Unbiased estimators
Markov chains and PageRank

CSE 312 Spring 26

Lecture 25

General Recipe (single parameter)

1. **Input** Given n i.i.d. samples $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta\right)$.
   - For discrete $\mathcal{L}(x_1, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
   - For continuous $\mathcal{L}(x_1, \ldots, x_n ; \theta) = \prod_{i=1}^{n} f(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots, x_n ; \theta)$
4. **Differentiate** Compute $\frac{\mathrm{d}}{\mathrm{d}\theta} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

Definition of Estimator

e.g. $\theta =$ parameter of an Exponential distin

θ: quantity we're trying to estimate

think of these as i. i. d. instances of $X \sim \text{Exp}(\theta)$

This is a constant

$X_1, X_2, \ldots, X_n$: i.i.d. data

$\hat{\theta}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$: estimation of θ based on specific instantiation of the data

This is a r.v. because it's a function of r.v.s

$\hat{\theta}(X_1, X_2, \ldots, X_n)$: estimator of the unknown θ

Sometimes just write $\hat{\theta}$

MLE for pink jelly beans

Q: What is the likelihood function $\boldsymbol{P}\{X = x \mid \theta\}$?

$$\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

X = # pink jelly beans in sample

Q: What is $\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$?

$$\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \frac{1000x}{n}$$

This holds $\forall x$

➔ $\hat{\theta}_{ML}(X) = \frac{1000X}{n}$

MLE estimation for normal distribution

$$\hat{\theta}_{\mu}\,(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

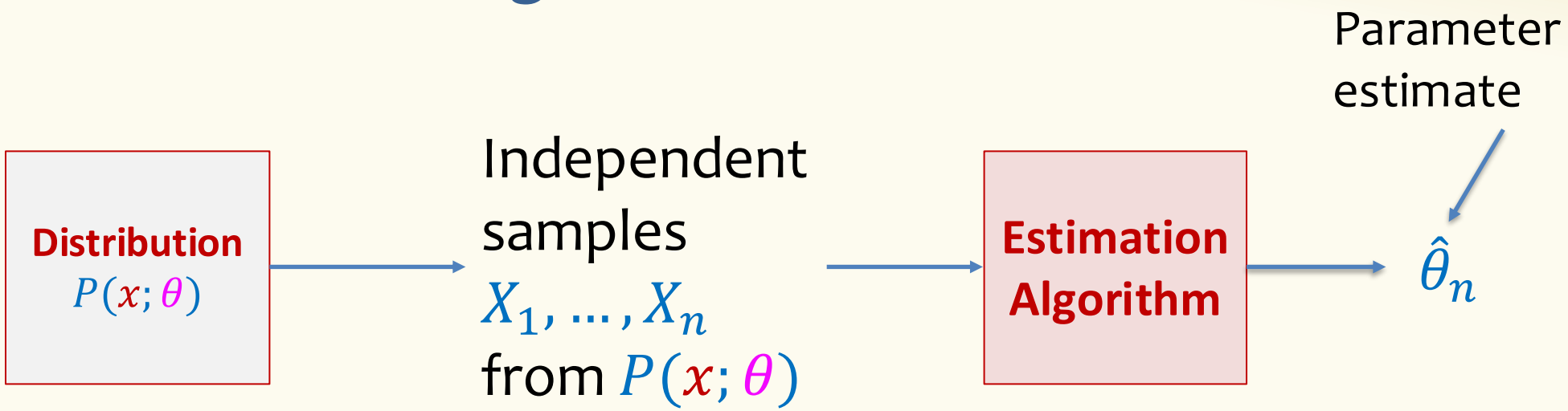
These hold for all $x_1, \dots, x_n$

$$\hat{\theta}_{\sigma^2}(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2$$

➔ $$\hat{\theta}_{\mu}\,(X_1, X_2, \dots, X_n) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

➔ $$\hat{\theta}_{\sigma^2}(X_1, X_2, \dots, X_n) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{X_1 + X_2 + \cdots + X_n}{n}\right)^2$$

When is an estimator good?



θ = unknown parameter

Definition. An estimator of parameter θ is an **unbiased estimator** if

$$\mathbb{E}\left[\hat{\theta}_n(X_1, \ldots, X_n)\right] = \theta.$$

Note: This expectation is over the samples $X_1, \ldots, X_n$

MLE for pink jelly beans

Q: What is the likelihood function $\boldsymbol{P}\{X = x \mid \theta\}$?

$$\boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \binom{n}{x} \left(\frac{\theta}{1000}\right)^{x} \cdot \left(1 - \frac{\theta}{1000}\right)^{n-x}$$

1000 jelly beans total

X = # pink jelly beans in sample

Q: What is $\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta)$?

$$\hat{\theta}_{ML}(X = x) = \arg\max_{\theta} \boldsymbol{P}(\mathrm{X} = \mathrm{x}; \theta) = \frac{1000x}{n}$$

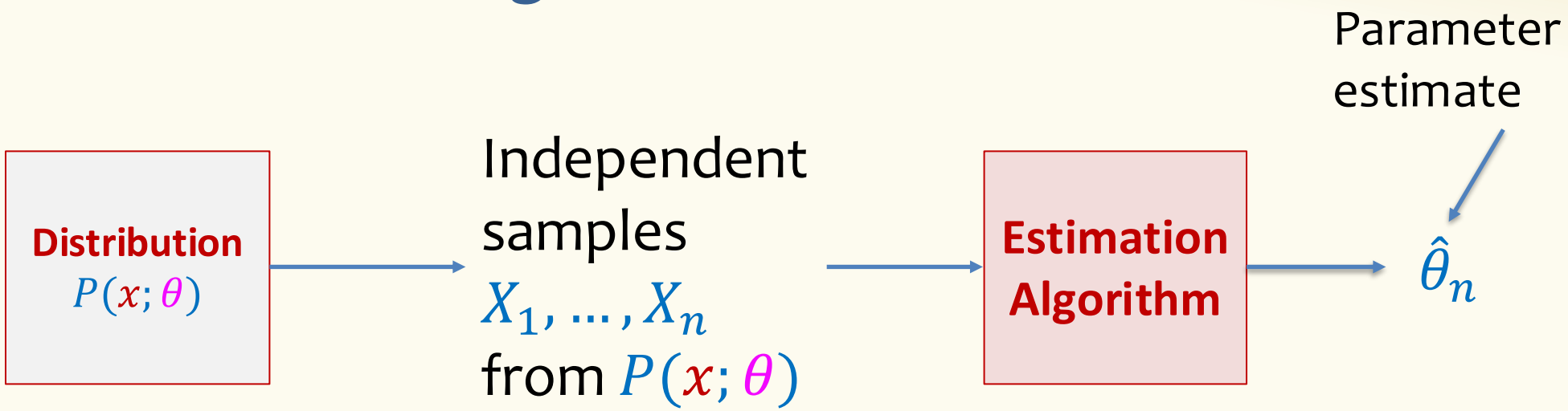
This holds $\forall x$

➔ $\hat{\theta}_{ML}(X) = \frac{1000X}{n}$

Fact. $\hat{\theta}_{ML}(X)$ is unbiased

Three samples from $U(0, \theta)$

When is an estimator good?



