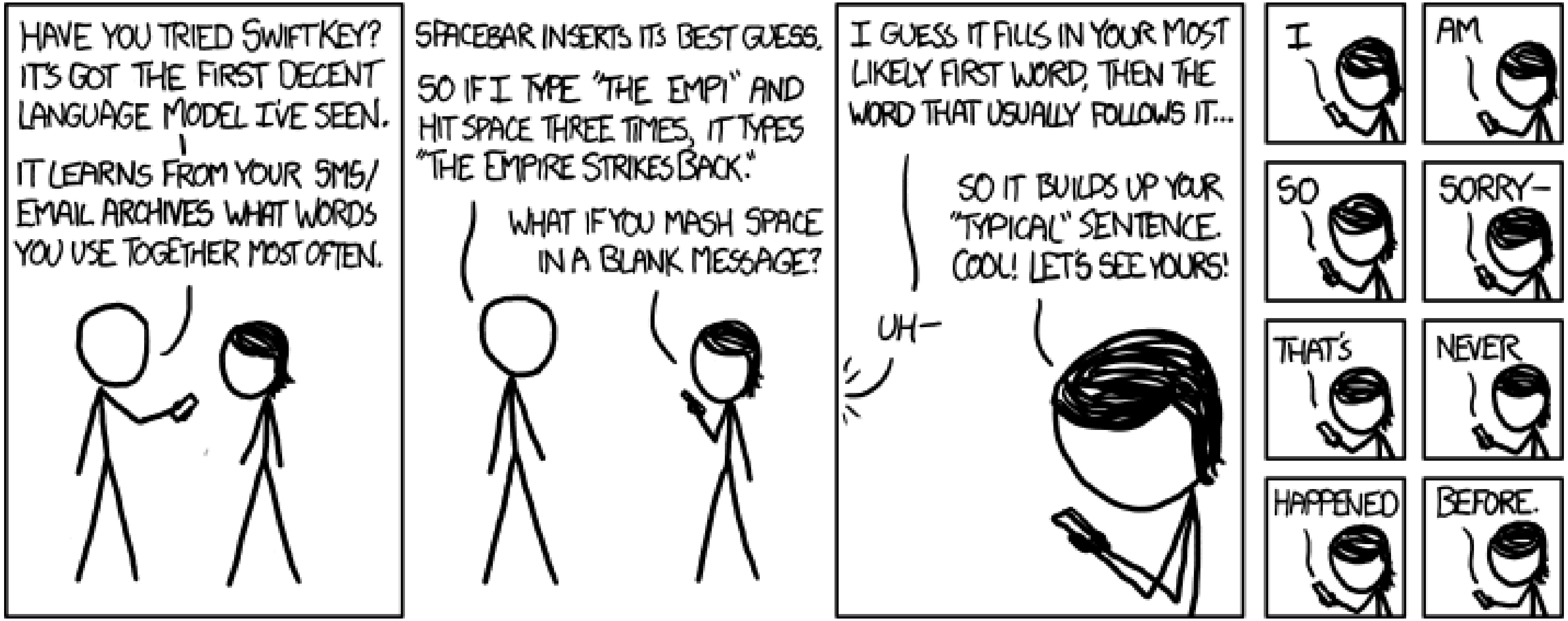



# Markov Chains

CSE 312 26Su

Lecture 23

https://xkcd.com/1068/

Slides adapted from Shreya Jayaraman

# So far: "one-shot" processes

Random Process → **Outcome Distribution** $D$

# So far: "one-shot" processes

Random Process → **Outcome Distribution** $D$

## More generally: randomness can enter over many steps and depend on previous outcomes

Random Process 1 → **Outcome Distribution** $D_1$ → Random Process 2 → **Outcome Distribution** $D_2$ → Random Process 3 → **Outcome Distribution** $D_3$ …

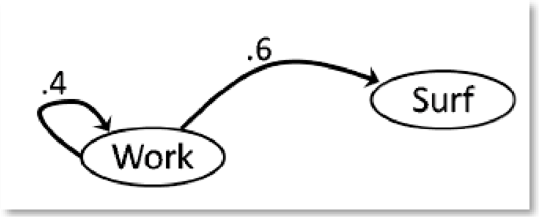
# Stochastic processes

**Definition.** A **discrete-time stochastic process** (DTSP) is a sequence of random variables $X^{(0)}, X^{(1)}, X^{(2)}, \ldots$ where $X^{(t)}$ is the value at time $t$.

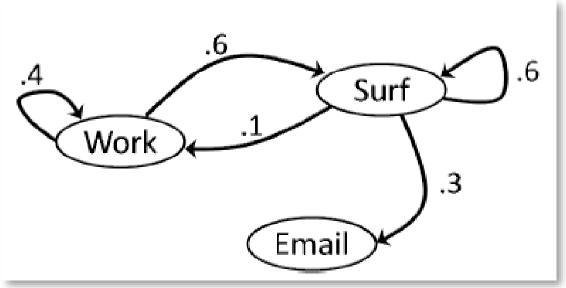
**Today:**
A very special type of DTSP called **Markov Chains**

# A typical day in my life

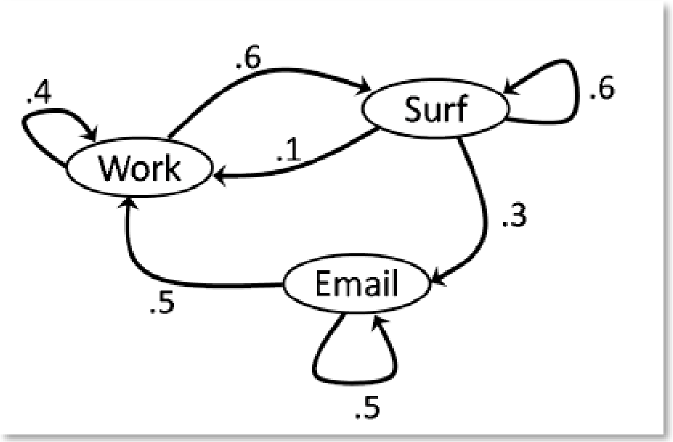

time $t = 0$




time $t = 1$




time $t = 2$

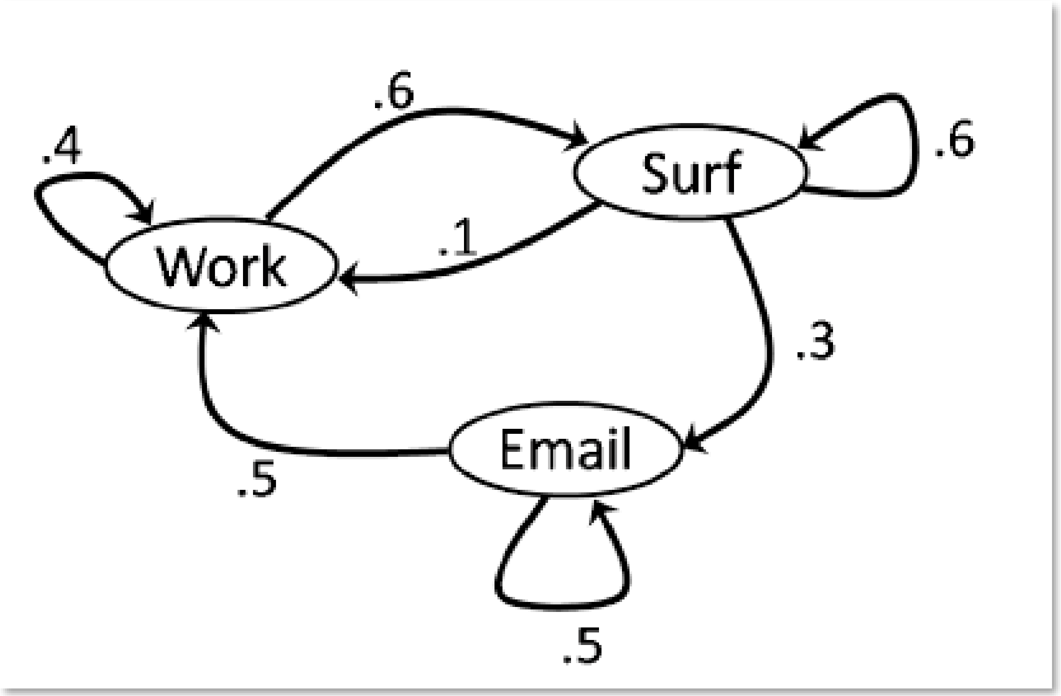



time $t = 3$

# Our first Markov Chain

At *each* time step $t$

- I can be in one of 3 **states** (Work, Surf, Email)
- If I am in some state $s$ at time $t$, the **labels of out-edges** of $s$ **give the probabilities** of moving to each of the states at time $t + 1$ (or staying at the same state)



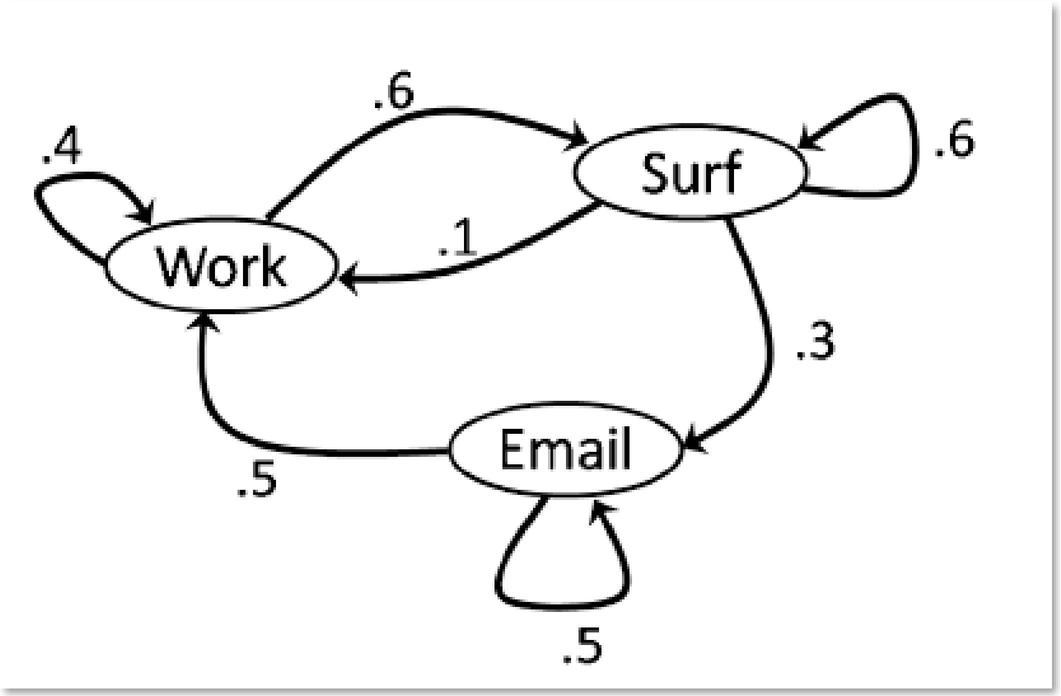
# Hey! That looks familiar to me!

