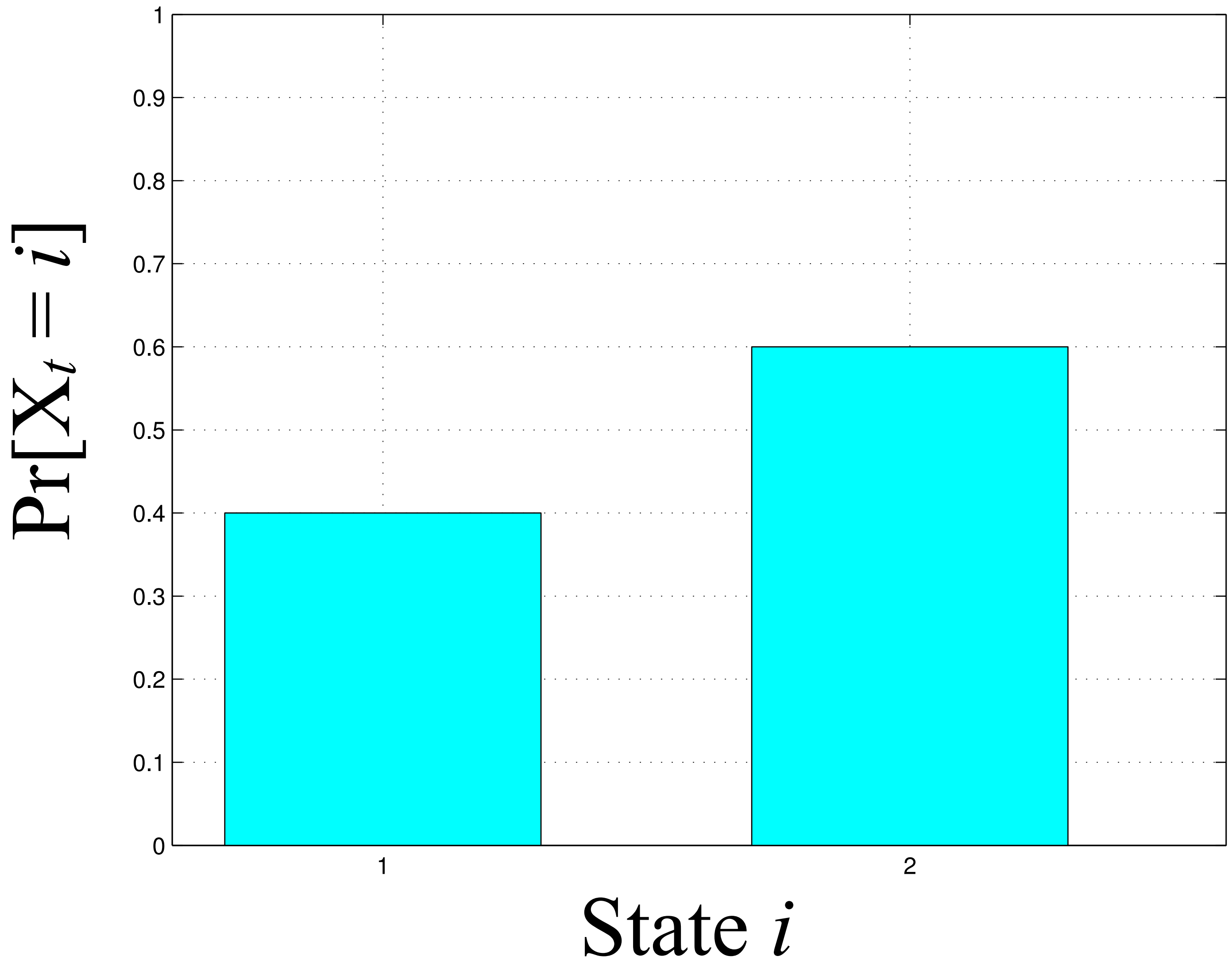




Example of stationary distribution



$t = 9$

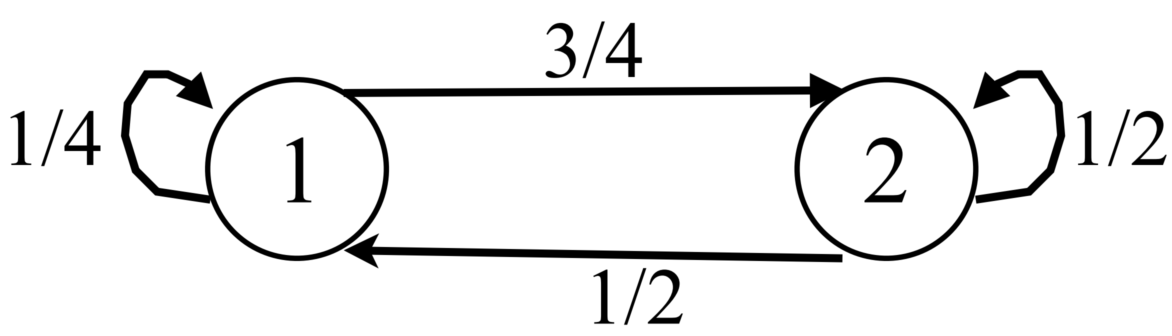


Three ways to find $\boldsymbol{\pi}$, part 1

- Stationary distribution is the limit probability
- We can find $\boldsymbol{\pi}$ by computing the probabilities over time until they converge
- The converged distribution is the stationary distribution
- This is called the *power method*
  - Start with arbitrary initial state $\boldsymbol{v}$
  - Compute $\boldsymbol{v}\mathbf{P}^1$, $\boldsymbol{v}\mathbf{P}^2$, $\boldsymbol{v}\mathbf{P}^3$, ..., until it converges
  - If convergence happens at step t, $\boldsymbol{\pi} = \boldsymbol{v}\mathbf{P}^t$
- We can define how accurately we want to compute $\boldsymbol{\pi}$

Three ways to find **π**, part 2

- **π** is stationary distribution if **πP** = **π**
- This defines a system of linear equations, which can be solved to find **π**
  - Add $\sum \pi_i = 1$ to get proper distribution
- Example:



$$\frac{1}{4}\pi_1 + \frac{1}{2}\pi_2 = \pi_1$$
$$\frac{3}{4}\pi_1 + \frac{1}{2}\pi_2 = \pi_2$$
$$\pi_1 + \pi_2 = 1$$

Three ways to find $\boldsymbol{\pi}$, part 3

- Given a square matrix **A**, vector $\boldsymbol{v}$ is its **left eigenvector** if $\boldsymbol{v}\mathbf{A} = \lambda\boldsymbol{v}$ for some scalar λ
  - Scalar λ is the **eigenvalue** associated to $\boldsymbol{v}$
- Therefore stationary distribution $\boldsymbol{\pi}$ of a Markov chain with transition matrix **P** is the normalized left eigenvector of **P** that has eigenvalue $\lambda = 1$
  - Very similar to solving the linear equation group, but more specialized

Three ways to find $\boldsymbol{\pi}$, bonus part

- If $\sum_i \pi_i = 1$ and $\pi_i p_{ij} = \pi_j p_{ji}$ for all i and j, then $\boldsymbol{\pi}$ is a stationary distribution
  - Sufficient but not necessary condition
  - If this holds, the Markov chain is (time) *reversible*
- In general, to check that some distribution $\boldsymbol{\pi}$ is a stationary distribution, just check that it satisfies

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$$

PageRank algorithm

- The random surfer:
  - A random surfer goes to a random web page
  - Clicks a random link to move to other web page
  - Repeats ad infinitum

PageRank algorithm

- The random surfer:
  - A random surfer goes to a random web page
  - Clicks a random link to move to other web page
  - Repeats ad infinitum
- This corresponds to a Markov chain with pages as states