θ = unknown parameter

Definition. An estimator is **unbiased** if $\mathbb{E}[\hat{\theta}_n] = \theta$ for all $n \geq 1$.

Definition. An estimator is **consistent** if $\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta$.

Theorem. MLE estimators are consistent.

(But not necessarily unbiased)

Example – Consistency

Normal outcomes $X_1, \ldots, X_n$ i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Population variance – Biased!

$\widehat{\Theta}_{\sigma^2}$ is "consistent"

Example – Consistency

Normal outcomes $X_1, \ldots, X_n$ i.i.d. according to $\mathcal{N}(\mu, \sigma^2)$ Assume: $\sigma^2 > 0$

$$\widehat{\Theta}_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Population variance – Biased!

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \widehat{\Theta}_{\mu}\right)^2$$

Sample variance – Unbiased!

$\widehat{\Theta}_{\sigma^2}$ converges to same value as S_n^2, i.e., σ^2, as $n \to \infty$.

$\widehat{\Theta}_{\sigma^2}$ is "consistent"

Why does it matter?

- When statisticians are estimating a variance from a sample, they usually divide by $n-1$ instead of n.

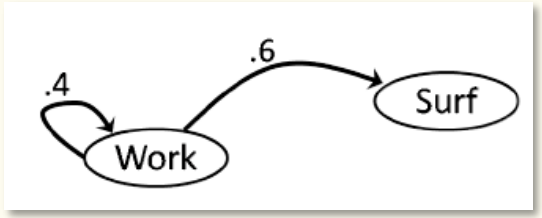
- They and we not only want good estimators (unbiased, consistent)
  - They/we also want **confidence bounds**
    - Upper bounds on the probability that these estimators are far the truth about the underlying distributions
  - Confidence bounds are just like what we wanted for our polling problems, but CLT is usually not the best thing to use to get them (unless the variance is known)

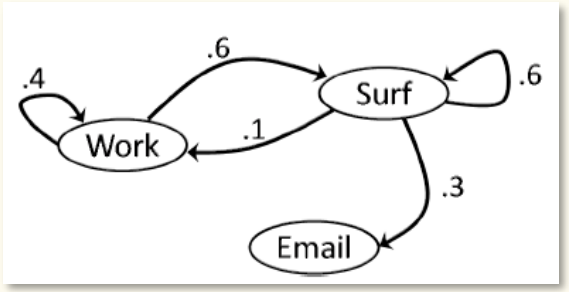
Agenda

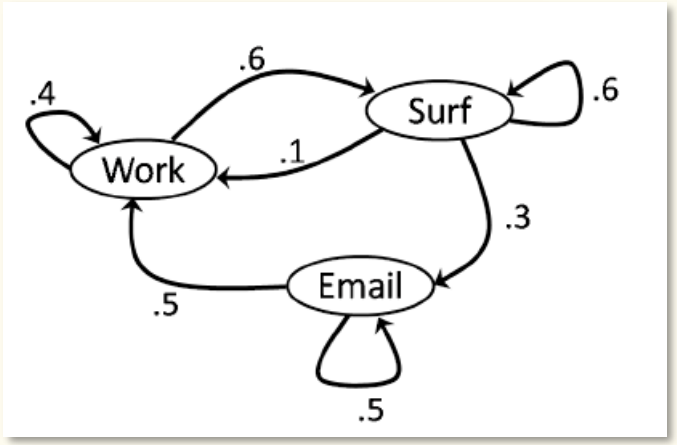
- Unbiased Estimation
- Markov Chains
- Application: PageRank

A typical day in my life….

time $t = 0$

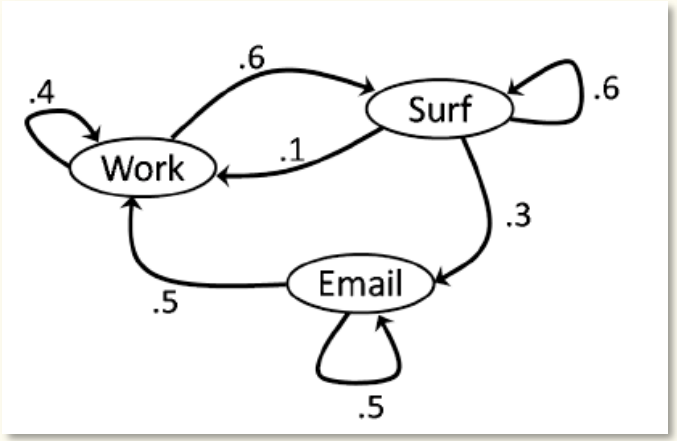






A typical day in my life

How do we interpret this diagram?



This kind of random process is called a **Markov Chain**

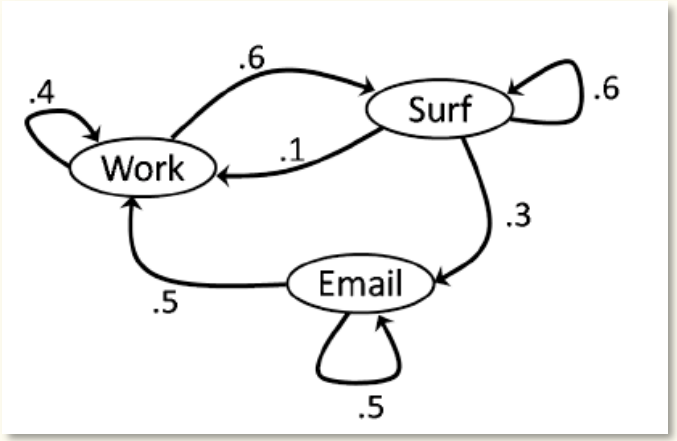
At *each* time step t

- 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 my moving to each of the states at time $t + 1$ (as well as staying the same)
    - so **labels on out-edges sum to** 1

e.g. If I am in Email, there is a 50-50 chance I will be in each of Work or Email at the next time step, but I will never be in state Surf in the next step.

