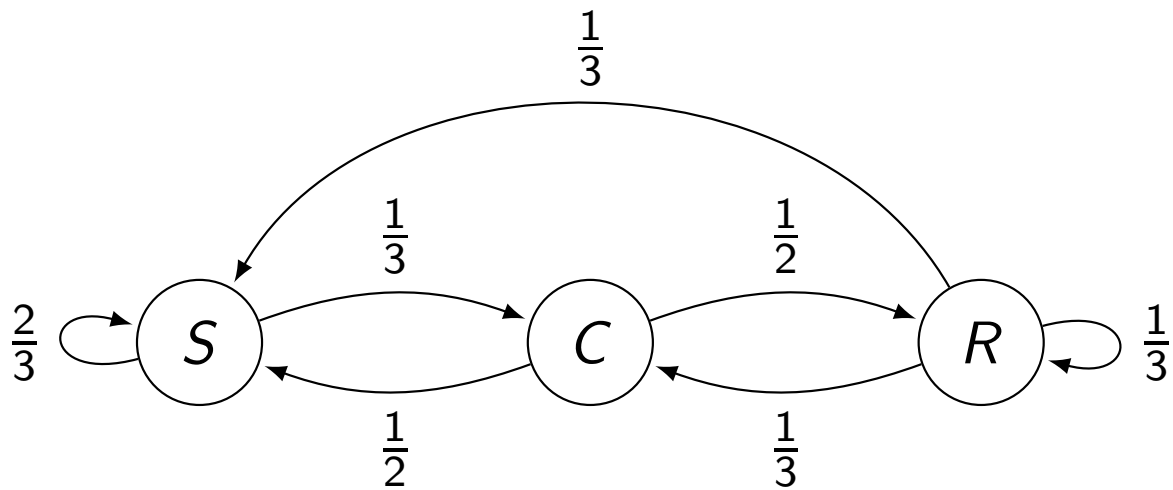
From Automata to Stochastic Systems

- So far: our models had *no notion of uncertainty* – the current state and the input uniquely fix what happens next.
- In practice, the behavior of many systems is uncertain. They are *stochastic*.
  - One example: the weather.
  - Many others: phone calls and requests arriving, computation times of tasks, queues in computer networks, failures, ...
- Modeling and reasoning about such systems is just as important as for automata.
  - We would like to answer questions such as: "What is the weather likely to be *in three days*?"
- How do we do that? $\rightarrow$ **Markov chains**

A First Example: The Weather in Zürich

- Each day is sunny (S), cloudy (C), or rainy (R).
- **Key assumption:** tomorrow's weather depends only on today's, with fixed probabilities:



An edge $a \to b$ is labelled with $\Pr[\text{tomorrow} = b \mid \text{today} = a]$.

The Transition Matrix and Initial Distribution

A handy way to collect the probabilities is through a **transition matrix** P:

$$P = \begin{array}{c|ccc} & S & C & R \\ \hline S & 2/3 & 1/3 & 0 \\ C & 1/2 & 0 & 1/2 \\ R & 1/3 & 1/3 & 1/3 \end{array}$$

- Entry $p_{i,j} = \Pr[\text{tomorrow} = j \mid \text{today} = i]$.
- Each row necessarily sums to 1.

- Alongside P, we typically also fix an **initial distribution** q_0 over the states, e.g. $q_0 = (1, 0, 0)$ if today is sunny.

State Machine vs. Markov Chain

	Finite automaton	**Markov chain**
description	$(Q, \Sigma, \delta, q_0, F)$	diagram, i.e. P and q_0
state	$q \in Q$	$X_t \in S$ — a *random variable*
# states	finite	finite *or* countably infinite
transition	$a \in \Sigma$	$p_{i,j} = \Pr[\text{next} \mid \text{now}]$
chosen by	an input symbol	chance
start	initial state q_0	initial distribution q_0
one run	a string $\to$ accept?	a sequence $X_0, X_1, X_2, \ldots$

- The Markov chain is the *random run* $X_0, X_1, X_2, \ldots$ the machine produces with one random variable per time step.