Yes, Markov chains are a special kind of probabilistic (finite) automation.

They look like Deterministic Finite Automata (DFAs) from 311, except…

- No input symbols on the edges (input is "tick of the clock")
- Out edges are labeled with probabilities.
- No start state or accept states.

But like DFAs, the only thing they remember about the past is the state they are currently in.

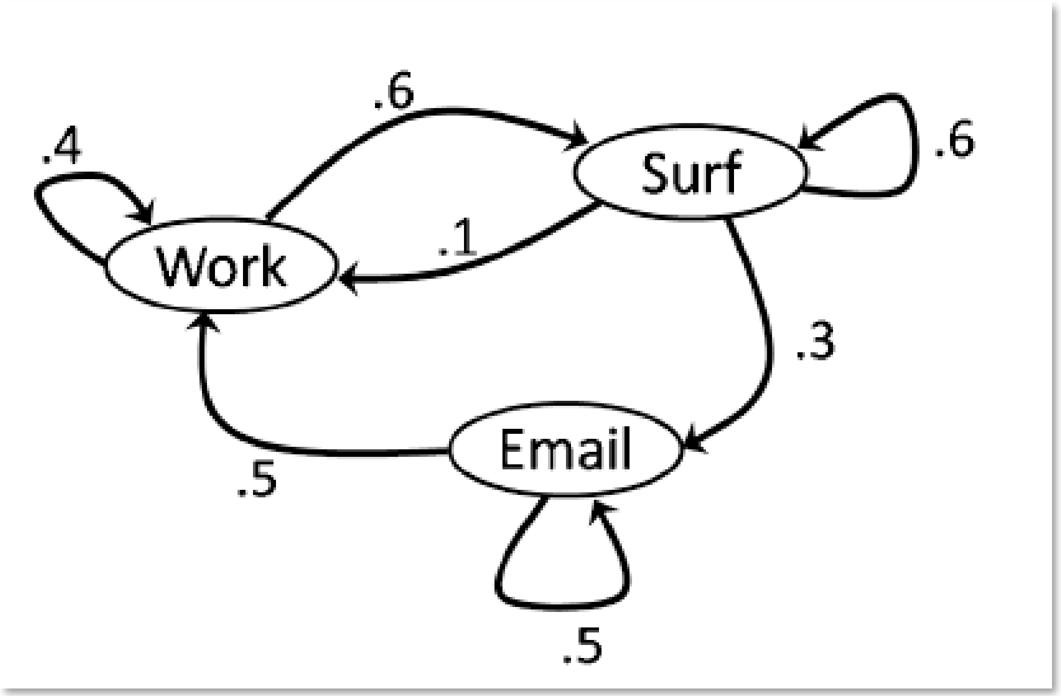


# State Space

State space $S = \{s_1, s_2, \dots, s_n\}$ is finite (or countably infinite)

Define $X^{(t)}$ be the state I am in at time $t$

^ This is a random variable, but it is a function from $\Omega \to S$ (where $S$ is the set of states), not $\Omega \to \mathbb{R}$.



# Markov Property

The future is conditionally independent of the past, given the present.

$$\mathbb{P}(X_{t+1} = x_{t+1} | X_0 = x_0, \dots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t)$$

# Stationary Transition Probabilities

We always transition from state $s_i$ to $s_j$ with probability independent of the current time.

The number of transition probabilities is the number of ways to move from one state to another: $n^2$.

# Transition Probability Matrix (TPM)

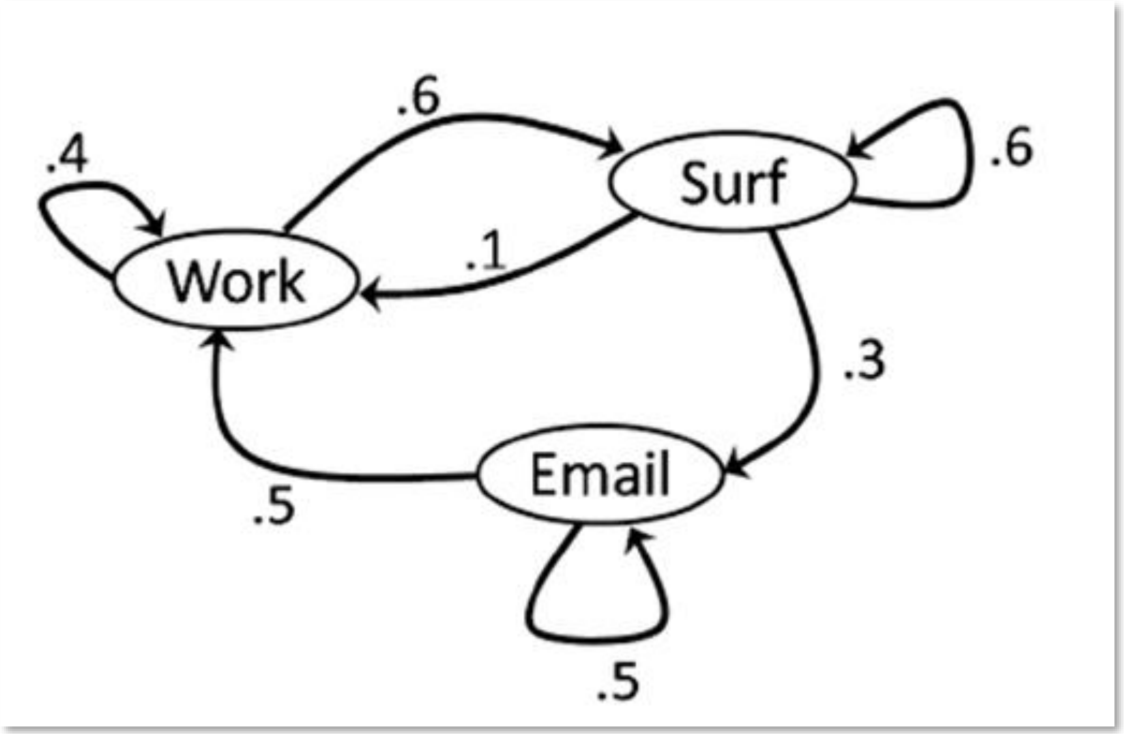
A TPM is a matrix that contains information about the probability of transitioning between states.

Col = next state

Row = current state

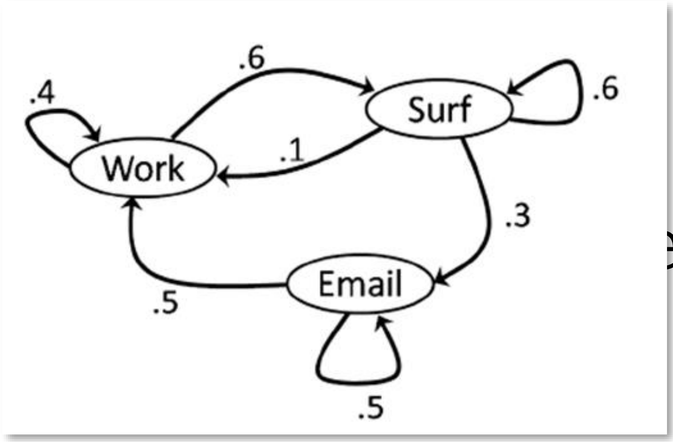
- $M_{i,j} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i)$ (i.e., row $i$ column $j$) is probability of transitioning from $i$ to $j$ (**assuming you are currently at $\boldsymbol{i}$**)
- Note that each row sums to 1 (it's a distribution!)

# Transition Probability Matrix (TPM)




- $M_{i,j} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i)$ (i.e., row $i$ column $j$) is probability of transitioning from $i$ to $j$ (**assuming you are currently at $\boldsymbol{i}$**)
- Note that each row sums to 1 (it's a distribution!)

# Transition Probability Matrix (TPM)

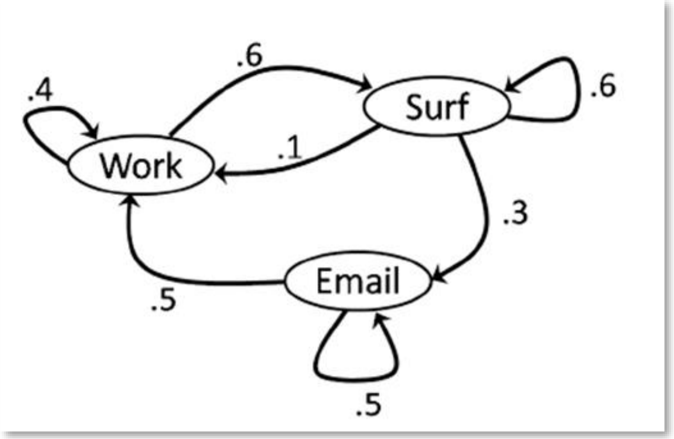



at contains information about the probability of
en states.

$$\begin{array}{cc} & \text{Col = next state} \\ \text{Row = current state} & \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$

- $M_{i,j} = \mathbb{P}(X_{t+1} = s_j | X_t = s_i)$ (i.e., row $i$ column $j$) is probability of transitioning from $i$ to $j$ (**assuming you are currently at $\boldsymbol{i}$**)
- Note that each row sums to 1 (it's a distribution!)

# Extending the time...

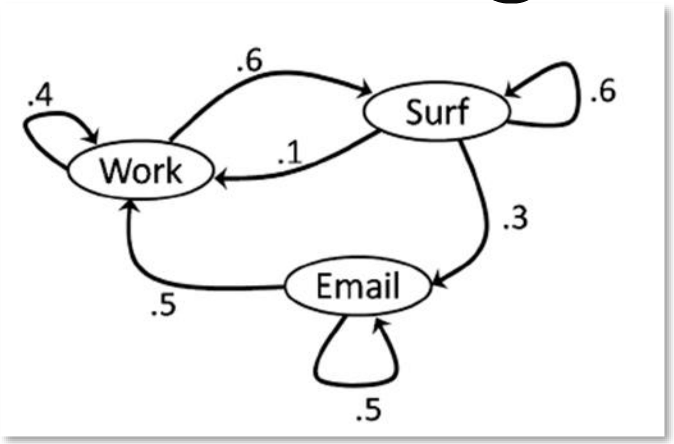



What is the probability that I am in state $s$ at time $1$?