This diagram looks vaguely familiar if you took CSE 311 …

Markov chains are a special kind of *probabilistic (finite) automaton*

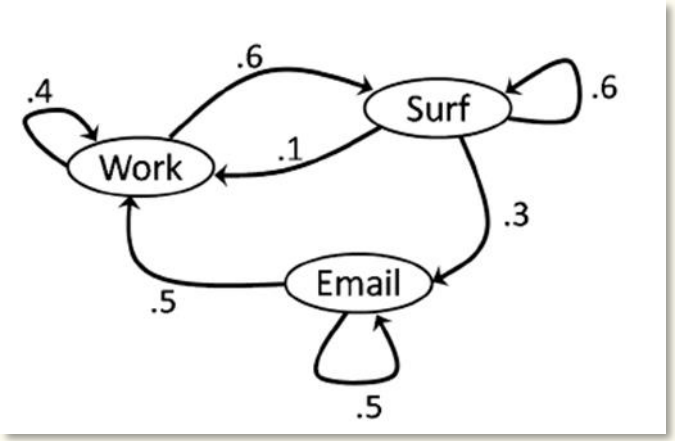


The diagrams look a bit like those of Deterministic Finite Automata (DFAs) you saw in 311 except that…

- There are no input symbols on the edges
  - Think of there being only one kind of input symbol "another tick of the clock" so no need to mark it on the edge
- They have multiple out-edges like an NFA, except that they come with probabilities

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

Many interesting questions about Markov Chains

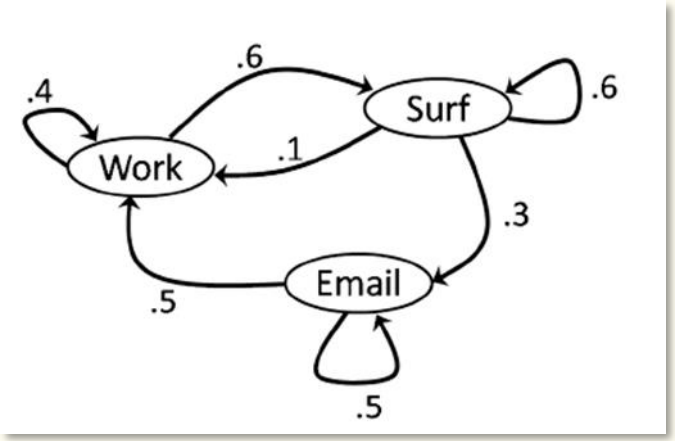


1. **What is the probability that I am in state s at time 1?**
2. **What is the probability that I am in state s at time 2?**
3. **What is the probability that I am in state s at some time t far in the future?**

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

To answer these questions, we need to understand how the probability distribution over states I could be in evolves over time.

Many interesting questions about Markov Chains



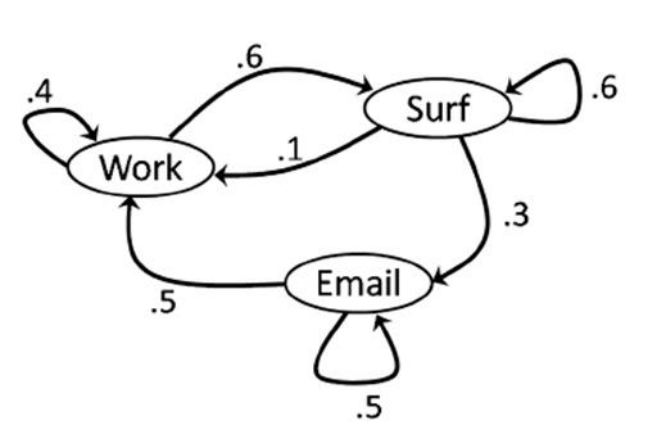
1. **What is the probability that I am in state s at time 1?**
2. **What is the probability that I am in state s at time 2?**
3. **What is the probability that I am in state s at some time t far in the future?**

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

To answer these questions, we need to understand how the probability distribution over states I could be in evolves over time.

$$M = \begin{array}{c} \\ \text{W} \\ \text{S} \\ \text{E} \end{array} \begin{array}{c} \begin{array}{ccc} \text{W} & \text{S} & \text{E} \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$

An organized way to understand how the distribution evolves



Suppose we know

$$q_W^{(t)} = P(\text{Working at time t})$$

$$q_S^{(t)} = P(\text{Surfing at time t})$$

$$q_E^{(t)} = P(\text{Emailing at time t})$$

By the law of total probability

$$\Pr(\text{Working at t+1}) = P(\text{Working at t}) \cdot 0.4 + \Pr(\text{Surfing at t}) \cdot 0.1 + \Pr(\text{Emailing at t}) \cdot 0.5$$

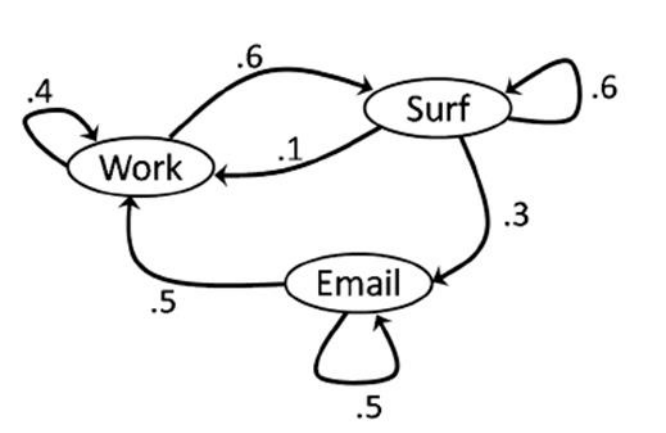
Pr (working at t+1 | working at t)

Pr (working at t+1 | surfing at t)

Pr (working at t+1 | emailing at t)

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

An organized way to understand how the distribution evolves



$q_W^{(t)} = P(\text{in state Work at time t})$

$q_S^{(t)} = P(\text{in state Surf at time t})$

$q_E^{(t)} = P(\text{in state Email at time t})$

Pr(Surfing at t+1) = P(Working at t) ·0.6 + Pr(Surfing at t) · 0.6 + Pr (email at t) · 0

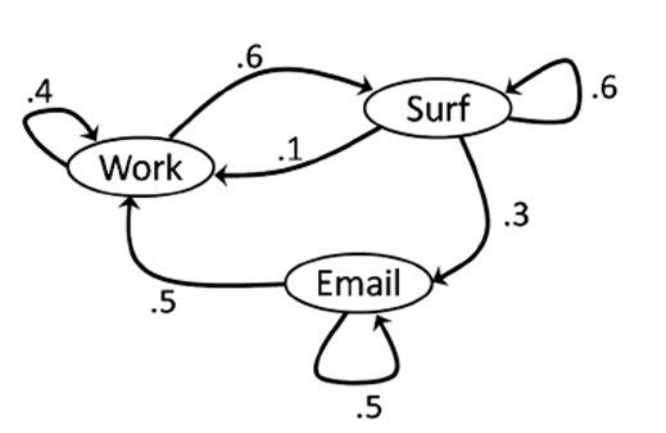
Pr (surfing at t+1 | working at t)

Pr (surfing at t+1 | surfing at t)

Pr (surfing at t+1 | emailing at t)

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

An organized way to understand the distribution at time t



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

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

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

$q_W^{(t)} =$ P(in state Work at time t)