Three Natural Questions . . .

Suppose today is sunny. We would like to answer:

Q1 What is the weather likely to be *in three days*?

Q2 How long until the *first rainy day*, on average?

Q3 Over many years, what *fraction* of days are sunny, cloudy, rainy?

. . . and their answers

Q1 Multiply the distribution by P once per day ($q_{t+1} = q_t \cdot P$):

day t	sunny	cloudy	rainy
0	1	0	0
1	0.667	0.333	0
2	0.611	0.222	0.167
3	0.574	0.259	0.167

$\Rightarrow$ in three days about 57% sunny.

$$\Pr[\text{sunny on day 2}] = \underbrace{\tfrac{2}{3} \cdot \tfrac{2}{3}}_{S \to S \to S} + \underbrace{\tfrac{1}{3} \cdot \tfrac{1}{2}}_{S \to C \to S} = 0.611.$$

Q2 Expected **hitting time** sunny $\to$ rainy $= 8$ days.

Q3 The **stationary distribution** $\pi = (\frac{9}{16}, \frac{4}{16}, \frac{3}{16})$: in the long run 9 of every 16 days are sunny.

The rest of the chapter makes these notions precise.

Definition: Markov Chain

- Let S be a finite or countably infinite set of states.
- A discrete-time **Markov chain** is a sequence of random variables $X_0, X_1, X_2, \ldots \in S$ with the **Markov property**: X_{t+1} depends only on X_t, not on the past $X_{t-1}, \ldots, X_0$ ("memoryless"):

  $$\Pr[X_{t+1} = s_{t+1} \mid X_0 = s_0, \ldots, X_t = s_t] = \Pr[X_{t+1} = s_{t+1} \mid X_t = s_t]$$

  for all t and all $s_0, \ldots, s_{t+1} \in S$.
- **Note:** memoryless is (usually) a simplification.
- The distribution of X_0 is the **initial distribution** $q_0 = (q_{0,s})_{s \in S}$ with $q_{0,s} = \Pr[X_0 = s]$.
- Time is discrete ($t \in \mathbb{N}_0$), for now. Continuous time ($t \in \mathbb{R}_0^+$) is the topic of the next chapter.

Time-Homogeneous Markov Chains

- If $\Pr[X_{t+1} = j \mid X_t = i]$ is independent of t for all $i, j \in S$, then the Markov chain is called **(time-)homogeneous**.
- We consider (almost) only time-homogeneous Markov chains.
- **Note:** real systems are often **not** homogeneous: rain after rain is likelier in winter, and a cafeteria sees more arrivals at noon. Homogeneity makes the chain tractable.
- For time-homogeneous Markov chains, the values

$$p_{i,j} := \Pr[X_{t+1} = j \mid X_t = i]$$

  are uniquely defined and yield the transition matrix P.
- **Note:** any non-homogeneous chain can be made homogeneous by making time part of the state.

Probability of a run

- Observation of a Markov chain from time 0 to time t_0.
- Possible run: states x_0, x_1, x_2, ..., x_{t_0}.
- Probability of this run (sample path):

$$q_{0,x_0} \cdot \Pr[X_1 = x_1 \mid X_0 = x_0] \cdot \ldots \cdot \Pr[X_{t_0} = x_{t_0} \mid X_{t_0-1} = x_{t_0-1}]$$

Example

Weather chain with $q_0 = (1, 0, 0)$ (today is sunny):

$\Rightarrow$ probability of the run (S, S, C, R) is

$$1 \cdot \tfrac{2}{3} \cdot \tfrac{1}{3} \cdot \tfrac{1}{2} = \tfrac{1}{9} \approx 0.111.$$

Probability distribution over the states (1)