PageRank algorithm

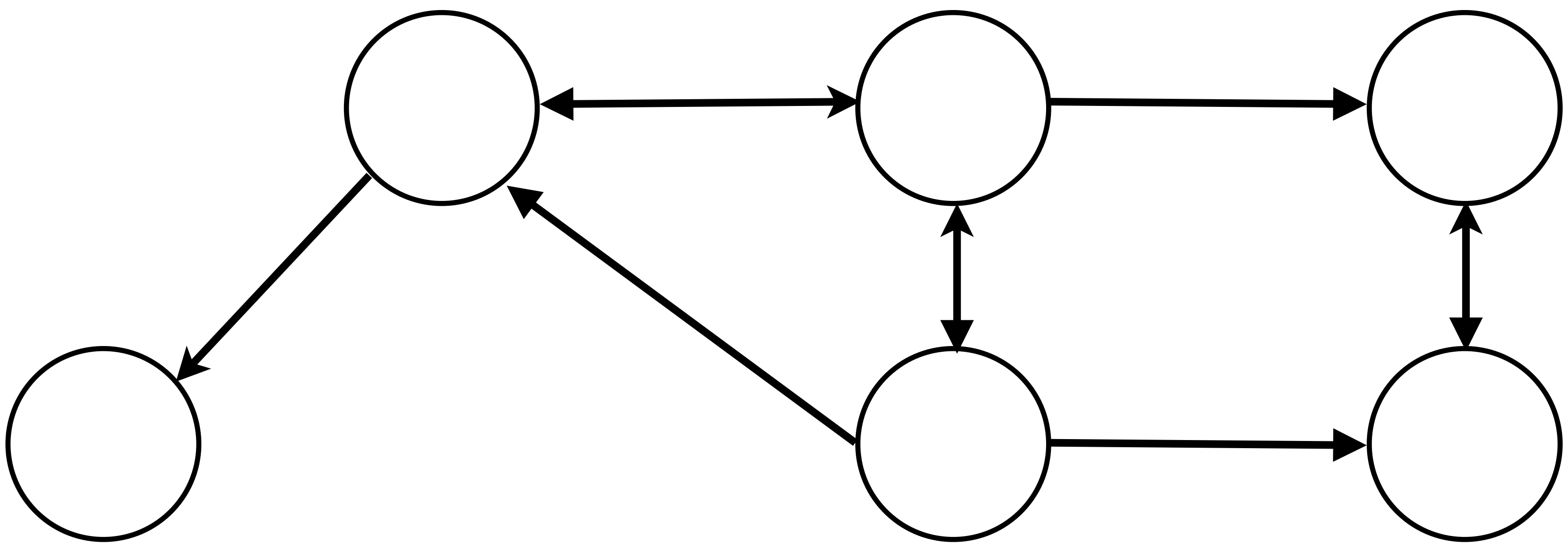
- The random surfer:
  - A random surfer goes to a random web page
  - Clicks a random link to move to other web page
  - Repeats ad infinitum
- This corresponds to a Markov chain with pages as states
- But we have a problem