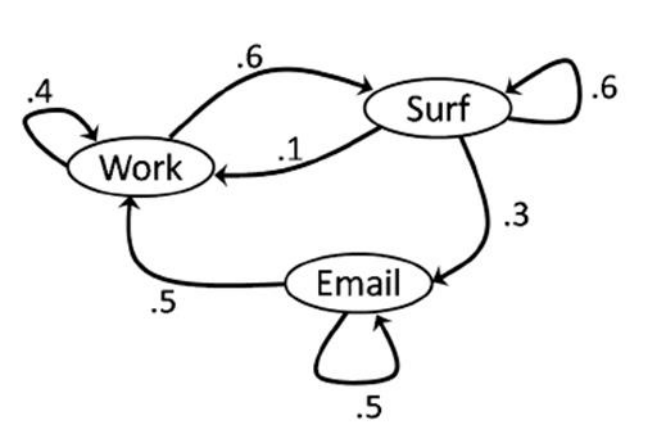
$q_S^{(t)} =$ P(in state Surf at time t)

$q_E^{(t)} =$ P(in state Email at time t)

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

$$M = \begin{array}{c} \\ \text{W} \\ \text{S} \\ \text{E} \end{array} \begin{array}{c} \begin{array}{ccc} \text{W} & \text{S} & \text{E} \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$

An organized way to understand the distribution at time t



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

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

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

$q_W^{(t)} =$ P(in state Work at time t)

$q_S^{(t)} =$ P(in state Surf at time t)

$q_E^{(t)} =$ P(in state Email at time t)

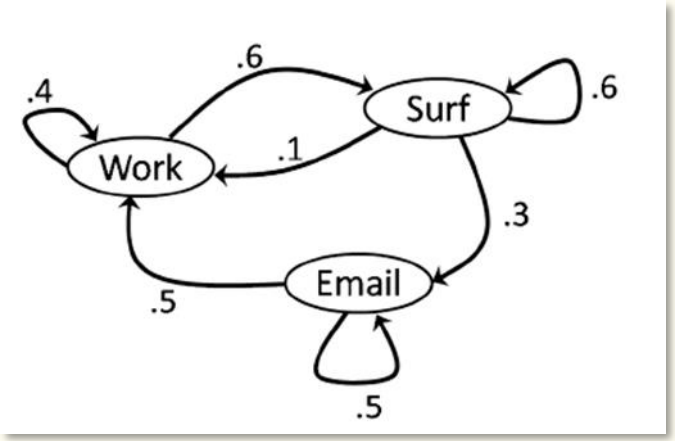
$$[q_W^{(t+1)}, q_S^{(t+1)}, q_E^{(t+1)}] = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}] \begin{matrix} \\ \boldsymbol{M} \end{matrix} \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Write $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$

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

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)} \boldsymbol{M}$$

Many interesting questions about Markov Chains



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

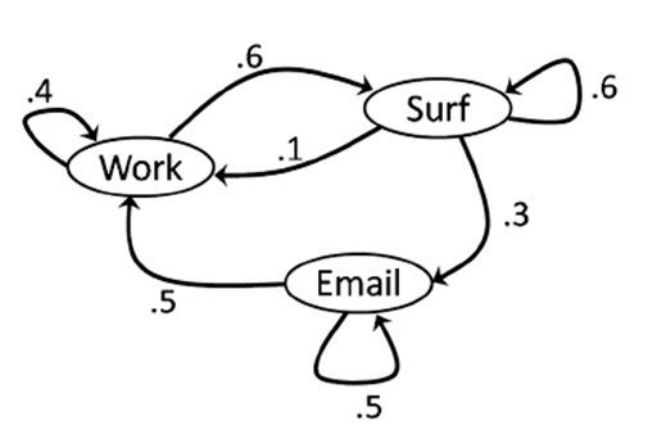
1. **What is the probability that I am in state s at time 1?**

In other words, what is $q^{(1)} = [q_W^{(1)}, q_S^{(1)}, q_E^{(1)}]$?

2. **What is the probability that I am in state s at time 2?**

In other words, what is $q^{(2)} = [q_W^{(2)}, q_S^{(2)}, q_E^{(2)}]$?

An organized way to understand the distribution at time t



$$q^{(0)} = [1,\ 0,\ 0]$$

Start out working

$$q^{(1)} = q^{(0)}M$$

$$q^{(1)} = [0.4,\ 0.6,\ 0] = [1,\ 0,\ 0]\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$q_W^{(t)} =$ P(in state Work at time t)

$q_S^{(t)} =$ P(in state Surf at time t)

$q_E^{(t)} =$ P(in state Email at time t)

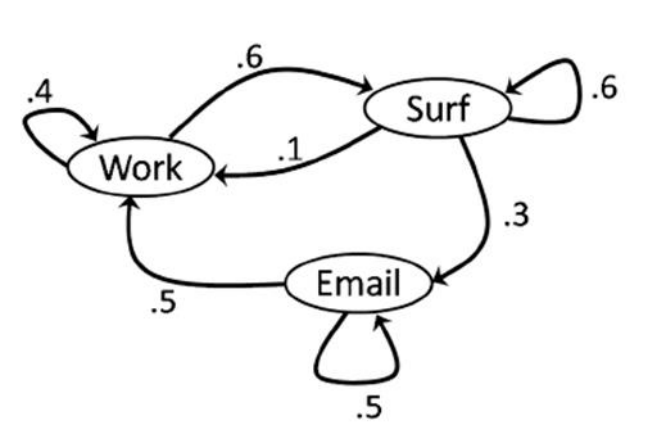
$$M = \begin{array}{c} \\ W \\ S \\ E \end{array}\begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$

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

$\boldsymbol{M}$

$$q^{(t+1)} = q^{(t)}M$$

An organized way to understand the distribution at time t



$q^{(0)} = [1, 0, 0]$ Start out working

$$q^{(1)} = q^{(0)}M$$

$$q^{(1)} = [0.4, 0.6, 0] = [1, 0, 0]\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$$q^{(2)} = q^{(1)}M$$

$$q^{(2)} = [0.22, 0.6, 0.18] = [0.4, 0.6, 0]\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

$q_W^{(t)} = P(\text{in state Work at time t})$

$q_S^{(t)} = P(\text{in state Surf at time t})$

$q_E^{(t)} = P(\text{in state Email at time t})$

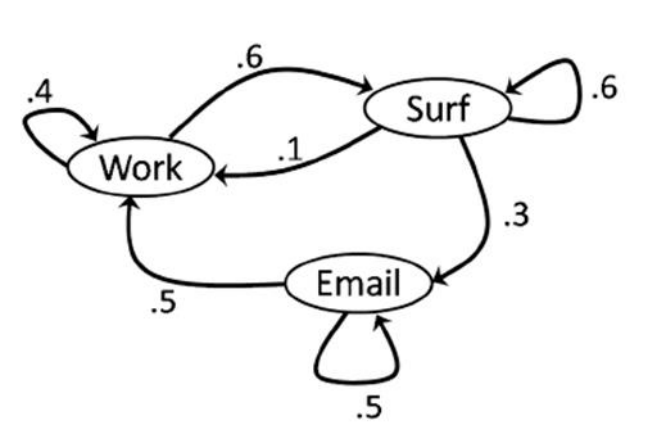
$$M = \begin{array}{c} \\ \text{W} \\ \text{S} \\ \text{E} \end{array}\begin{array}{c}\begin{array}{ccc} \text{W} & \text{S} & \text{E} \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}\end{array}$$