- **Goal:** compute the probability distribution over the states at an arbitrary time t from the initial distribution.
- Initial distribution: q_0 (row vector with n entries)
- Distribution at time t: $q_t = (q_{t,i})_{i \in S}$ with $q_{t,i} = \Pr[X_t = i]$
- Computing q_{t+1} from q_t:

$$\begin{aligned} q_{t+1,j} &= \Pr[X_{t+1} = j] \\ &= \sum_{i \in S} \Pr[X_t = i] \cdot \Pr[X_{t+1} = j \mid X_t = i] = \sum_{i \in S} q_{t,i} \cdot p_{i,j} \end{aligned}$$

Written as a vector-matrix multiplication:

$$q_{t+1} = q_t \cdot P$$

Probability distribution over the states (2)

- So we know: $q_{t+1} = q_t \cdot P$.
- Then it must hold that:
$$q_1 = q_0 \cdot P$$
$$q_2 = q_1 \cdot P = q_0 \cdot P \cdot P = q_0 \cdot P^2$$
$$q_3 = q_2 \cdot P = q_0 \cdot P^2 \cdot P = q_0 \cdot P^3$$
$$\vdots$$
$$\boxed{q_t = q_0 \cdot P^t}$$
- Likewise: $\boxed{q_{t+k} = q_t \cdot P^k \text{ for all } k \geq 0}$
- The entry in row i and column j of P^k, denoted $p_{i,j}^{(k)} := (P^k)_{i,j}$, gives the probability of going from state i to state j in k steps.

Transient Analysis

- What is the probability of being in state j after k steps?
- How likely is it to eventually get from i to j?
- How many steps does the chain need on average to get from i to j?

Many of these questions can be answered with the help of the equation

$$q_{t+k} = q_t \cdot P^k \text{ for all } k \geq 0$$

Sojourn Time

- It is sunny today, how many days will it *stay* sunny?
- The number of time steps the chain stays in state i (from time t on) is a random variable: the **sojourn time** T_i.
- It holds that $\Pr[T_i = k] = p_{i,i}^{k-1}(1 - p_{i,i})$ and $\Pr[T_i > k] = p_{i,i}^{k}$:
  T_i is **geometrically distributed**, so $\mathbf{E}[T_i] = \frac{1}{1-p_{i,i}}$.
- **Note:** The sojourn time is independent of how long the chain has already been in state i (memoryless).
- E.g. the weather chain stays sunny for $\frac{1}{1-2/3} = 3$ days on average; cloudy has no self-loop ($p_{C,C} = 0$), so it is left after a single day.

Hitting Time and Arrival Probability

- If it is *sunny* today, how long until the first *rainy* day?
- **Hitting time** $T_{i,j}$ from state i to j:

$$T_{i,j} := \min\{t \geq 1 \mid X_t = j \text{ if } X_0 = i\}$$

(if state j is never reached, $T_{i,j} = \infty$)

- $h_{i,j} := \mathbf{E}[T_{i,j}]$ is the **expected hitting time** from i to j
- $f_{i,j} := \Pr[T_{i,j} < \infty]$ is the **arrival probability** from i to j.
- $c_{i,j} := h_{i,j} + h_{j,i}$ is the **commute time** between i and j (there and back)

Computing the Hitting Time and Arrival Probability

Lemma

For the ***expected hitting time****, for all* $i, j \in S$ *it holds*

$$h_{i,j} = 1 + \sum_{k \neq j} p_{i,k} h_{k,j},$$

if $h_{i,j}$ *and* $h_{k,j}$ *exist.*

For the ***arrival probabilities****, analogously,*

$$f_{i,j} = p_{i,j} + \sum_{k \neq j} p_{i,k} f_{k,j}.$$

Proof for $h_{i,j} = 1 + \sum_{k \neq j} p_{i,k} h_{k,j}$

$$
\begin{aligned}
h_{i,j} &= \mathbf{E}[T_{i,j}] = \sum_{k \in S} \mathbf{E}[T_{i,j} \mid X_1 = k] \cdot p_{i,k} \\
&= \mathbf{E}[T_{i,j} \mid X_1 = j] \cdot p_{i,j} + \sum_{k \neq j} \mathbf{E}[T_{i,j} \mid X_1 = k] \cdot p_{i,k} \\
&= 1 \cdot p_{i,j} + \sum_{k \neq j} (1 + \mathbf{E}[T_{k,j}]) \cdot p_{i,k} \\
&= 1 + \sum_{k \neq j} \mathbf{E}[T_{k,j}] \cdot p_{i,k} = 1 + \sum_{k \neq j} p_{i,k} h_{k,j}
\end{aligned}
$$