Dead ends

Images: Wikipedia

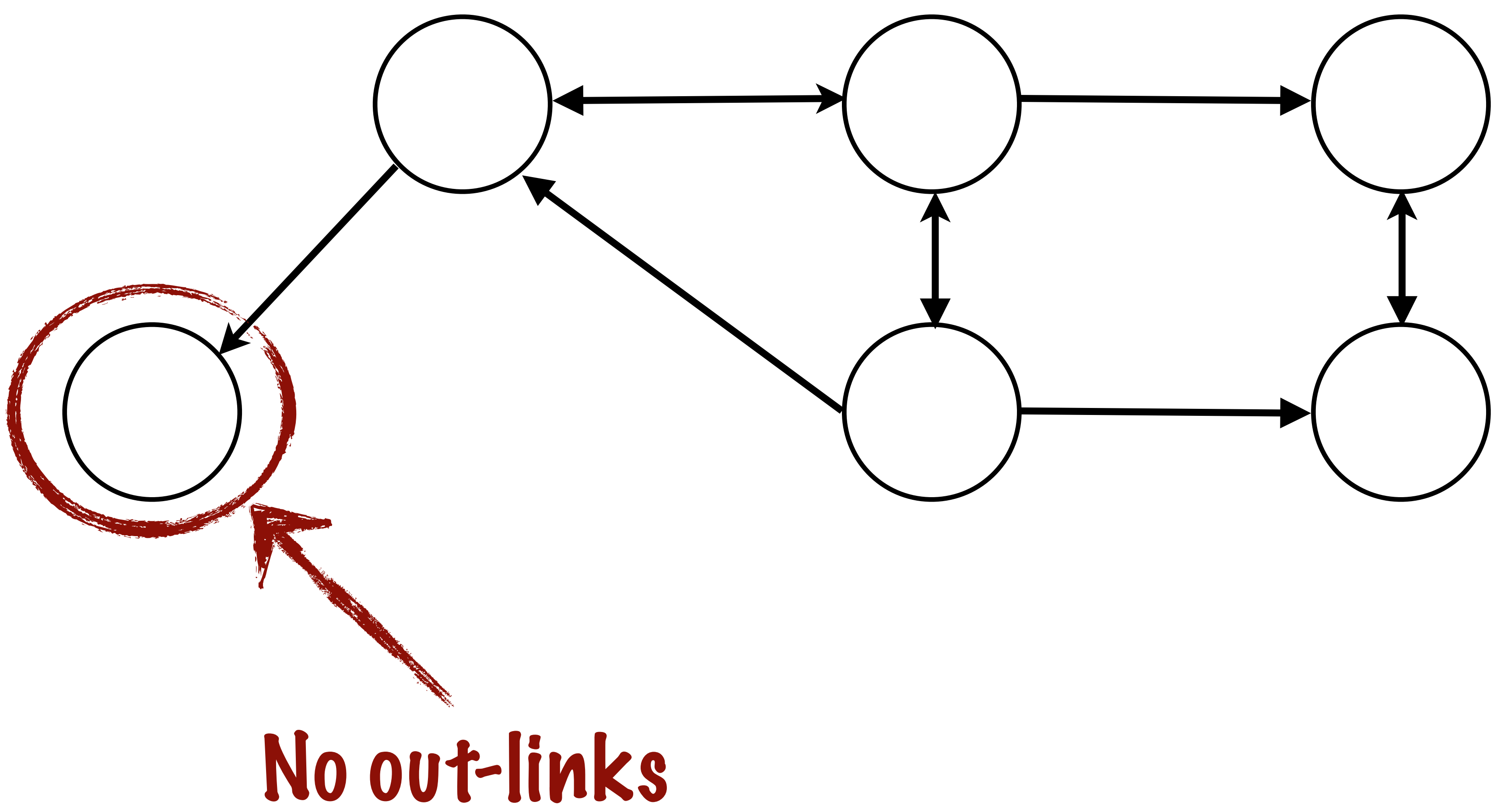
Dead ends

- Surfer can end in pages that have no out-links
- Surfer can end in a part of we where he cannot return to the other parts of the web