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

$\boldsymbol{M}$

$$q^{(t+1)} = q^{(t)}M$$

An organized way to understand the distribution at time t



$q^{(0)} = [1, 0, 0]$ Start out working

$q^{(1)} = q^{(0)}M$

$q^{(1)} = [0.4, 0.6, 0] = [1, 0, 0]\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$

$q^{(2)} = q^{(1)}M$

$q^{(2)} = [0.22, 0.6, 0.18] = [0.4, 0.6, 0]\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$

$q_W^{(t)} =$ P(in state Work at time t)

$q_S^{(t)} =$ P(in state Surf at time t)

$q_E^{(t)} =$ P(in state Email at time t)

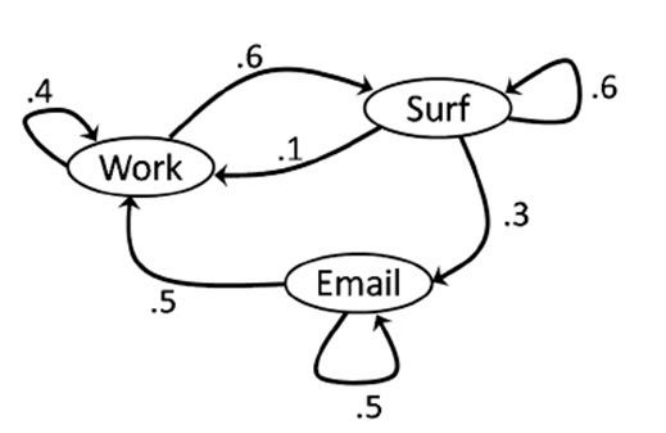
$q^{(2)} = [1, 0, 0]\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}\begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix}$

$= q^{(0)}M^2$

$$M = \begin{array}{c} \\ \text{W} \\ \text{S} \\ \text{E} \end{array}\begin{array}{c} \begin{array}{ccc} \text{W} & \text{S} & \text{E} \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$

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

An organized way to understand the distribution at time t



M^2 captures 2-step transition probabilities

Start out working

$$q^{(1)} = [0.4,\ 0.6,\ \quad] = [1,\ 0,\ 0]\begin{bmatrix}0.4 & 0.6 & 0\\ 0.1 & 0.6 & 0.3\\ 0.5 & 0 & 0.5\end{bmatrix}$$

$$q^{(2)} = q^{(1)}M$$

$$q^{(2)} = [0.22,\ 0.6,\ 0.18] = [0.\ \ 0.6,\ 0]\begin{bmatrix}0.4 & 0.6 & 0\\ 0.1 & 0.6 & 0.3\\ 0.5 & 0 & 0.5\end{bmatrix}$$

$q_W^{(t)}$ = P(in state Work at time t)

$q_S^{(t)}$ = P(in state Surf at time t)

$q_E^{(t)}$ = P(in state Email at time t)

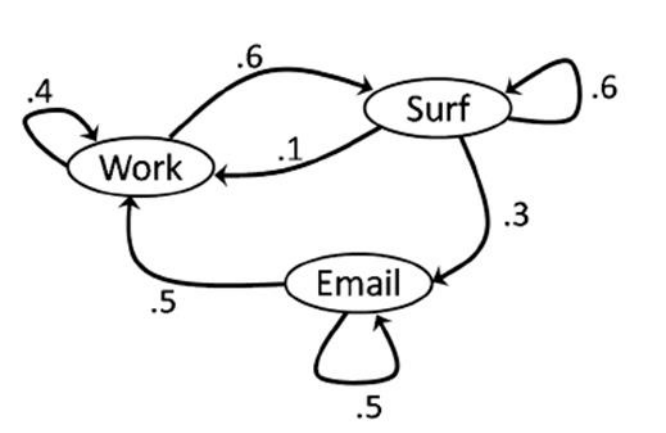
$$q^{(2)} = [1,\ 0,\ 0]\begin{bmatrix}0.4 & 0.6 & 0\\ 0.1 & 0.6 & 0.3\\ 0.5 & 0 & 0.5\end{bmatrix}\begin{bmatrix}0.4 & 0.6 & 0\\ 0.1 & 0.6 & 0.3\\ 0.5 & 0 & 0.5\end{bmatrix}$$

$$= q^{(0)}M^2$$

$$M = \begin{array}{c} \\ \text{W} \\ \text{S} \\ \text{E}\end{array}\begin{array}{c}\begin{array}{ccc}\text{W} & \text{S} & \text{E}\end{array}\\ \begin{bmatrix}0.4 & 0.6 & 0\\ 0.1 & 0.6 & 0.3\\ 0.5 & 0 & 0.5\end{bmatrix}\end{array}$$

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

An organized way to understand the distribution at time t



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

Write $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$

Then for all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

$q_W^{(t)} =$ P(in state Work at time t)

$q_S^{(t)} =$ P(in state Surf at time t)

$q_E^{(t)} =$ P(in state Email at time t)

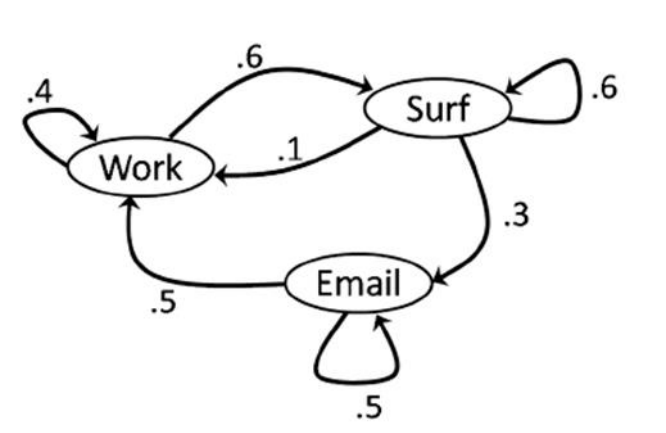
$$\boldsymbol{q}^{(1)} = \boldsymbol{q}^{(0)}\boldsymbol{M}$$

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

$$\boldsymbol{q}^{(3)} = \boldsymbol{q}^{(2)}\boldsymbol{M} = \boldsymbol{q}^{(0)}\boldsymbol{M}^2\boldsymbol{M} = \boldsymbol{q}^{(0)}\boldsymbol{M}^3$$