□

Example: Applied to the weather chain (S, C, R)

We want the expected hitting time $h_{S,R}$. With $h_{i,R} = 1 + \sum_{k \neq R} p_{i,k} h_{k,R}$:

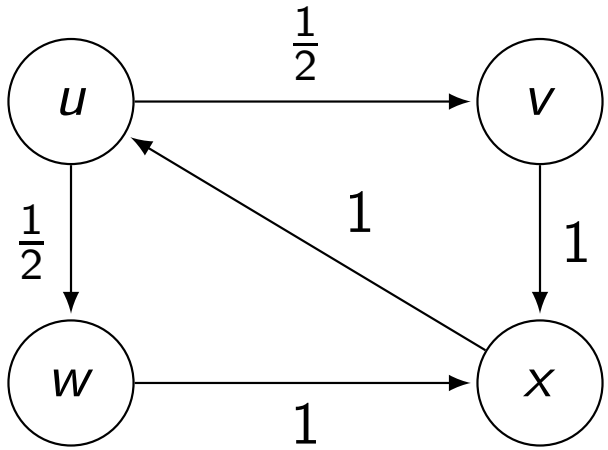
$$\begin{aligned} h_{S,R} &= 1 + \tfrac{2}{3} h_{S,R} + \tfrac{1}{3} h_{C,R} \\ h_{C,R} &= 1 + \tfrac{1}{2} h_{S,R} \end{aligned}$$

Solving gives $h_{C,R} = 5$ and $\boxed{h_{S,R} = 8}$: starting on a sunny day, it takes on average **8 days** until the first rainy day.

Stationary Analysis

- Real dynamical systems often run for a long time.
- Consider the behavior for $t \to \infty$.
- The probability distribution over the states of the Markov chain at time t is $q_t = q_0 \cdot P^t$. **Convergence?**
- Intuitively if q_t converges to a vector π as $t \to \infty$, then π should satisfy the equation $\pi = \pi \cdot P$.
- **Definition.** A state vector π with $\pi_j \geq 0$ and $\sum_{j \in S} \pi_j = 1$ is called a **stationary distribution** of the Markov chain with transition matrix P if $\pi = \pi \cdot P$.
- **Note:** The stationary distribution π is an eigenvector of P for the eigenvalue 1.

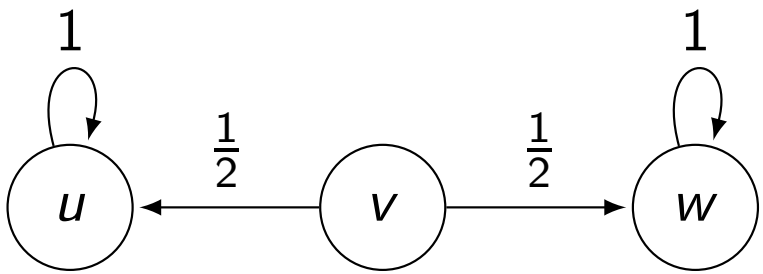
Stationary Distribution



What is π_v? $\frac{1}{4}$ $\frac{1}{5}$ $\frac{1}{6}$ none of these

(Non-)Uniqueness of the Stationary Distribution

- Consider this Markov chain:



$$P = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \end{pmatrix}$$

- This Markov chain has several stationary distributions: for example $(1, 0, 0)$, $(0, 0, 1)$, and $(0.2, 0, 0.8)$
- Reason: states u and w are "absorbing".