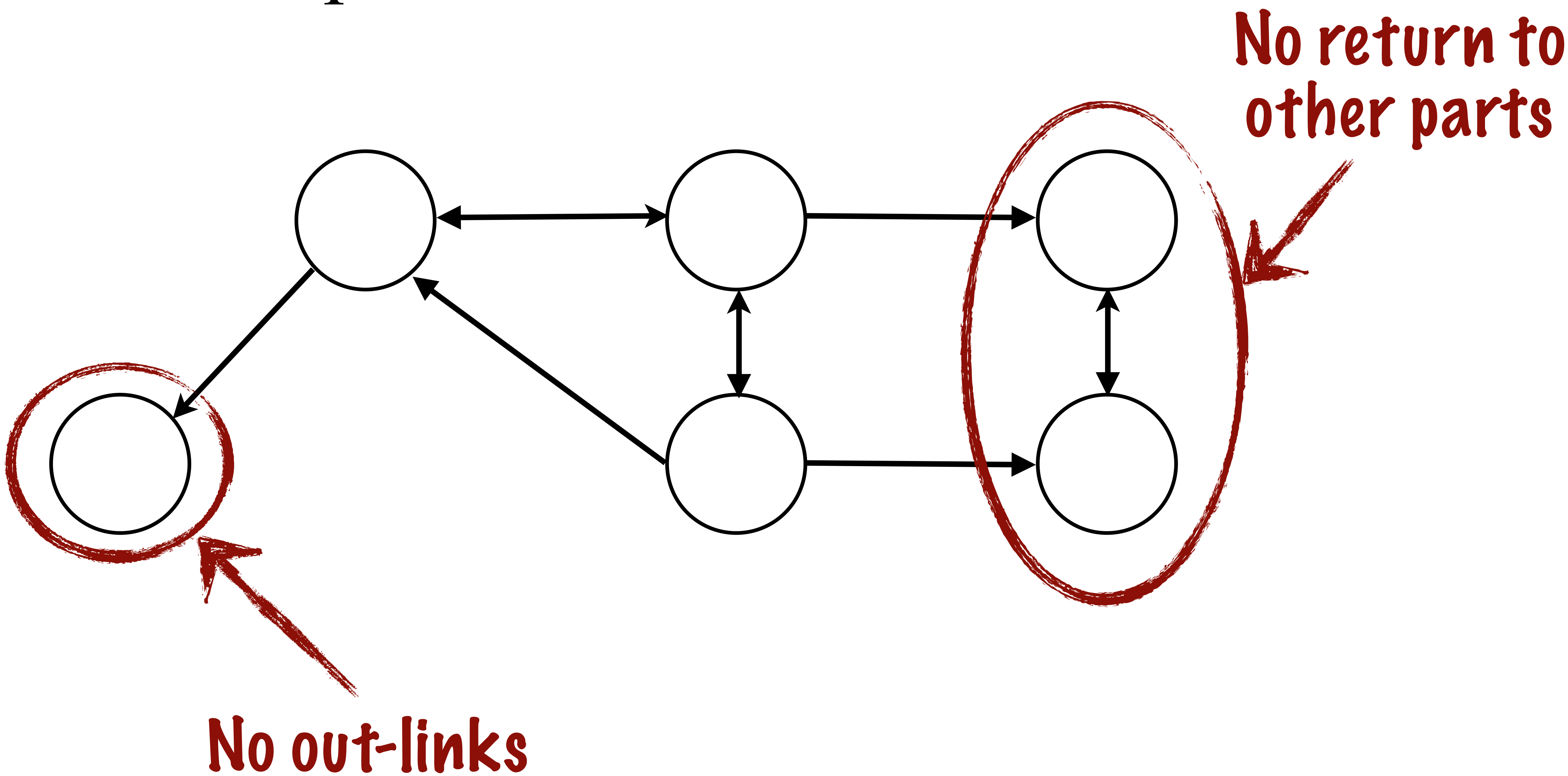
Dead ends

- Surfer can end in pages that have no out-links
- Surfer can end in a part of we where he cannot return to the other parts of the web