Given: In state Work at time $t = 0$

| State \ Time | $t = 0$ | $t = 1$ | $t = 2$ |
|---|---|---|---|
| $\mathbb{P}(X^{(t)} = \text{Work})$ | 1 | | |
| $\mathbb{P}(X^{(t)} = \text{Surf})$ | 0 | | |
| $\mathbb{P}(X^{(t)} = \text{Email})$ | 0 | | |

# Extending the time...

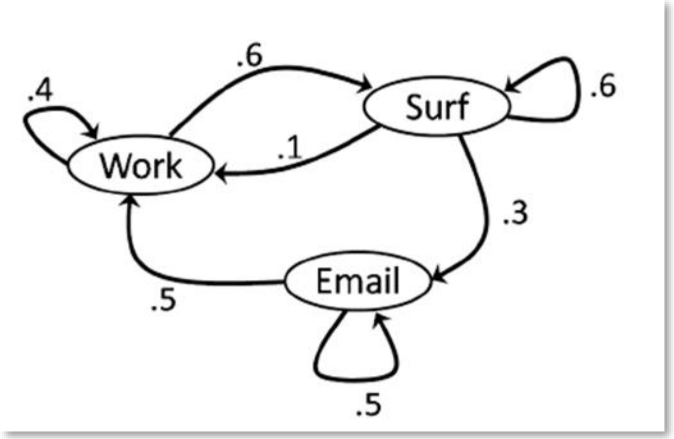



What is the probability that I am in state $s$ at time $1$?

Given: In state Work at time $t = 0$

| State \ Time | $t = 0$ | $t = 1$ | $t = 2$ |
|---|---|---|---|
| $\mathbb{P}(X^{(t)} = \text{Work})$ | 1 | 0.4 | |
| $\mathbb{P}(X^{(t)} = \text{Surf})$ | 0 | 0.6 | |
| $\mathbb{P}(X^{(t)} = \text{Email})$ | 0 | 0 | |

# Extending the time…

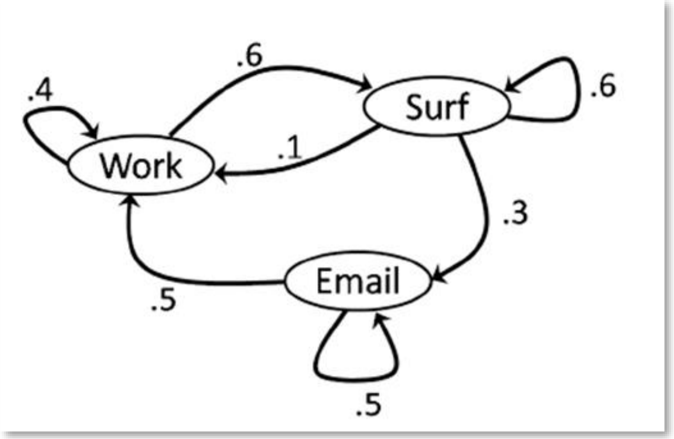



What is the probability that I am in state $s$ at time 1?
What is the probability that I am in state $s$ at time 2?

Given: In state Work at time $t = 0$

| State \ Time | $t = 0$ | $t = 1$ | $t = 2$ |
|---|---|---|---|
| $\mathbb{P}(X^{(t)} = \text{Work})$ | 1 | 0.4 | $= 0.4 \cdot 0.4 + 0.6 \cdot 0.1 = 0.22$ |
| $\mathbb{P}(X^{(t)} = \text{Surf})$ | 0 | 0.6 | $= 0.4 \cdot 0.6 + 0.6 \cdot 0.6 = 0.60$ |
| $\mathbb{P}(X^{(t)} = \text{Email})$ | 0 | 0 | $= 0.4 \cdot 0 + 0.6 \cdot 0.3 = 0.18$ |

# A Linear Algebra Perspective

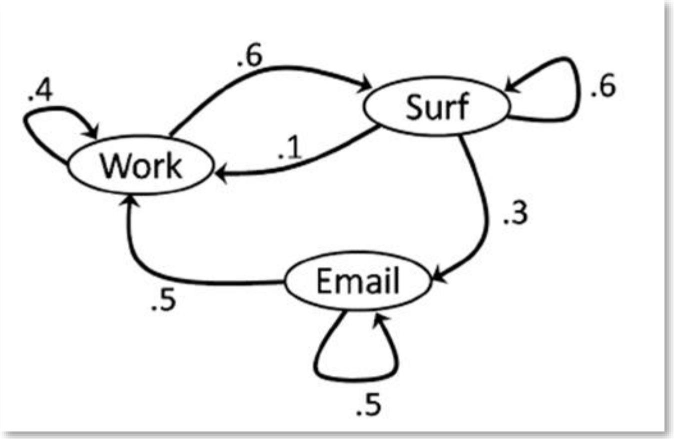



Let $q_s^{(t)}$ be the probability of being in state $s$ at time $t$.
E.g., in the table, $q_E^{(2)} = 0.18$.
Common to think of these as a row vector, e.g.:

$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$$

| State \ Time | $t = 0$ | $t = 1$ | $t = 2$ |
|---|---|---|---|
| $\mathbb{P}(X^{(t)} = \text{Work})$ | 1 | 0.4 | $= 0.4 \cdot 0.4 + 0.6 \cdot 0.1 = 0.22$ |
| $\mathbb{P}(X^{(t)} = \text{Surf})$ | 0 | 0.6 | $= 0.4 \cdot 0.6 + 0.6 \cdot 0.6 = 0.60$ |
| $\mathbb{P}(X^{(t)} = \text{Email})$ | 0 | 0 | $= 0.4 \cdot 0 + 0.6 \cdot 0.3 = 0.18$ |

# A Linear Algebra Perspective

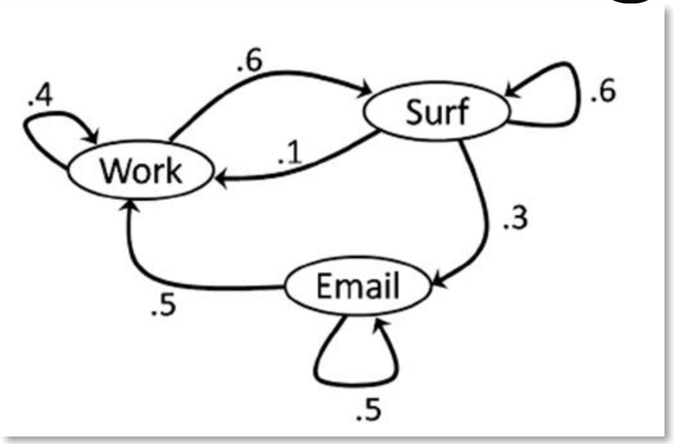



Let $q_s^{(t)}$ be the probability of being in state $s$ at time $t$.
E.g., in the table, $q_E^{(2)} = 0.18$.
Common to think of these as a row vector, e.g.:

$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$$

| **State \ Time** | $t = 0$ | $t = 1$ | $t = 2$ |
|---|---|---|---|
| $\mathbb{P}(X^{(t)} = \textbf{Work}) = q_W^{(t)}$ | 1 | 0.4 | $= 0.4 \cdot 0.4 + 0.6 \cdot 0.1 = 0.22$ |
| $\mathbb{P}(X^{(t)} = \textbf{Surf}) = q_s^{(t)}$ | 0 | 0.6 | $= 0.4 \cdot 0.6 + 0.6 \cdot 0.6 = 0.60$ |
| $\mathbb{P}(X^{(t)} = \textbf{Email}) = q_E^{(t)}$ | 0 | 0 | $= 0.4 \cdot 0 + 0.6 \cdot 0.3 = 0.18$ |

# A Linear Algebra Perspective

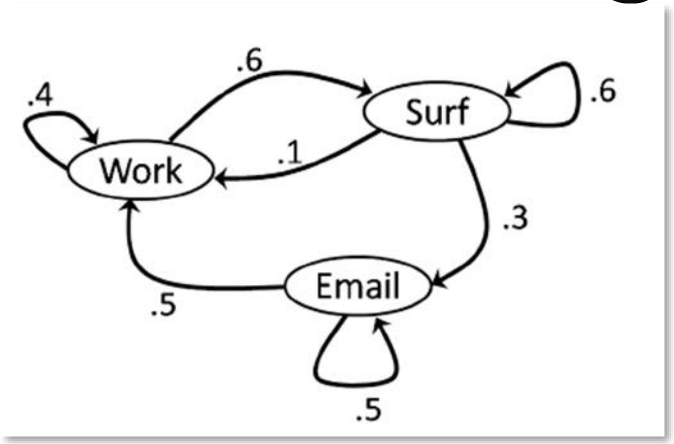



$$M = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Let $[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ be a row vector

| $t$ | 0 | 1 | 2 |
|---|---|---|---|
| $q_W^{(t)}$ | 1 | 0.4 | 0.22 |
| $q_S^{(t)}$ | 0 | 0.6 | 0.60 |
| $q_E^{(t)}$ | 0 | 0 | 0.18 |

# A Linear Algebra Perspective




$$M = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Let $[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ be a row vector

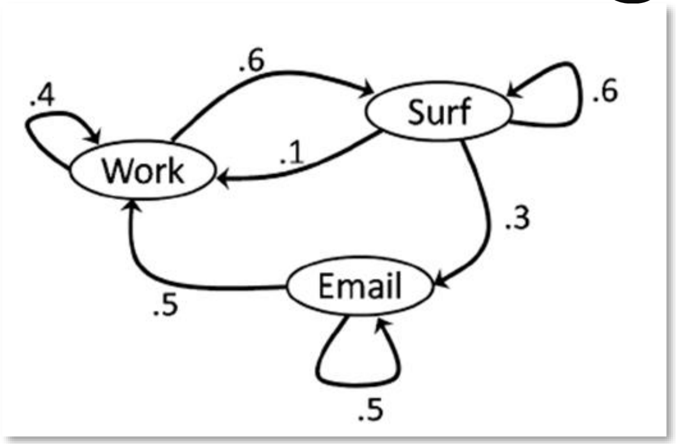
| $t$ | 0 | 1 | 2 |
|---|---|---|---|
| $q_W^{(t)}$ | 1 | 0.4 | 0.22 |
| $q_S^{(t)}$ | 0 | 0.6 | 0.60 |
| $q_E^{(t)}$ | 0 | 0 | 0.18 |

How would you compute the probability of ending up in Work at t=2?