Irreducible Markov Chains

Definition. A Markov chain is called **irreducible** if all states are reachable from all other states. That is, if for all states $i, j \in S$ there is some $t \in \mathbb{N}$ such that $p_{i,j}^{(t)} > 0$.

Theorem

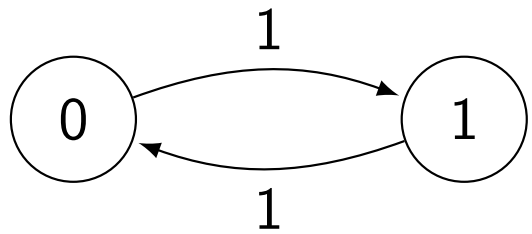
*An irreducible finite Markov chain has a **unique stationary distribution** π, and $\pi_j = 1/h_{j,j}$ for all $j \in S$.*

Intuition. On average, state j is visited every $h_{j,j}$ steps.

Question: Does an irreducible finite Markov chain always converge to its stationary distribution?

Convergence?

Question: Does an irreducible finite Markov chain always converge to its stationary distribution? **NO!**



This chain is irreducible and finite, but the state vector q_t does not necessarily converge as $t \to \infty$:

$q_0 = (1, 0),\ q_1 = (0, 1),\ q_2 = (1, 0),\ q_3 = (0, 1),\ \ldots$

Reason: periodicity!

Aperiodic Markov Chains

- The **period** of a state $j \in S$ is the largest number $\xi \in \mathbb{N}$ s.t.:

$$\{n \in \mathbb{N} \mid p_{j,j}^{(n)} > 0\} \subseteq \{i \cdot \xi \mid i \in \mathbb{N}\}$$

  i.e. ξ is the greatest common divisor of all possible return times to j.

- A state with period $\xi = 1$ is called **aperiodic**.
- A Markov chain is called **aperiodic** if all states are aperiodic.
- **Useful test condition:** state j is aperiodic if one of the following two conditions holds:
  - $p_{j,j} > 0$
  - $\exists n, m \in \mathbb{N}: \; p_{j,j}^{(m)}, p_{j,j}^{(n)} > 0$ and $\gcd(m, n) = 1$
- **Example:** a state on a 2-cycle and a 3-cycle can return in $2, 3, 4, 5, \ldots$ steps: aperiodic.

Ergodic Markov Chains

- **Finite**, irreducible, and aperiodic Markov chains are called **ergodic**.

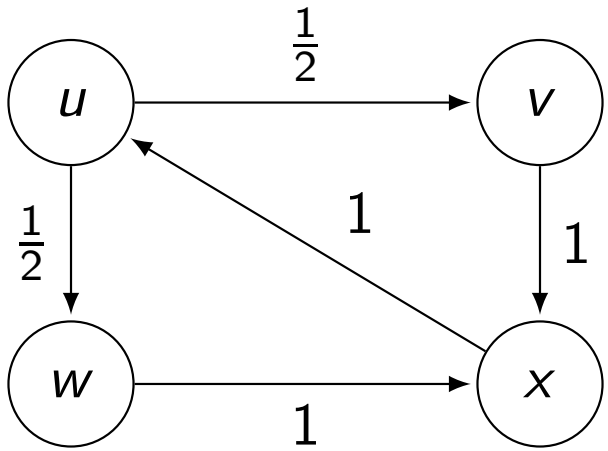
Theorem (Fundamental Theorem for Ergodic Markov Chains)

For every ergodic Markov chain, independently of the initial state,

$$\lim_{t \to \infty} q_t = \pi,$$

where π is the unique stationary distribution of the chain.

Irreducible? Aperiodic? Ergodic?



Is this chain irreducible, aperiodic, ergodic, or none of these?