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

An organized way to understand the distribution at time t



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

$\boldsymbol{M^t}$ captures t -step transition probabilities

ite $\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$

The all $t \geq 0$, $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

$q_W^{(t)} =$ P(in state Work at time t)

$q_S^{(t)} =$ P(in state Surf at time t)

$q_E^{(t)} =$ P(in state Email at time t)

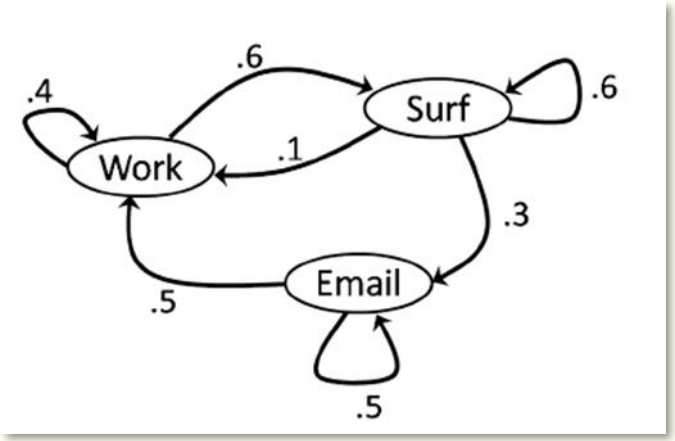
$$\boldsymbol{q}^{(1)} = \boldsymbol{q}^{(0)}\boldsymbol{M}$$

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

$$\boldsymbol{q}^{(3)} = \boldsymbol{q}^{(2)}\boldsymbol{M} = \boldsymbol{q}^{(0)}\boldsymbol{M}^2\boldsymbol{M} = \boldsymbol{q}^{(0)}\boldsymbol{M}^3$$

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

Many interesting questions about Markov Chains



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?**

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

$$\boldsymbol{q}^{(t)} = [q_W^{(t)}, q_S^{(t)}, q_E^{(t)}]$$

$q_W^{(t)} =$ P(in state Work at time t)

$q_S^{(t)} =$ P(in state Surf at time t)

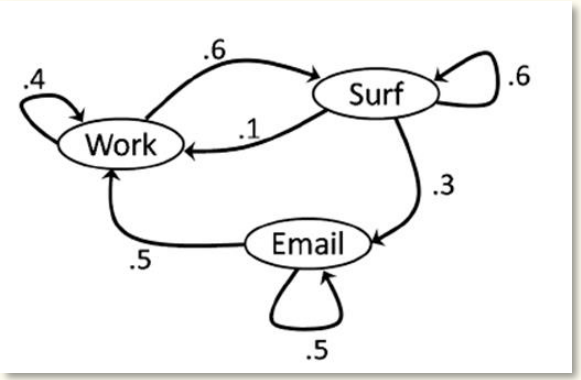
$q_E^{(t)} =$ P(in state Email at time t)

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

What does $\boldsymbol{q}^{(t)}$ look like for really big t ?

M^t as t grows

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



M

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

M^2

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

M^3

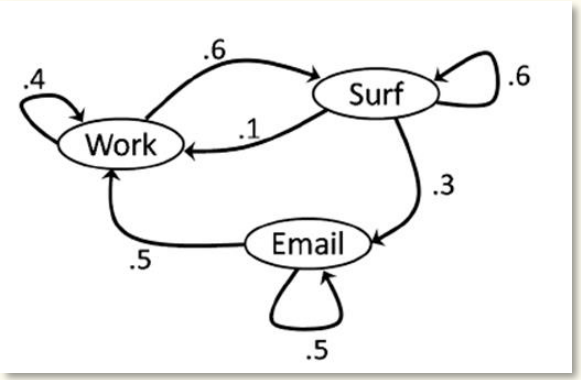
$$\begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .238 & .492 & .270 \\ .307 & .402 & .291 \\ .335 & .450 & .215 \end{pmatrix} \end{array}$$

M^{10}

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

M^t as t grows

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



M

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

M^2

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

M^3

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

M^{10}

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

M^{30}

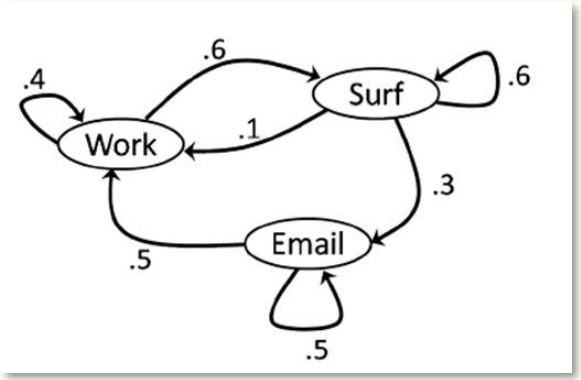
$$\begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .29411764705 & .44117647059 & .26470588235 \\ .29411764706 & .44117647058 & .26470588235 \\ .29411764706 & .44117647059 & .26470588235 \end{pmatrix} \end{array}$$

M^{60}

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

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

M^t as t grows



$$q^{(60)} = q^{(0)} M^{60}$$

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

$q_W^{(t)} = P(\text{in state Work at time t})$

$q_S^{(t)} = P(\text{in state Surf at time t})$

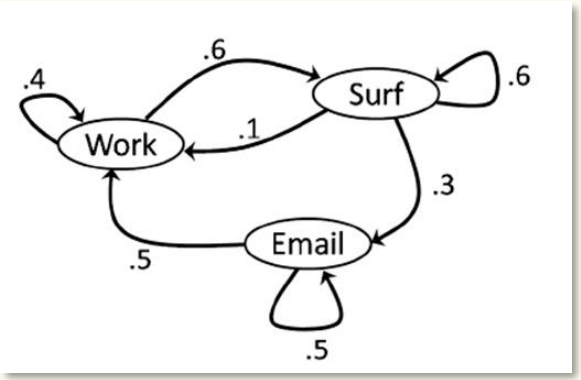
$q_E^{(t)} = P(\text{in state Email at time t})$

$q^{(0)}$

$$[1, 0, 0] \cdot \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

$$= \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \end{pmatrix}$$

M^t as t grows



$$q^{(60)} = q^{(0)} M^{60}$$

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

$q_W^{(t)} = P(\text{in state Work at time t})$

$q_S^{(t)} = P(\text{in state Surf at time t})$

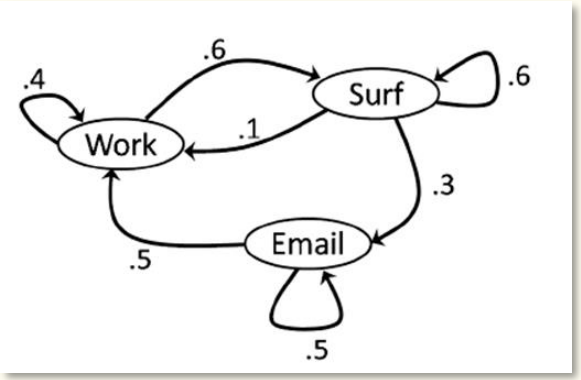
$q_E^{(t)} = P(\text{in state Email at time t})$