Use the LTP!

$$q_W^{(2)} = q_W^{(1)} \cdot M_{W,W} + q_S^{(1)} \cdot M_{S,W} + q_E^{(1)} \cdot M_{E,W}$$

# A Linear Algebra Perspective




$$M = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Let $[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ be a row vector

| $t$ | 0 | 1 | 2 |
|---|---|---|---|
| $q_W^{(t)}$ | 1 | 0.4 | 0.22 |
| $q_S^{(t)}$ | 0 | 0.6 | 0.60 |
| $q_E^{(t)}$ | 0 | 0 | 0.18 |

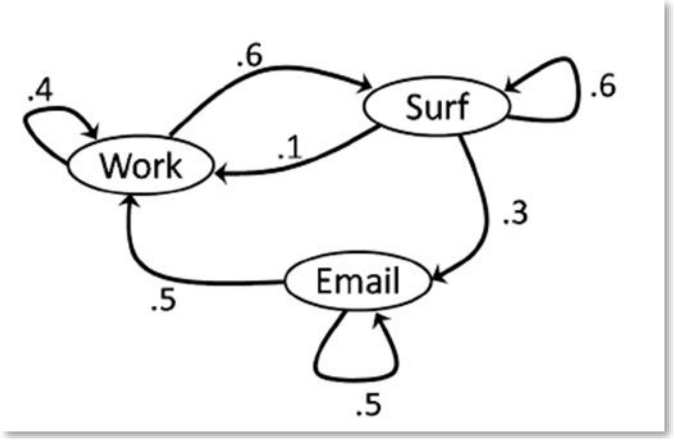
$$[q_W^{(1)}, q_S^{(1)}, q_E^{(1)}]\mathrm{M}$$

$$= [0.4 \quad 0.6 \quad 0] \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$$= [0.22 \quad 0.60 \quad 0.18]$$

$$= [q_W^{(2)} \quad q_S^{(2)} \quad q_E^{(1)}]$$

# A Linear Algebra Perspective




$$M = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Let $[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ be a row vector

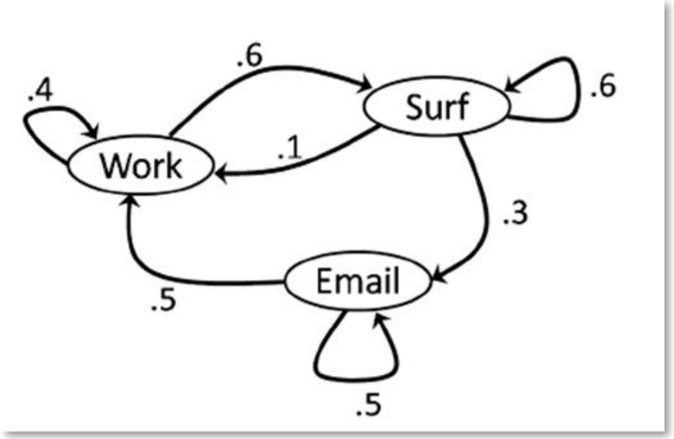
$$[q_W^{(1)}, q_S^{(1)}, q_E^{(1)}]\mathrm{M}$$

$$= \begin{bmatrix} 0.4 & 0.6 & 0 \end{bmatrix} \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

| State \ Time | 0 | 1 | 2 |
|---|---|---|---|
| $\mathbb{P}(\boldsymbol{X}^{(t)} = \textbf{Work}) = q_W^{(t)}$ | 1 | 0.4 | 0.22 |
| $\mathbb{P}(\boldsymbol{X}^{(t)} = \textbf{Surf}) = q_S^{(t)}$ | 0 | 0.6 | 0.60 |
| $\mathbb{P}(\boldsymbol{X}^{(t)} = \textbf{Email}) = q_E^{(t)}$ | 0 | 0 | 0.18 |

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} ad + bg & ae + bh & af + bi \end{bmatrix}$$

# General Vector-Matrix Multiplication




That Vector-Matrix multiplication representation works for any step.

$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \overset{\boldsymbol{M}}{\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}} = [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]$$

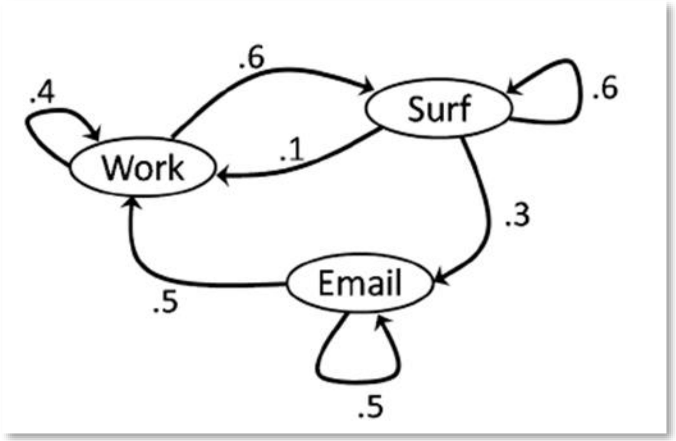
Vector-matrix multiplication

$$q_W^{(t)} \cdot 0.4 + q_S^{(t)} \cdot 0.1 + q_E^{(t)} \cdot 0.5 = q_W^{(t+1)}$$

$$q_W^{(t)} \cdot 0.6 + q_S^{(t)} \cdot 0.6 + q_E^{(t)} \cdot 0 = q_S^{(t+1)}$$

$$q_W^{(t)} \cdot 0 + q_S^{(t)} \cdot 0.3 + q_E^{(t)} \cdot 0.5 = q_E^{(t+1)}$$

# A little bit of induction




$$
[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \;
\begin{array}{c}
\boldsymbol{M} \\
\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}
\end{array}
= [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]
$$

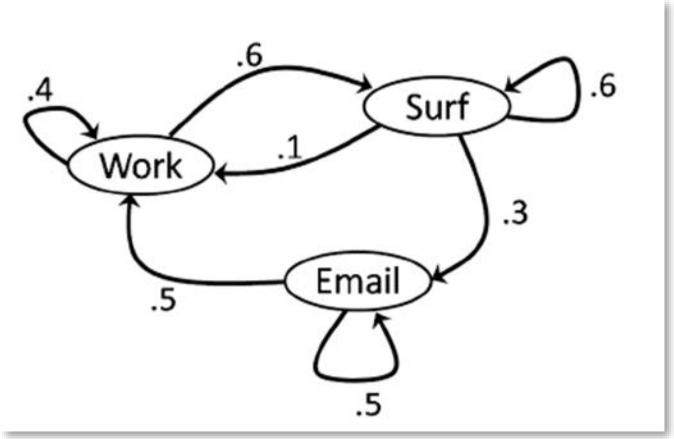
Let $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ Then for all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

So $\boldsymbol{q}^{(1)} = \boldsymbol{q}^{(0)}\boldsymbol{M}$

$\boldsymbol{q}^{(2)} = \boldsymbol{q}^{(1)}\boldsymbol{M} = (\boldsymbol{q}^{(0)}\boldsymbol{M})\boldsymbol{M} = \boldsymbol{q}^{(0)}\boldsymbol{M}^2$

…

# A little bit of induction




$$[q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \overset{\boldsymbol{M}}{\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}} = [q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}]$$

Let $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$ Then for all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

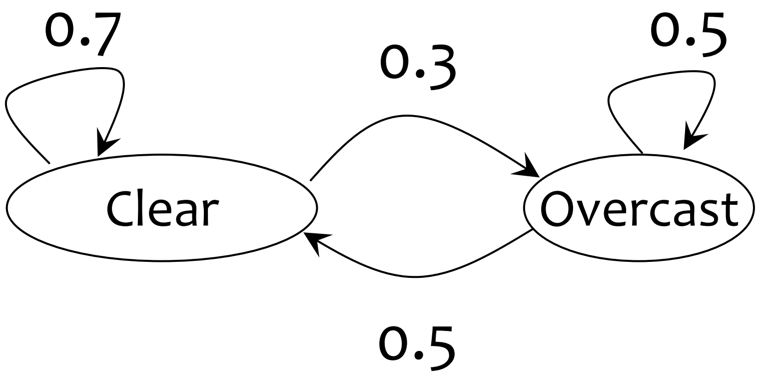
So $\boldsymbol{q}^{(1)} = \boldsymbol{q}^{(0)}\boldsymbol{M}$

$\boldsymbol{q}^{(2)} = \boldsymbol{q}^{(1)}\boldsymbol{M} = (\boldsymbol{q}^{(0)}\boldsymbol{M})\boldsymbol{M} = \boldsymbol{q}^{(0)}\boldsymbol{M}^2$

…

$\boxed{\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t \text{ for all } t \geq 0}$

# Another example:




Suppose that $\boldsymbol{q}^{(0)} = \left[q_C^{(0)}, q_O^{(0)}\right] = [0,1]$

We have $\boldsymbol{M} = \begin{bmatrix} 0.7 & 0.3 \\ 0.5 & 0.5 \end{bmatrix}$

What is $\boldsymbol{q}^{(2)}$ ?

a. $[0.3, 0.7]$

b. $[0.6, 0.4]$

c. $[0.7, 0.3]$

d. $[0.5, 0.5]$

e. $[0.4, 0.6]$

# Markov Chain Questions

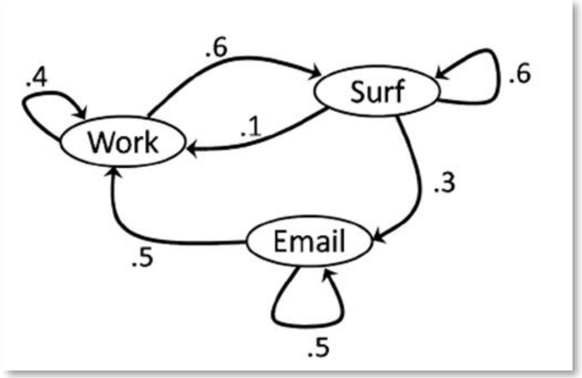
1. What is the probability that I am in state $s$ at time $1$?
2. What is the probability that I am in state $s$ at time $2$?
3. What is the probability that I am in state $s$ at some time $t$ far in the future?

$$\boxed{\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t \text{ for all } t \geq 0}$$

What does $\boldsymbol{M}^t$ look like for really big $t$ ?

# $\boldsymbol{M}^t$ as $t$ grows

$\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t$ for all $t \geq 0$




$\boldsymbol{M}$

$$\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$\boldsymbol{M}^2$

$$\begin{array}{c c} & \begin{array}{ccc} W & S & E \end{array} \\ \begin{array}{c} W \\ S \\ E \end{array} & \begin{pmatrix} .22 & .6 & .18 \\ .25 & .42 & .33 \\ .45 & .3 & .25 \end{pmatrix} \end{array}$$

$\boldsymbol{M}^3$

$$\begin{array}{c c} & \begin{array}{ccc} W & S & E \end{array} \\ \begin{array}{c} W \\ S \\ E \end{array} & \begin{pmatrix} .238 & .492 & .270 \\ .307 & .402 & .291 \\ .335 & .450 & .215 \end{pmatrix} \end{array}$$

$\boldsymbol{M}^{10}$

$$\begin{array}{c c} & \begin{array}{ccc} W & S & E \end{array} \\ \begin{array}{c} W \\ S \\ E \end{array} & \begin{pmatrix} .2940 & .4413 & .2648 \\ .2942 & .4411 & .2648 \\ .2942 & .4413 & .2648 \end{pmatrix} \end{array}$$

$\boldsymbol{M}^{30}$

$$\begin{array}{c c} & \begin{array}{ccc} W & S & E \end{array} \\ \begin{array}{c} W \\ S \\ E \end{array} & \begin{pmatrix} .29411764705 & .44117647059 & .26470588235 \\ .29411764706 & .44117647058 & .26470588235 \\ .29411764706 & .44117647059 & .26470588235 \end{pmatrix} \end{array}$$

$\boldsymbol{M}^{60}$

$$\begin{array}{c c} & \begin{array}{ccc} W & S & E \end{array} \\ \begin{array}{c} W \\ S \\ E \end{array} & \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

What does this say about $\boldsymbol{q}^{(t)}$?