Example: Long-run behavior of the weather chain

- The chain is irreducible and aperiodic, hence **ergodic**.
- The matrix powers converge quickly:

$$P^3 \approx \begin{pmatrix} 0.574 & 0.259 & 0.167 \\ 0.556 & 0.222 & 0.222 \\ 0.537 & 0.259 & 0.204 \end{pmatrix}, \quad P^{10} \approx \begin{pmatrix} 0.563 & 0.250 & 0.187 \\ 0.562 & 0.250 & 0.187 \\ 0.562 & 0.250 & 0.188 \end{pmatrix}$$

- Solving $\pi = \pi P$ with $\pi_S + \pi_C + \pi_R = 1$ gives the unique stationary distribution

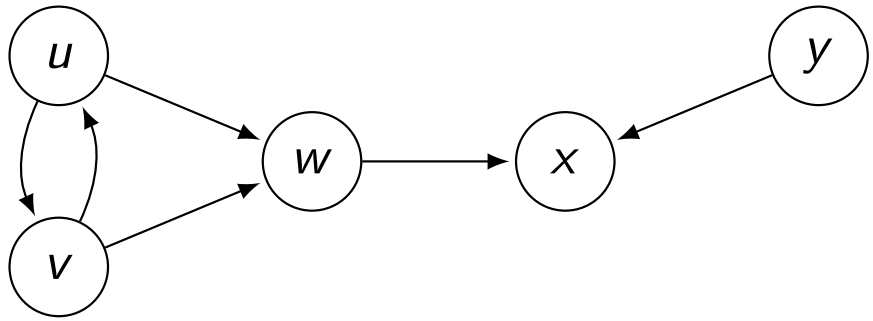
$$\pi = \left(\tfrac{9}{16}, \tfrac{4}{16}, \tfrac{3}{16}\right) = (0.5625,\ 0.25,\ 0.1875).$$

- In the long run 9 of every 16 days are sunny, independently of today's weather.

PageRank: How Does Google Rank Web Pages?

- A Google search does two things:
  - *filter*: find the pages that match the search terms (easy);
  - *rank*: show the "important" pages first (hard).
- The ranking is computed *once*, independently of the query, from the **web graph** alone: nodes are pages, edge (u, v) means u has a hyperlink to v.
- **Google's answer (Brin & Page, 1998): PageRank**, which models the web graph as a Markov chain and ranks pages by its stationary distribution.
- **Next:** what is the Markov chain, and what can go wrong with it?

Example: Which Page Is the Most Important?

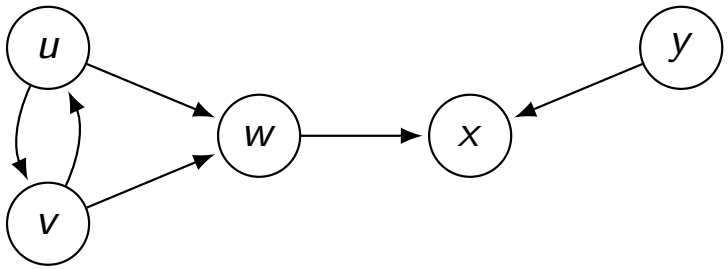


- A tiny web graph with 5 pages. **Which page is the most important?**
- *Naïve approach:* rank by the number of incoming links. Here w and x tie with 2 links each, u and v have 1, y has none.
- *But:* x is linked by w, an important page; y is linked by nobody. Shouldn't a link from an important page count more?
- And the importance of w depends in turn on who links to w, ... a *recursive* notion.

Google's Idea: The Random Surfer

- A **random surfer** starts on some page and, in each step, clicks on a *uniformly random* link of the current page.
- Pages visited often are important: many links point to them, from pages that are themselves visited often.
- This is a **random walk** on the web graph, i.e. a Markov chain with the **random surfer matrix** W: $w_{i,j} = 1/\delta(i)$ if page i links to page j, else 0, where $\delta(i)$ is the number of links on page i.
- The **importance of a page** is the fraction of time the surfer spends on it in the long run: the **stationary distribution** $\pi = \pi \cdot W$.