$q^{(0)}$

$$[0.5, 0.5, 0\,] \cdot \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

$$= \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \end{pmatrix}$$

M^t as t grows



$$q^{(60)} = q^{(0)} M^{60}$$

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

$q_W^{(t)} = P(\text{in state Work at time t})$

$q_S^{(t)} = P(\text{in state Surf at time t})$

$q_E^{(t)} = P(\text{in state Email at time t})$

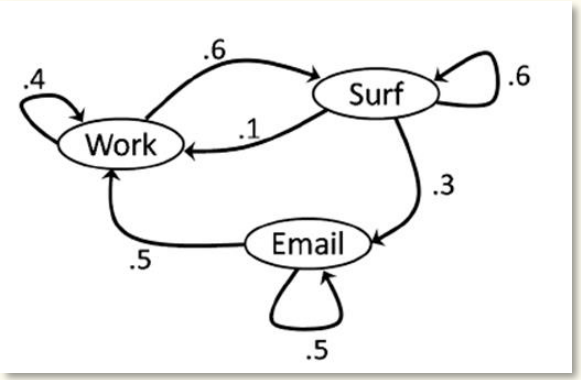
$q^{(0)}$

$$[0.4, 0.5, 0.1\ ] \cdot \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

$$= \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \end{pmatrix}$$

M^t as t grows

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



$$q^{(60)} = q^{(0)} M^{60}$$

$q_W^{(t)} = P(\text{in state Work at time t})$

$q_S^{(t)} = P(\text{in state Surf at time t})$

$q_E^{(t)} = P(\text{in state Email at time t})$

$$[q_W^{(60)}, q_S^{(60)}, q_E^{(60)}] = [q_W^{(0)}, q_S^{(0)}, q_E^{(0)}] \cdot$$

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

$\forall q^{(0)}$ $\quad [q_W^{(60)}, q_S^{(60)}, q_E^{(60)}] = [\ .294117647058823 \quad .441176470588235 \quad .264705882352941\ ]$

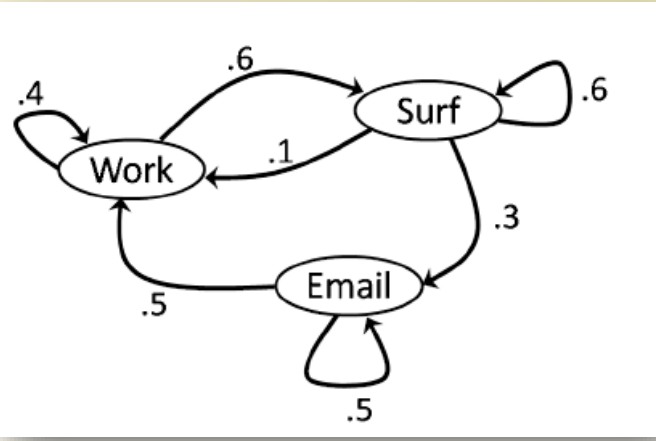
- In the long run, the starting state doesn't really matter!!
- In particular, at any point in the (slightly distant) future, the chance I'm surfing the web is about 44%
- The distribution on states converges to $[\ .294117647058823 \quad .441176470588235 \quad .264705882352941\ ]$

- Suppose that we believe the distribution on states converges to some fixed probability vector $\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$

  In other words, $\lim_{t \to \infty} \boldsymbol{q}^{(t)} = \boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$

- Can we figure out what $\boldsymbol{\pi}$ vis just by looking at M?

Observation



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

Proof: $\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{M}$

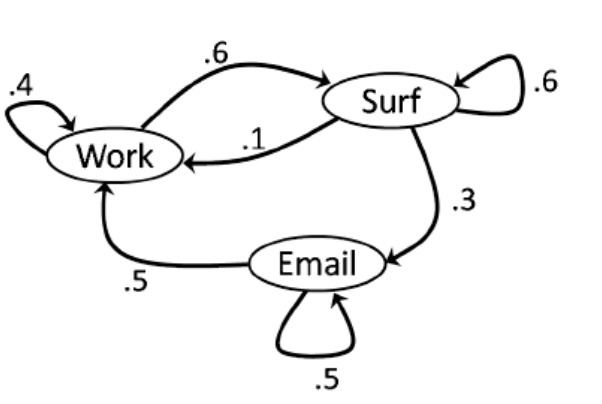
$$\boldsymbol{q}^{(t+2)} = \boldsymbol{q}^{(t+1)}\boldsymbol{M} = \boldsymbol{q}^{(t)}\boldsymbol{M} = \boldsymbol{q}^{(t+1)}$$

Such a $\boldsymbol{q}^{(t)}$ (that never changes) is called a **stationary distribution** and has a special name $\boldsymbol{\pi} = (\pi_W, \pi_S, \pi_E)$

It is the solution to $\boldsymbol{\pi} = \boldsymbol{\pi}\,\boldsymbol{M}$

Solving for Stationary Distribution

$$(\pi_W, \pi_S, \pi_E) = (\pi_W, \pi_S, \pi_E) \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$



$$\pi_W = \pi_W \cdot 0.4 + \pi_S \cdot 0.1 + \pi_E \cdot 0.5$$

$$\pi_S = \pi_W \cdot 0.6 + \pi_S \cdot 0.6 + \pi_E \cdot 0$$

$$\pi_E = \pi_W \cdot 0 + \pi_S \cdot 0.3 + \pi_E \cdot 0.5$$

$$\pi_W + \pi_S + \pi_E = 1$$

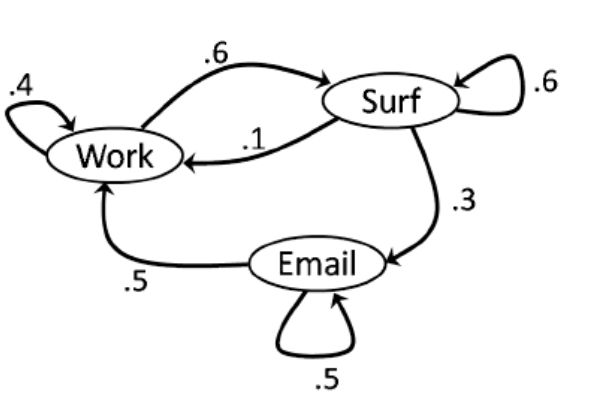
Solving for Stationary Distribution

$$(\pi_W, \pi_S, \pi_E) \begin{pmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{pmatrix} = (\pi_W, \pi_S, \pi_E)$$