# What does this say about $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\boldsymbol{M}^t$ ?

Note that no matter what probability distribution $\boldsymbol{q}^{(0)}$ is …

$\boldsymbol{q}^{(\mathbf{0})}\boldsymbol{M}^t$ is just a <u>weighted average</u> of the rows of $\boldsymbol{M}^t$

If every row of $\boldsymbol{M}^t$ were *exactly* the same …that would mean that $\boldsymbol{q}^{(\mathbf{0})}\boldsymbol{M}^t$ would be equal to the common row

So $\boldsymbol{q}^{(t)}$ wouldn't depend on $\boldsymbol{q}^{(\mathbf{0})}$

The rows aren't exactly the same, but they are very close

So $\boldsymbol{q}^{(t)}$ barely depends on $\boldsymbol{q}^{(\mathbf{0})}$ after very few steps

# Observation

If $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}$ then it will never change again!

Called a **stationary distribution** and has a special symbol

$$\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$$

Solution to $\boldsymbol{\pi} = \boldsymbol{\pi}\,\boldsymbol{M}$

Typically happens after a “long time” (mixing time)

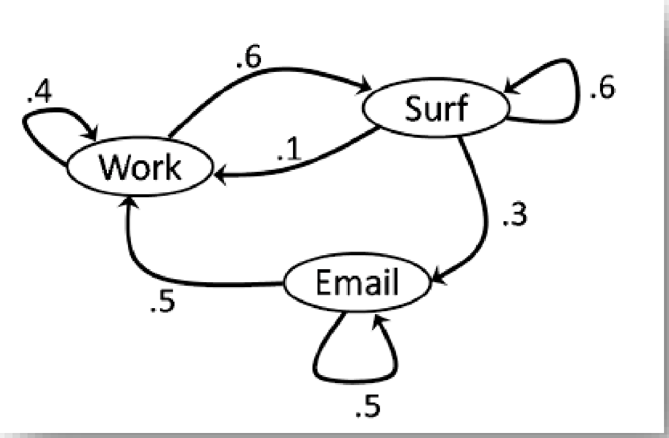
# Solving for Stationary Distribution

$$\boldsymbol{M} = \begin{pmatrix} .4 & .6 & 0 \\ .1 & .6 & .3 \\ .5 & 0 & .5 \end{pmatrix}$$

*Stationary Distribution satisfies*

- $\boldsymbol{\pi} = \boldsymbol{\pi M}$, where $\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$
- $\pi_W + \pi_S + \pi_E = 1$

$\Rightarrow \pi_W = \frac{10}{34}, \ \pi_S = \frac{15}{34}, \ \pi_E = \frac{9}{34}$




As $t \to \infty$, $\boldsymbol{q}^{(t)} \to \boldsymbol{\pi}$ no matter what distribution $\boldsymbol{q}^{(0)}$ is !!

# Markov Chains in general

A set of $n$ **states** $S = \{s_1, s_2, \ldots, s_n\}$

The state at time $t$ is denoted by $X^{(t)}$

A $n \times n$ **transition matrix** $\boldsymbol{M}$, with $\boldsymbol{M}_{ij} = P\left(X^{(t+1)} = j \;\middle|\; X^{(t)} = i\right)$

$\boldsymbol{q}^{(t)} = (q_1^{(t)}, q_2^{(t)}, \ldots, q_n^{(t)})$ where $q_i^{(t)} = P(X^{(t)} = i)$

Transition: LTP $\Rightarrow \boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\,\boldsymbol{M}$ so $\boldsymbol{q}^{(t)} = \boldsymbol{q}^{(0)}\,\boldsymbol{M}^t$

A **stationary distribution** $\boldsymbol{\pi}$ is the solution to:

$$\boldsymbol{\pi} = \boldsymbol{\pi}\,\boldsymbol{M}, \text{ normalized so that } \Sigma_{i\in[n]}\pi_i = 1$$

# The Fundamental Theorem of Markov Chains

**The Fundamental Theorem of Markov Chains**

Every Markov chain that is irreducible and aperiodic has a unique stationary distribution $\pi$.

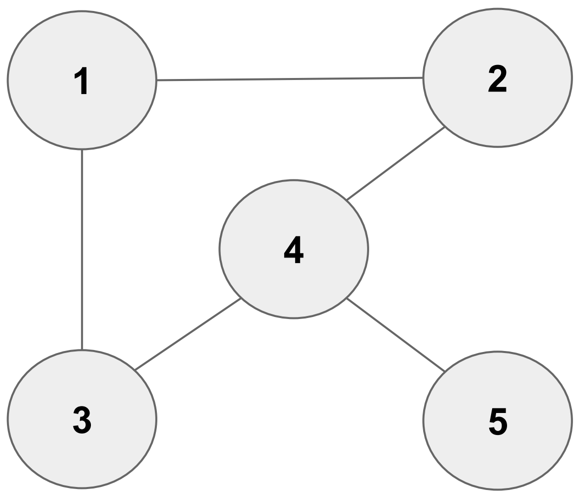
Moreover, as $t \to \infty$: for all $i, j$: $\left(M_{ij}\right)^t = \pi_j$

Details of "irreducible" and "aperiodic" are beyond our scope; but suffice it to say that practically relevant Markov chains usually meet both these conditions (or can be rearranged to meet them).

# Example: Random Walks

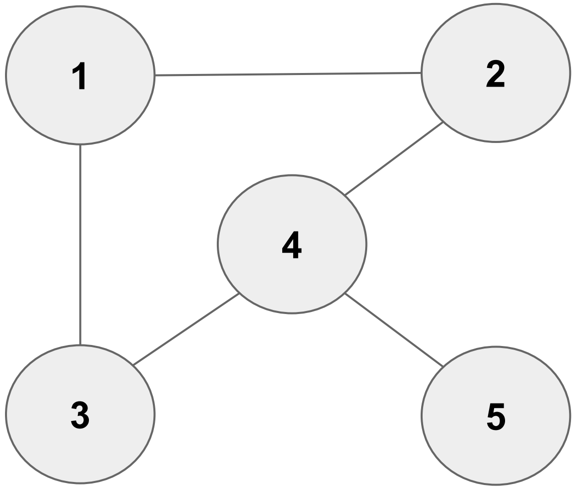
Suppose we start at node 1, and at each step transition to a neighboring node with equal probability.

This is called a “random walk” on this graph.



# Example: Random Walks on an Undirected Graph

Start by defining transition probs.



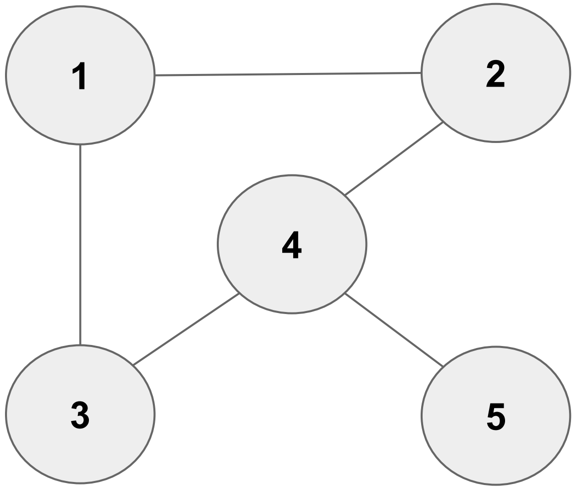

$$
\begin{array}{c|ccccc|}
 & \text{To } \mathbf{1} & \text{To } \mathbf{2} & \text{To } \mathbf{3} & \text{To } \mathbf{4} & \text{To } \mathbf{5} \\
\hline
\text{From } \mathbf{1} & & & & & \\
\text{From } \mathbf{2} & & & & & \\
\text{From } \mathbf{3} & & & & & \\
\text{From } \mathbf{4} & & & & & \\
\text{From } \mathbf{5} & & & & & \\
\end{array}
$$

$$\boldsymbol{M}_{ij} = P\left(X^{(t+1)} = j \mid X^{(t)} = i\right)$$

$$\boldsymbol{q}_i^{(t)} = P\left(\mathrm{X}^{(t)} = i\right) = \left(\boldsymbol{q}^{(0)} \boldsymbol{M}^t\right)_i$$

# Example: Random Walks on an Undirected Graph

Start by defining transition probs.




$$
\begin{array}{cc}
 & \begin{array}{ccccc} \textbf{To 1} & \textbf{To 2} & \textbf{To 3} & \textbf{To 4} & \textbf{To 5} \end{array} \\
\begin{array}{c} \text{From } \mathbf{1} \\ \text{From } \mathbf{2} \\ \text{From } \mathbf{3} \\ \text{From } \mathbf{4} \\ \text{From } \mathbf{5} \end{array} &
\begin{bmatrix}
0 & 1/2 & 1/2 & 0 & 0 \\
1/2 & 0 & 0 & 1/2 & 0 \\
1/2 & 0 & 0 & 1/2 & 0 \\
0 & 1/3 & 1/3 & 0 & 1/3 \\
0 & 0 & 0 & 1 & 0
\end{bmatrix}
\end{array}
$$

$$\boldsymbol{M}_{ij} = P\left(X^{(t+1)} = j \mid X^{(t)} = i\right)$$

$$\boldsymbol{q}_i^{(t)} = P\left(X^{(t)} = i\right) = \left(\boldsymbol{q}^{(0)} \boldsymbol{M}^t\right)_i$$

# Markov Chain Monte Carlo (MCMC)

# Markov Chain Monte Carlo

A modelling technique.

Two common uses:

> Sampling from a (hard) distribution (Bayesian analysis!)

> Approximate solutions to complex optimization problems

# Markov Chain Monte Carlo

A modelling technique.

Two common uses:

> Sampling from a (hard) distribution (Bayesian analysis!)

> Approximate solutions to complex optimization problems

Simulate a random walk that moves among possible configurations of a system and will converge to a useful distribution over these configurations

# The 0-1 knapsack problem

Given $n$ items with weights $w_1, \dots, w_n > 0$ and values $v_1, \dots, v_n > 0$

- Knapsack with weight limit $W$

Find the most valuable subset of items which satisfy the weight constraint of the knapsack.

# The 0-1 knapsack problem

Given $n$ items with weights $w_1, \ldots, w_n > 0$ and values $v_1, \ldots, v_n > 0$

- Knapsack with weight limit $W$

Find the most valuable subset of items which satisfy the weight constraint of the knapsack.