Example: The Random Surfer Matrix



$$W = \begin{array}{c} \\ u \\ v \\ w \\ x \\ y \end{array} \begin{array}{c} \begin{array}{ccccc} u & v & w & x & y \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

- Row i spreads the surfer evenly over the links of page i: u has two links, so $w_{u,v} = w_{u,w} = 1/2$; w has one, so $w_{w,x} = 1$.
- Each row sums to 1, as for a transition matrix ... except row x, which sums to 0: page x has no outgoing link.

Pages Without Outgoing Links

- Some pages have **no outgoing links** ("sinks"), such as x in the example: row x of W is all zeros, so W is not a transition matrix. The surfer is stuck.
- **What would you do?**

Fixing the Random Surfer

- **Idea:** let the surfer jump to a *uniformly random page* instead of following a link. Google uses this twice:
- **1. On a sink, always jump.** Row x of W becomes $(1/n, \ldots, 1/n)$, so W is a transition matrix again and the surfer is never stuck.
- **2. Anywhere, jump with a small probability** $1 - \alpha$; with probability α follow a random link as before (the surfer "gets bored").
  - Why? Even with the sink fix, nothing guarantees that the web chain is **ergodic**, so $q_t = q_0 \cdot W^t$ need not converge to a unique π.
  - The random jumps add an edge between every two pages: the chain becomes ergodic.

The Google Matrix

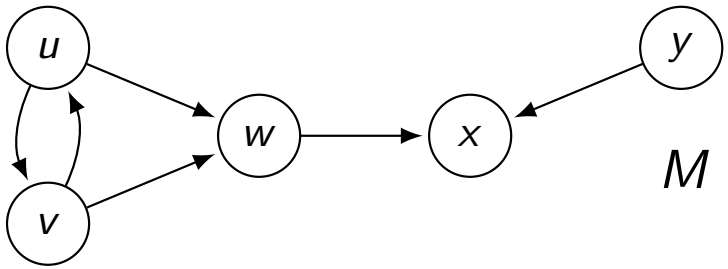
- **Definition.** For $\alpha \in (0,1)$ and R the $n \times n$ matrix with all entries $1/n$, the **Google matrix** is

$$M = \alpha \cdot W + (1-\alpha) \cdot R.$$

- M is ergodic: every page reaches every page (irreducible), and self-jumps exist (aperiodic). So $\pi = \pi \cdot M$ exists, is unique, and $q_t \to \pi$.

- The random jumps change π, so α is chosen close to 1; typically $\alpha = 0.85$.

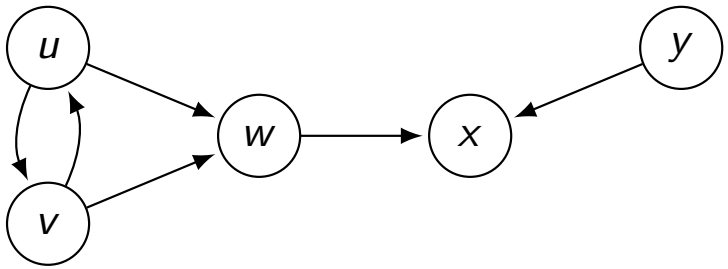
Example: Google Matrix of the Small Web Graph



$$M = \begin{array}{c} \\ u \\ v \\ w \\ x \\ y \end{array} \begin{array}{c} \begin{array}{ccccc} u & v & w & x & y \end{array} \\ \begin{pmatrix} 0.03 & 0.455 & 0.455 & 0.03 & 0.03 \\ 0.455 & 0.03 & 0.455 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.03 \\ 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.03 \end{pmatrix} \end{array}$$

- With $\alpha = 0.85$ and $n = 5$:
  $m_{i,j} = 0.85 \cdot w_{i,j} + 0.15/5 = 0.85 \cdot w_{i,j} + 0.03$.
- Row x (the sink) was already $(1/5, \ldots, 1/5)$, and stays so:
  $0.85 \cdot 0.2 + 0.03 = 0.2$.
- Every entry is now positive, so the chain is ergodic. But the links still dominate: from w the surfer goes to x with probability 0.88, elsewhere with 0.03 each.