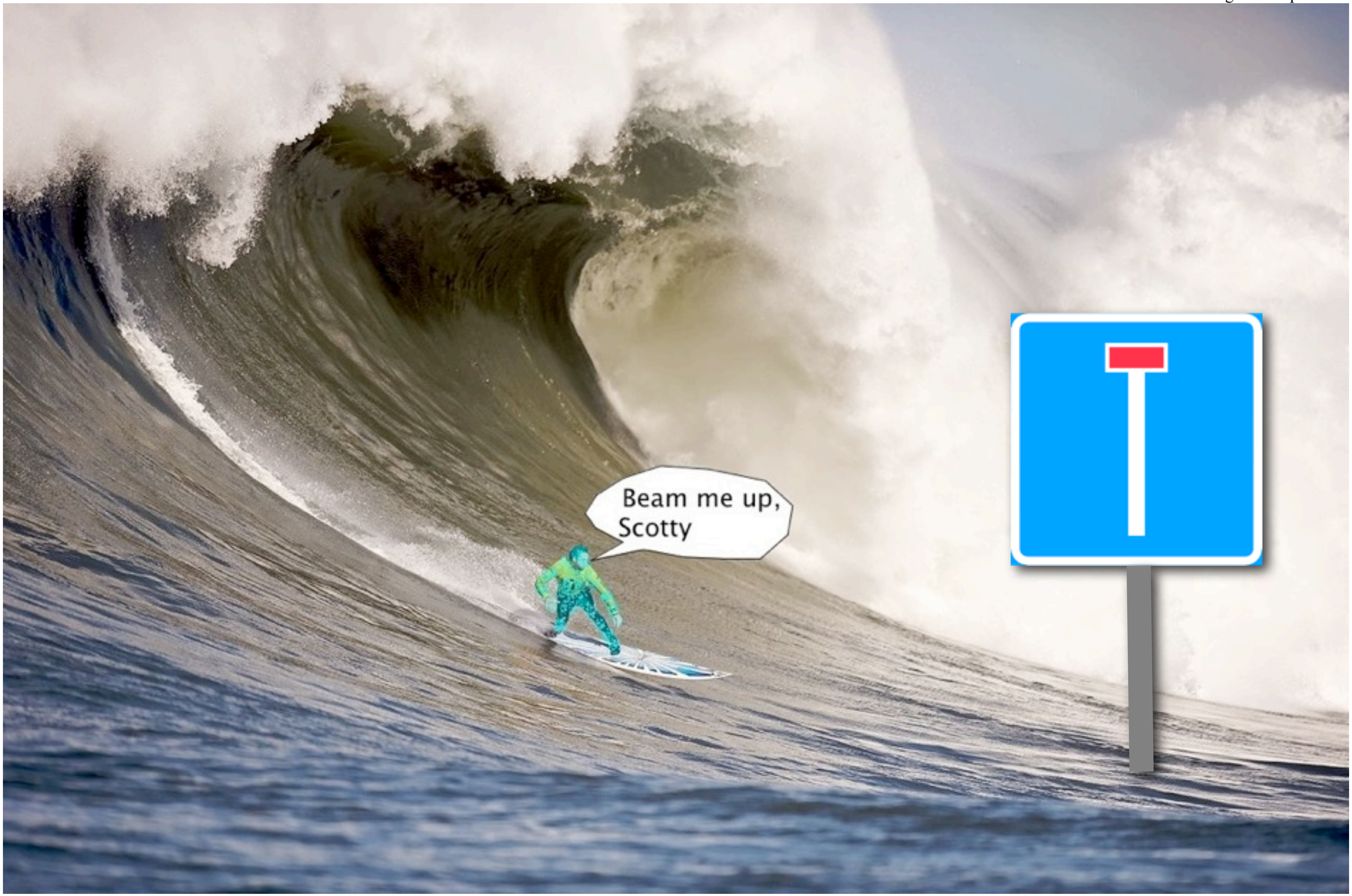
Dead ends

- Surfer can end in pages that have no out-links
- Surfer can end in a part of we where he cannot return to the other parts of the web



Answer: Teleportation

Images: Wikipedia



Teleportation

- At each step, if page has out-links, random surfer
  - with probability α selects new page uniformly at random among *all* web pages
  - with probability $1 - \alpha$, selects new page uniformly at random among the pages this page links to
- Otherwise random surfer selects new page u.a.r. among all pages
- Parameter α, $0 < \alpha < 1$, is fixed (e.g. $\alpha = 0.1$)
- Teleportation corresponds to the user typing new address in the address bar of the browser

Computing the PageRank

- Given a directed graph of N hyperlinked documents
  - Form the N-by-N adjacency matrix $\mathbf{A} = (a_{ij})$, where $a_{ij} = 1$ if page i links to page j
  - For rows of $\mathbf{A}$ that have no 1s
    - Replace each element with $1/N$
  - For other rows
    - If row has k 1s, multiply every entry with $(1 - \alpha)/k$
    - Add α/N to every entry
  - The resulting matrix $\mathbf{P}$ is a transition matrix of N-state, irreducible, and ergodic Markov chain that has stationary distribution $\boldsymbol{\pi}$
  - The PageRank of page i is π_i

PageRank and queries

- PageRank does *not* depend on the query
  - Establishes a static ordering between web pages
- To rank the query results, search engines need to combine query-dependent rankings with static ranking such as PageRank