Item 1: 2lb, $5

Item 2: 4lb, $9

Item 3: 2lb, $10

Limit $W$ = 4lb

Then best is choosing items 1 and 3

# MCMC formalization

Let $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ represent whether we took item $i$ or not

We want to maximize total value: $\sum_{i=1}^{n} v_i x_i$

With weight constraint: $\sum_{i=1}^{n} w_i x_i \leq W$

# MCMC formalization

Let $x = (x_1, \dots, x_n) \in \{0, 1\}^n$ represent whether we took item $i$ or not

We want to maximize total value: $\sum_{i=1}^{n} v_i x_i$

With weight constraint: $\sum_{i=1}^{n} w_i x_i \leq W$

Define a Markov chain.

The states are the possible solutions.

How many?

# MCMC formalization

Let $x = (x_1, \dots, x_n) \in \{0,1\}^n$ represent whether we took item $i$ or not

We want to maximize total value: $\sum_{i=1}^{n} v_i x_i$

With weight constraint: $\sum_{i=1}^{n} w_i x_i \le W$

Define a Markov chain.

The states are the possible solutions.

How many? $2^n$

# MCMC formalization

Let $x = (x_1, \dots, x_n) \in \{0,1\}^n$ represent whether we took item $i$ or not

We want to maximize total value: $\sum_{i=1}^{n} v_i x_i$

With weight constraint: $\sum_{i=1}^{n} w_i x_i \leq W$

Define a Markov chain.

The states are the possible solutions.

How many? $2^n$

Implicit TPM. Higher probability of a "good" solution

One "good" definition: solutions that satisfy the weight constraint.

# Knapsack: Starter Algorithm

**Algorithm 3** (First Attempt) MCMC for 0-1 Knapsack Problem

x $\leftarrow$ vector of $n$ zeros, where $x_i$ is a binary vector in $\{0, 1\}^n$ which represents whether or not we have item $i$. (Initially, start with an empty knapsack).
best_x $\leftarrow$ x
**for** $t = 1, \ldots,$ NUM_ITER **do**
    $k \leftarrow$ a random integer in $\{1, 2, \ldots, n\}$.
    new_x $\leftarrow$ x but with x$[k]$ flipped ($0 \to 1$ or $1 \to 0$).
    **if** new_x satisfies weight constraint **then**
        x $\leftarrow$ new_x
    **if** value(x) $>$ value(best_x) **then**
        best_x $\leftarrow$ x

# MCMC Strategy (for optimization)

Define a Markov chain:

> States = possible solutions

> Implicitly defined TPM (higher probabilities on “good” solutions)

Run MCMC: Simulate Markov Chain for many iterations until we reach a “good” state (transitioning according to the TPM until stationary distribution).

# A Better Algorithm

**Algorithm 5** MCMC for 0-1 Knapsack Problem

subset $\leftarrow$ vector of $n$ zeros, where subset is always a binary vector in $\{0,1\}^n$ which represents whether or not we have each item. (Initially, start with an empty knapsack).
best_subset $\leftarrow$ vector of $n$ zeros
**for** $t = 1, \ldots,$ NUM_ITER **do**
    $k \leftarrow$ a random uniform integer in $\{1, 2, \ldots, n\}$.
    new_subset $\leftarrow$ subset but with subset$[k]$ flipped $(0 \to 1$ or $1 \to 0)$.
    $\Delta \leftarrow$ value(new_subset) $-$ value(subset)
    **if** new_subset satisfies weight constraint (total weight $\leq W$) **then**
        **if** $\Delta > 0$ OR $(T > 0$ AND $\mathrm{Unif}(0,1) < e^{\Delta/T})$ **then**
            subset $\leftarrow$ new_subset
    **if** value(subset) $>$ value(best_subset) **then**
        best_subset $\leftarrow$ subset

# A Better Algorithm

**Algorithm 5** MCMC for 0-1 Knapsack Problem

subset $\leftarrow$ vector of $n$ zeros, where subset is always a
or not we have each item. (Initially, start with an er
best_subset $\leftarrow$ vector of $n$ zeros
**for** $t = 1, \ldots,$ NUM_ITER **do**
  $k \leftarrow$ a random uniform integer in $\{1, 2, \ldots, n\}$.
  new_subset $\leftarrow$ subset but with subset$[k]$ flipped ($0 \rightarrow 1$ or $1 \rightarrow 0$).
  $\Delta \leftarrow$ value(new_subset) $-$ value(subset)
  **if** new_subset satisfies weight constraint (total weight $\leq W$) **then**
    **if** $\Delta > 0$ OR ($T > 0$ AND Unif$(0, 1) < e^{\Delta/T}$) **then**
      subset $\leftarrow$ new_subset
  **if** value(subset) $>$ value(best_subset) **then**
    best_subset $\leftarrow$ subset

Check multiple "better" features:
- Does it satisfy the weight constraint?
- Is it a solution of higher value?

# A Better Algorithm

**Algorithm 5** MCMC for 0-1 Knapsack Problem

subset $\leftarrow$ vector of $n$ zeros, where subset is always a
or not we have each item. (Initially, start with an er
best_subset $\leftarrow$ vector of $n$ zeros
**for** $t = 1, \ldots,$ NUM_ITER **do**
  $k \leftarrow$ a random uniform integer in $\{1, 2, \ldots, n\}$.
  new_subset $\leftarrow$ subset but with subset$[k]$ flipped ($0 \to 1$ or $1 \to 0$).
  $\Delta \leftarrow$ value(new_subset) $-$ value(subset)
  **if** new_subset satisfies weight constraint (total weight $\leq W$) **then**
    **if** $\Delta > 0$ OR ($T > 0$ AND Unif$(0,1) < e^{\Delta/T}$) **then**
      subset $\leftarrow$ new_subset
  **if** value(subset) $>$ value(best_subset) **then**
    best_subset $\leftarrow$ subset

Exploitation: Transition to higher total values

Exploration: random transitions occasionally, to avoid local optima

# PageRank

# PageRank: Some History

The year was 1997

- Bill Clinton in the White House
- Deep Blue beat world chess champion (Kasparov)

The Internet was not like it was today. Finding stuff was hard!

- In Nov 1997, only one of the top 4 search engines found itself when you searched for it.

# The Problem

Search engines worked by matching words in your queries to documents.

Not bad in theory, but in practice there are lots of documents that match a query.

- Search for ‘Bill Clinton’, top result is ‘Bill Clinton Joke of the Day’
- Susceptible to spammers and advertisers

# The Fix: Ranking Results

Start by doing filtering to relevant documents (with decent textual match).

Then **rank** the results based on some measure of 'quality' or 'authority'.

**Key question:** How to define 'quality' or 'authority'?

Enter two groups:

Jon Kleinberg (professor at Cornell)

Larry Page and Sergey Brin (Ph.D. students at Stanford)

# Both groups had the same brilliant idea

Larry Page and Sergey Brin (Ph.D. students at Stanford)

- Took the idea and founded Google, making billions

Jon Kleinberg (professor at Cornell)

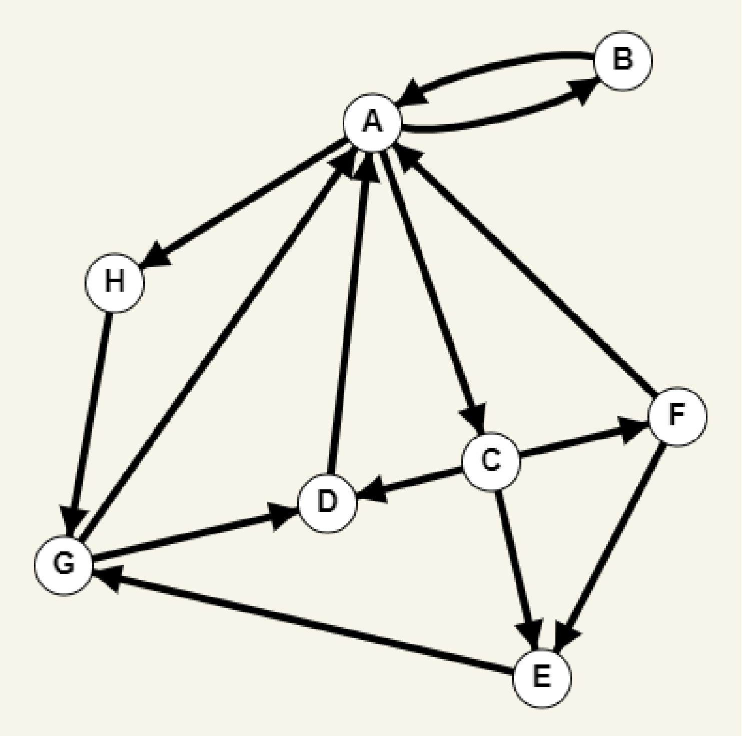
- MacArthur genius prize, Nevanlinna Prize and many other academic honors

# PageRank - Idea

Take into account the directed graph structure of the web.

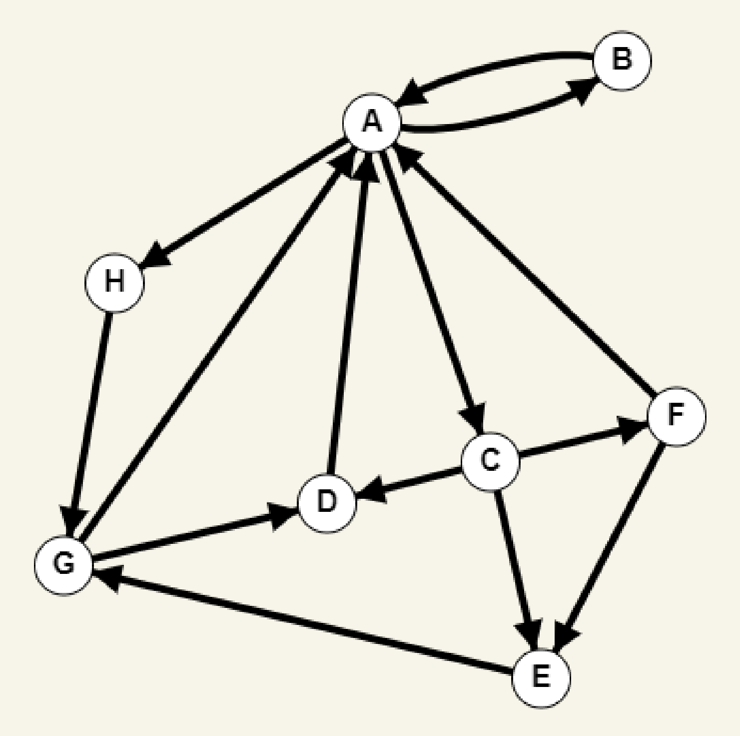
Use hyperlink analysis to compute what pages are high quality or have high authority.

Trust the Internet itself define what is useful via its links.