Example: PageRank of the Small Web Graph



Website	Rank
x	0.384615
w	0.230769
u	0.153846
v	0.153846
y	0.0769231

- The ranks are the stationary distribution of the random surfer.
- x beats w although both have two incoming links, and y, linked by nobody, comes last.
- u and v tie: they are symmetric in the graph.

Computing PageRank

- M is ergodic, so $q_{t+1} = q_t \cdot M$ converges to π from *any* q_0: **just iterate.**
  - Fast: the error shrinks by a factor α per step, so after 100 steps it is $0.85^{100} \approx 10^{-7}$.
  - Cheap: $q \cdot M = \alpha\, q \cdot W + \frac{1-\alpha}{n}(1, \ldots, 1)$, and W has only one non-zero entry per link.
- The alternatives are worse for n in the billions:
  - solving $\pi = \pi \cdot M$ as a linear system costs $O(n^3)$;
  - simulating random surfers and counting visits needs far more steps than there are pages.
- This is how PageRank (Brin & Page, 1998) works. Google's recipe today is a company secret.

From PageRank to Random Walks

- The random surfer is a **random walk**: a Markov chain defined by a graph, where in each step one moves from the current node to a random neighbor.
- PageRank: a *directed* graph (the links), and π had to be computed numerically.
- Now the simplest case: an *undirected* graph, every neighbor equally likely: a **simple random walk**. Every edge can be used in both directions (though not with the same probability).
- Then much more can be said: π in closed form, and how long it takes to visit *every* node (the **cover time**).

Roomba: "Random Walk"



Random Walk in Graphs

- Given an **undirected connected graph** $G = (V, E)$ with n vertices and m edges, and a start vertex $s \in V$.
- In each time step, one moves **uniformly** to one of the neighboring vertices: a **simple random walk**.
- **Cover time** $\mathrm{cov}(s)$ = expected number of steps until, starting from s, all vertices in V have been visited.
- Connection between **cover time** and **hitting time**:
  - Markov chain with states (v, I) (v = current vertex, I = visited set) plus a sink state t; once everything is visited, $(v, V) \to t$ with probability 1
  - Then exactly $\mathrm{cov}(s) = h_{(s,\{s\}),\, t} - 1$

Stationary Distribution of Random Walks

Lemma

The stationary distribution is given by

$$\pi_u = \frac{\delta(u)}{2m} \quad (\delta(u) = \textit{degree of } u)$$

Proof.

Assumption: for all j, $\pi_j = \delta(j)/2m$. Then, for arbitrary u, it follows

$$\pi_u = \sum_{v \in N(u)} \pi_v \cdot p_{v,u} = \sum_{v \in N(u)} \frac{\delta(v)}{2m} \cdot \frac{1}{\delta(v)} = \frac{\delta(u)}{2m}.$$

As $\sum_u \delta(u) = 2m$, $\sum_u \pi_u = 1$: π is a distribution, the unique one by irreducibility. Hence $h_{u,u} = 1/\pi_u = \frac{2m}{\delta(u)}$. □

Cover Time of Random Walks

Lemma

For the ***cover time*** *of a graph* $G = (V, E)$ *starting from an arbitrary vertex* $s \in V$*, we have* $\mathrm{cov}(s) < 4m(n-1)$.

Proof.

- For $(u, v) \in E$, $h_{u,v} < 2m$ follows from $h_{u,u} = \frac{2m}{\delta(u)}$, since

$$\frac{2m}{\delta(u)} = h_{u,u} = \frac{1}{\delta(u)} \sum_{w \in N(u)} (h_{w,u} + 1).$$

- There exists a traversal of the vertices of G starting from s that uses $2n-2$ edges (traverse any spanning tree rooted at s twice). With $h_{u,v} < 2m$ for $(u, v) \in E$, it follows

$$\mathrm{cov}(s) < (2n-2) \cdot 2m = 4m(n-1). \qquad \square$$