# PageRank – Idea (try 1)

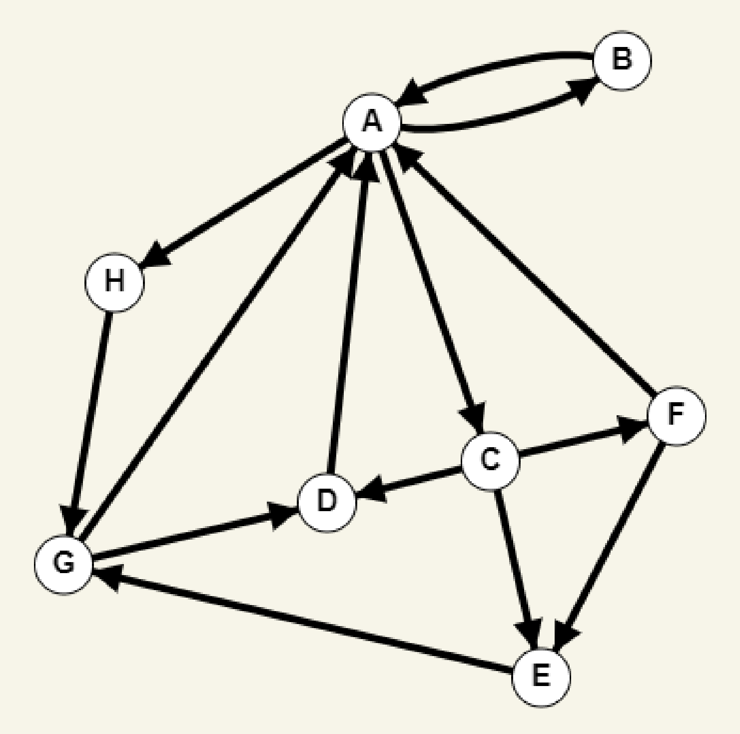
Idea 1 : Think of each link as a citation "vote of quality"




Rank pages by in-degree?

# PageRank – Idea 1 doesn't work

Idea 1 : Think of each link as a citation
"vote of quality"




Rank pages by in-degree?

Problems:

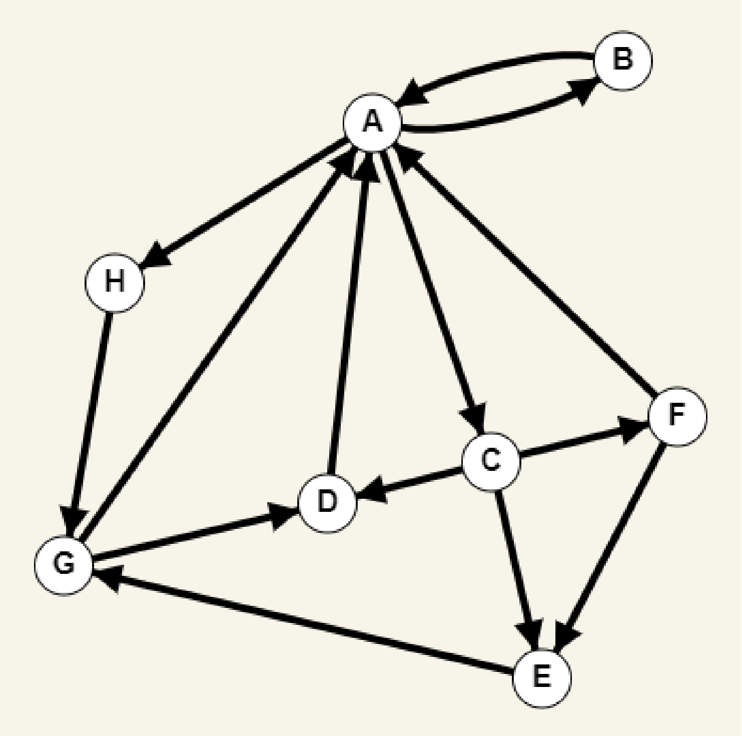
Spamming

Some linkers are not discriminating

Not all links created equal

# PageRank – Idea (try 2)

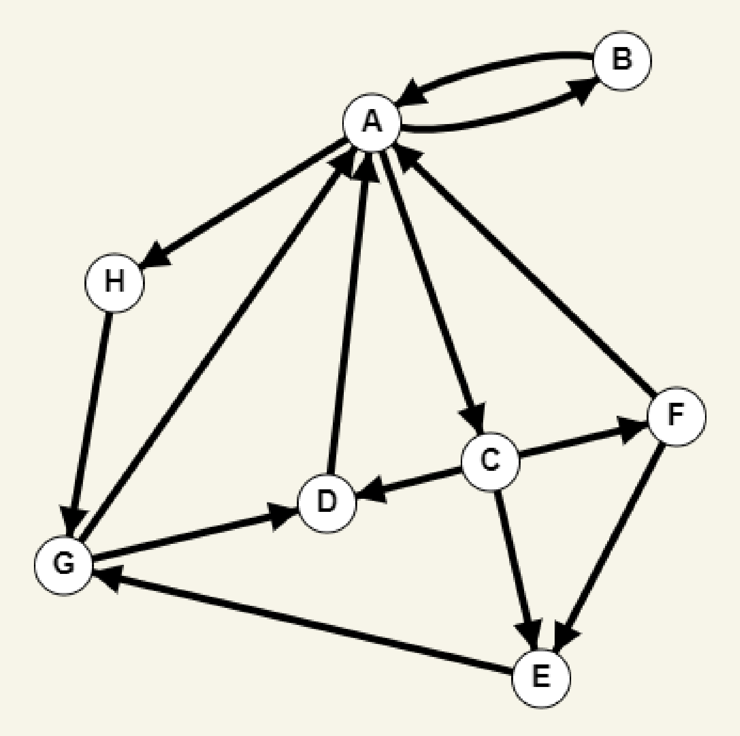
Idea 2 : Perhaps we should weight the links somehow and then use the weights of the in-links to rank pages



# Inching towards PageRank

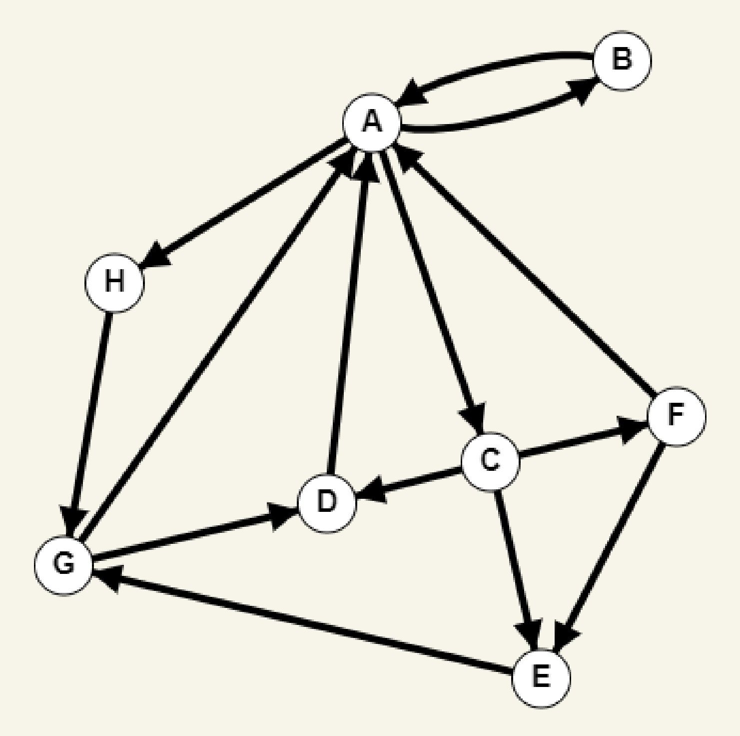
Web page has high quality if it's linked to by lots of high quality pages

That's a recursive definition!



# Inching towards PageRank (importance)

If web page $x$ has $d$ outgoing links, one of which goes to $y$, this contributes $1/d$ to the importance of $y$




But $1/d$ of what?

We want to take into account the importance of $x$ too…

…so it actually contributes $1/d$ of the importance of $x$

# This gives the following equations

Idea: Use the transition matrix $\boldsymbol{M}$ defined by a *random walk* on the web to compute quality of webpages.

Namely: Find $\boldsymbol{q}$ such that $\boldsymbol{qM} = \boldsymbol{q}$

This is the stationary distribution for the Markov chain defined by a random web surfer

- Starts at some node (webpage) and randomly follows a link to another.
- Use stationary distribution of her surfing patterns after a long time as notion of quality

# Issues with PageRank

How to handle dangling nodes (dead ends that don’t link to anything) ?

How to handle Rank sinks – group of pages that only link to each other ?

Both solutions can be solved by “teleportation”

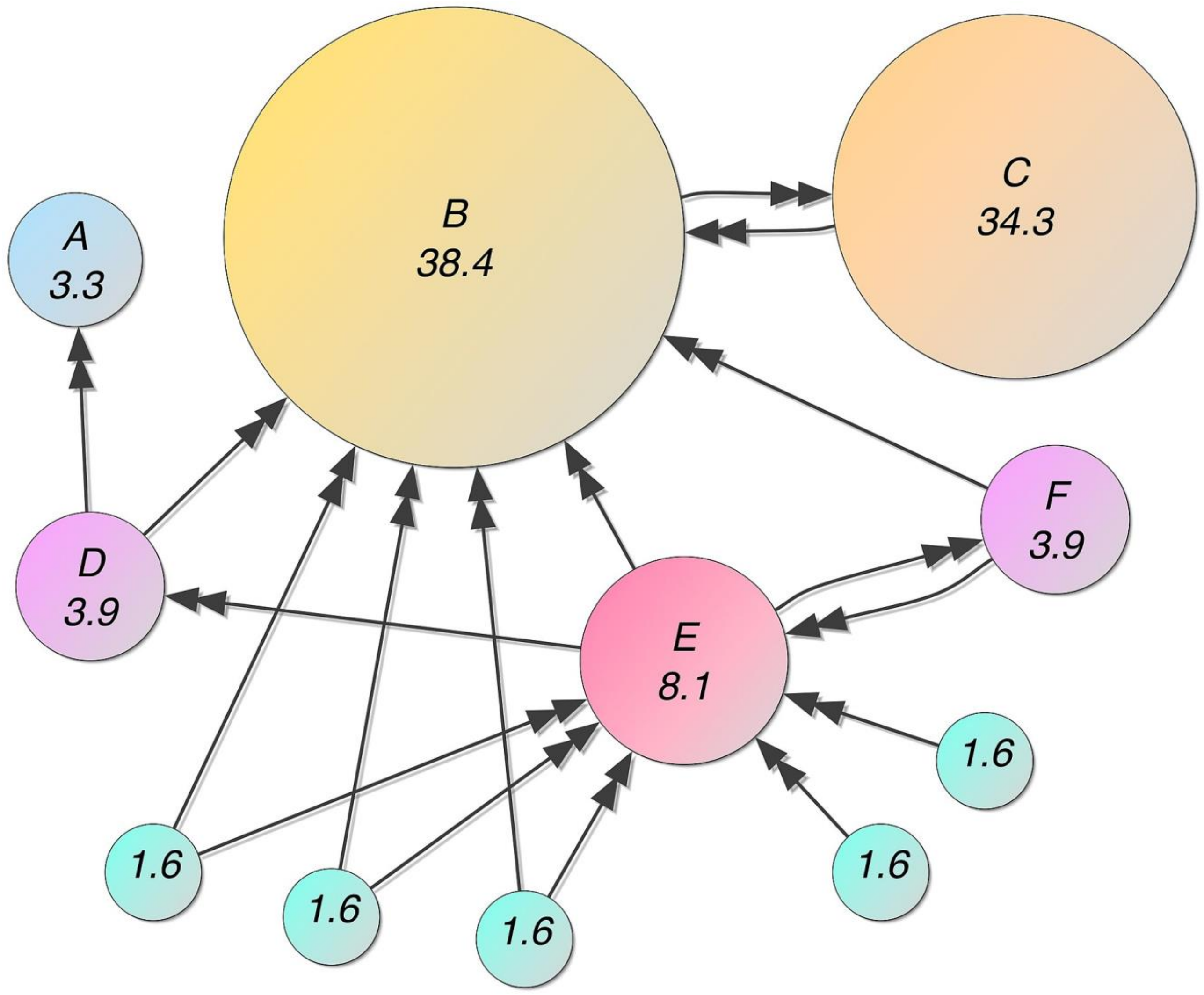
# Final PageRank Algorithm

1. Make a Markov Chain with one state for each webpage on the Internet with the transition probabilities $\boldsymbol{M}_{ij} = \frac{1}{\mathrm{outdeg}(i)}$.
2. Use a modified random walk. At each point in time if the surfer is at some webpage $i$:

   If $i$ has outlinks:
   - With probability $p$, take a step to one of the neighbors of $i$ (equally likely)
   - With probability $1 - p$, "teleport" to a uniformly random page in the whole Internet.

   Otherwise, always "teleport"
3. Compute stationary distribution $\boldsymbol{\pi}$ of this perturbed Markov chain.
4. Define the PageRank of a webpage $i$ as the stationary probability $\pi_i$.
5. Find all pages with decent textual match to search and then order those pages by PageRank!

# PageRank - Example



# It Gets More Complicated

While this basic algorithm was the defining idea that launched Google on their path to success, this is far from the end to optimizing search

Nowadays, Google and other web search engines have a LOT more secret sauce to rank pages, most of which they don’t reveal 1) for competitive advantage and 2) to avoid gaming of their algorithms.

# Differential Privacy

# Privacy Preservation

A real-world example (adapted from *The Ethical Algorithm* by Kearns and Roth; based on protocol by Warner [1965]).

And gives a sense of how randomness is actually used to protect privacy.

# Privacy Preservation with Randomness

You're working with a social scientist. They want to get accurate data on the rate at which people cheat on their romantic partners.

We know about polling accuracy!

Do a poll, call up a random sample of adults and ask them "have you ever cheated on your romantic partner?"

Use a tail-bound to estimate the needed number $n$ get a guaranteed good estimate, right?

You do that, and somehow, no one says they cheated.

# What’s the problem?

People lie.

Or they might be concerned about you keeping this data.

Databases can be leaked (or infiltrated. Or subpoenaed).

You don’t want to hold this data, and the people you’re calling don’t want you to hold this data.

# Doing Better With Randomness

You don’t really need to know **who** was cheating. Just how many people were.

Here’s a protocol:

Please flip a coin.

- If the coin is heads, or you have ever cheated, please tell me “heads”
- If the coin is tails and you have not ever cheated, please tell me “tails”

# Will it be private?

If you are someone who has cheated, and you report heads can that be used against you? Not substantially – just say “no the coin came up heads!”

You discover your partner said heads, what’s the probability that they cheated?

# Will it be private?

If you are someone who has cheated on your spouse, and you report heads can that be used against you? Not substantially – just say “no the coin came up heads!”

$$\mathbb{P}(C|H) = \frac{\mathbb{P}(H|C)\cdot\mathbb{P}(C)}{\mathbb{P}(H)} = \frac{1\cdot\mathbb{P}(C)}{\frac{1}{2}\mathbb{P}(\overline{C}) + 1\cdot\mathbb{P}(C)}$$

Is this a substantial change?

No. For real world values (~15%) of $\mathbb{P}(C)$, the probability estimate would increase (to ~26%). But that isn’t too damaging.

# But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you asked 100 people the "heads/tails" question, and 60 people said "heads." What do you predict would be the number of people who cheated on a partner?

Can you generalize your idea for $n$ people polled, and $X$ the number of people that said "heads"?

# But will it be accurate?

But we've lost our data haven't we? People answered a different question. Can we still estimate how many people cheated?

Suppose you poll $n$ people, and let $X$ be the number of people who said "heads" We'll find an estimate $Y$ of the number of people who cheated in the sample, and let $p$ be the true probability of cheating in the population.
What should $Y$ be? Can we draw a margin of error around $Y$?

$\mathbb{P}(X_i = 1) = \frac{1}{2} + \frac{1}{2} \cdot p$

$\mathbb{E}[X] = \frac{n}{2} + \frac{1}{2}\mathbb{E}[Y]$

We'll define $Y$ to be: $Y = 2\left(X - \frac{n}{2}\right)$.
This is a **definition**, based on how the $\mathbb{E}[Y]$ should relate to the $\mathbb{E}[X]$.

# But will it be accurate?

$Y = 2\left(X - \frac{n}{2}\right)$

$\mathrm{Var}(X) = \mathrm{Var}(\sum X_i) = \sum \mathrm{Var}(X_i)$

$\mathrm{Var}(X_i)$? It's an indicator with parameter $p + (1-p) \cdot \frac{1}{2} = \frac{1}{2} + \frac{p}{2}$

So $\mathrm{Var}(X_i) = \left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right)$

$\mathrm{Var}(Y) = 4\mathrm{Var}(X) = 4n\mathrm{Var}(X_i) = 4n\left(\frac{1}{2} + \frac{p}{2}\right)\left(\frac{1}{2} - \frac{p}{2}\right) \leq \frac{4n}{4} = n$

The variance is 4 times as much as it would have been for a non-anonymous poll.

# Can we use Chernoff?

## (Multiplicative) Chernoff Bound

Let $X_1, X_2, \dots, X_n$ be *independent* Bernoulli random variables.
*Let* $X = \sum X_i$, and $\mu = \mathbb{E}[X]$. For any $0 \le \delta \le 1$

$$\mathbb{P}(X \ge (1+\delta)\mu) \le \exp\left(-\frac{\delta^2\mu}{3}\right) \text{ and } \mathbb{P}(X \le (1-\delta)\mu) \le \exp\left(-\frac{\delta^2\mu}{2}\right)$$

What happens with $\mathrm{n} = 1000$ people?

What range will we be within at least 95% of the time?

# A different inequality

If we try to use Chernoff, we'll hit a frustrating block.

Since $\mu$ depends on $p$, $p$ appears in the formula for $\delta$. And we wouldn't get an absolute guarantee unless we could plug in a $p$.

And it'll turn out that as $p \to 0$ that $\delta \to \infty$ so we don't say anything then.

Luckily, there's always another bound…

# ☹ Can't bound $\delta$ without bounding $p$

The right tail is the looser bound, so ensuring the right tail is less than 2.5% gives us the needed guarantee.

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) = \exp\left(-\frac{\delta^2 1000p}{3}\right) \leq .025$$

$$-\frac{\delta^2 1000p}{3} \leq \ln(.025)$$

$$-\delta^2 \leq \frac{3 \cdot \ln(.025)}{1000p}$$

$$\delta \geq \sqrt{\frac{-3\ln(.025)}{1000p}}$$

As $p \to 0$, $\delta \to \infty$ – we're not actually making a claim anymore.

# Hoeffding's Inequality

## Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
*Let* $\bar{X} = \sum X_i/n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$. Why?

$$Y = 2\left(X - \frac{n}{2}\right) \text{ or } X = \frac{Y+n}{2}$$

$$\begin{aligned}
&|X - \mathbb{E}[X]| \\
&= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y+n}{2}\right]\right| \\
&= \left|\frac{Y+n}{2} - \mathbb{E}\left[\frac{Y}{2}\right] - \frac{n}{2}\right| \\
&= \left|\frac{Y}{2} - \mathbb{E}\left[\frac{Y}{2}\right]\right| \\
&= \frac{1}{2}|Y - \mathbb{E}[Y]|
\end{aligned}$$

So $|X - \mathbb{E}[X]| \geq t$ if and only if $\frac{1}{2}|Y - \mathbb{E}[Y]| \geq t$ iff $|Y - \mathbb{E}[Y]| \geq 2t$.

# Hoeffding's Inequality

## Hoeffding's Inequality

Let $X_1, X_2, \ldots, X_n$ be *independent* RVs, each with range $[0,1]$.
*Let* $\bar{X} = \sum X_i / n$, and $\mu = \mathbb{E}[\bar{X}]$. For any $t \geq 0$

$$\mathbb{P}(|\bar{X} - \mathbb{E}[\bar{X}]| \geq t) \leq 2\exp(-2nt^2)$$

How close will we be with n=1000 with probability at least .95?

$|X - \mathbb{E}[X]| \geq t$ if and only if $|Y - \mathbb{E}[Y]| \geq 2t$.

# Margin of Error

$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]| \geq t/2) \leq 2\exp(-2nt^2) \leq .05$

For $n = 1000$, we get:

$2\exp\left(-2n\left(\frac{t}{2}\right)^2\right) \leq .05 \Rightarrow -\frac{2000t^2}{4} \leq \ln(.025) \Rightarrow t \leq .086.$

$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq .086) \leq .05$

So our margin of error is about 8.6%.

To get a margin-of-error of 5% need $2\exp\left(-2n\left(\frac{.05}{2}\right)^2\right) \leq .05$

$n \geq 2952$

# How much do we lose?

We lose a factor of two in the length of the margin (equivalently, we'd need to talk to $4$ times as many people to have the same confidence.

You can also control this tradeoff.

Want more accuracy? Make it roll a die: report 1 if cheated (truth o/w)

Want more security? Make it Bernoulli with probability $p \gg \frac{1}{2}$ or cheated have the same report (e.g. report "die roll 1 [and didn't cheat]" or "die roll 2-6 [or did cheat]"

# In The Real World

Injecting randomness to preserve privacy is a real thing.

Instead of having everyone flip a coin, “random noise” can be inserted after all the data has been collected.

**Differential privacy** is being used to protect the 2020 Census data.

The overall count of people in each state is exact (well, exactly the data they collected). But the data per block or per city will be randomized to protect against revealing who lives where.

This video nicely explains what’s involved. Notice that the accuracy guarantees come in the same “inside-margin-of-error-with-probability” guarantees we’ve been giving for our randomness (just much stronger).