$$\pi_W = \pi_W \cdot 0.4 + \pi_S \cdot 0.1 + \pi_E \cdot 0.5$$

$$\pi_S = \pi_W \cdot 0.6 + \pi_S \cdot 0.6 + \pi_E \cdot 0$$

$$\pi_E = \pi_W \cdot 0 + \pi_S \cdot 0.3 + \pi_E \cdot 0.5$$



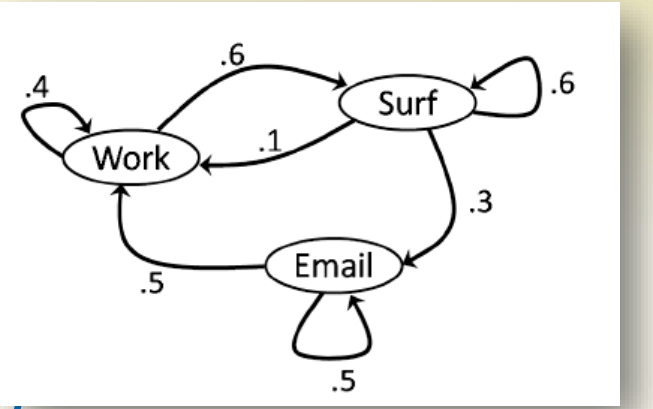
$$\pi_W + \pi_S + \pi_E = 1$$

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

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

Markov Chains summary

$$M = \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.1 & 0.6 & 0.3 \\ 0.5 & 0 & 0.5 \end{bmatrix} \end{array}$$



- A set of n **states** $\{1, 2, 3, \ldots\ n\}$
- A square **transition matrix** $\boldsymbol{M}$, dimension $n \times\ n$

$$\boldsymbol{M}_{ij} = P(\text{transition from } i \text{ to } j)$$

- $\boldsymbol{M}^t{}_{ij}$ = Pr (in state j after t steps| start in state i).
- Nice Markov chains are not sensitive to initial distribution of states. $\boldsymbol{M}^t \rightarrow \boldsymbol{W}$, where all rows in $\boldsymbol{W}$ are the same probability vector $\boldsymbol{\pi}$
- A **stationary distribution** $\boldsymbol{\pi}$ is the solution to:

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

$$\boldsymbol{M}^{60} = \begin{array}{c} \\ W \\ S \\ E \end{array} \begin{array}{c} \begin{array}{ccc} W & S & E \end{array} \\ \begin{pmatrix} .294117647058823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \\ .294117647068823 & .441176470588235 & .264705882352941 \end{pmatrix} \end{array}$$

The Fundamental Theorem of Markov Chains

Theorem. Any nice* Markov chain

has a unique stationary distribution $\boldsymbol{\pi}$.

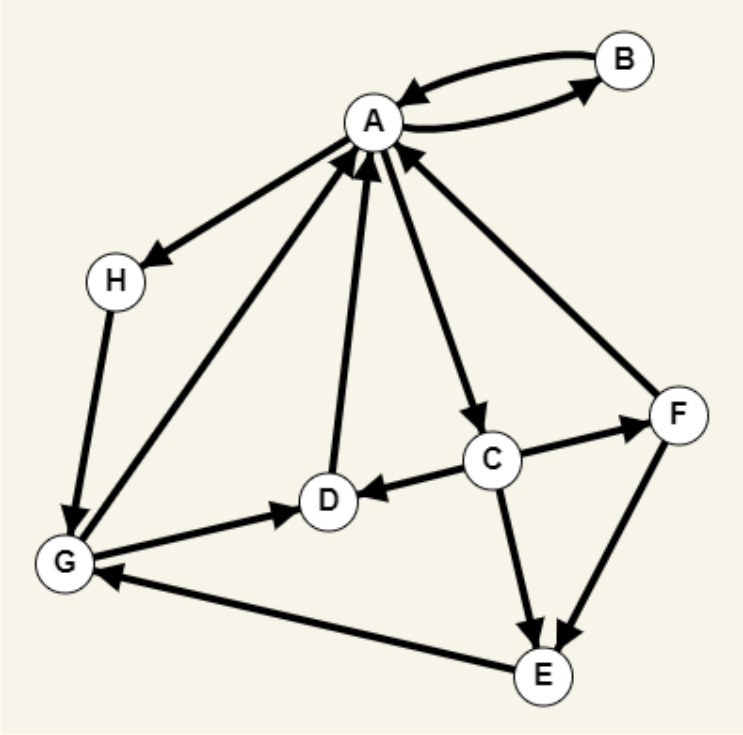
Moreover, as $t \to \infty$, for all i, j, $\lim_{t\to\infty} \boldsymbol{M}_{ij}^{t} = \pi_j$

aperiodic and irreducible: these concepts are beyond us but they turn out to cover a very large class of Markov chains of practical importance.

Another Example of a Markov Chain: Random Walks on Graph

Suppose we start at some node, and at each step transition to a neighboring node with each neighbor equally likely.

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



Agenda

- Unbiased Estimation
- Markov Chains
- Application: 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 actually 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)

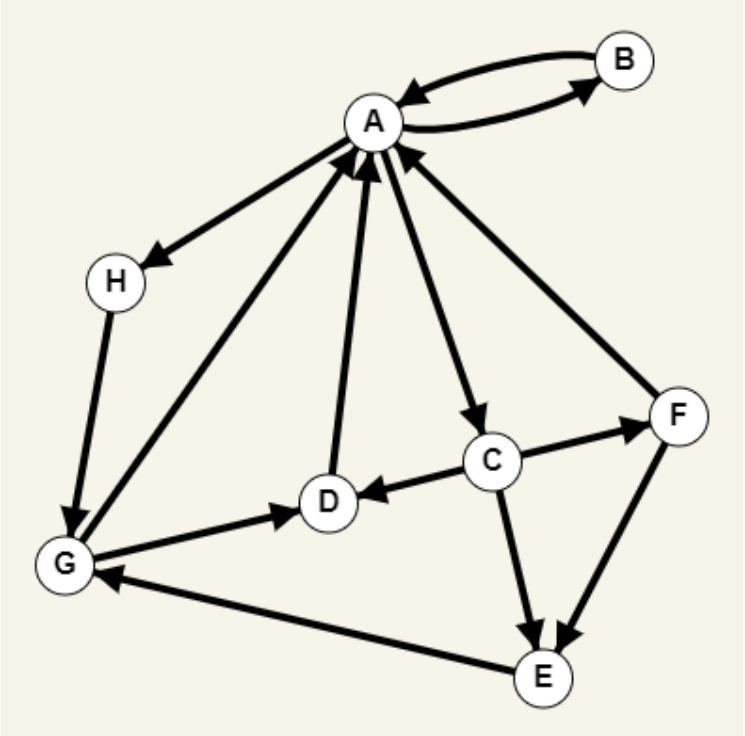
- MacArthur genius prize, Nevanlinna Prize and many other academic honors

PageRank - Idea

Take into account the directed graph structure of the web.

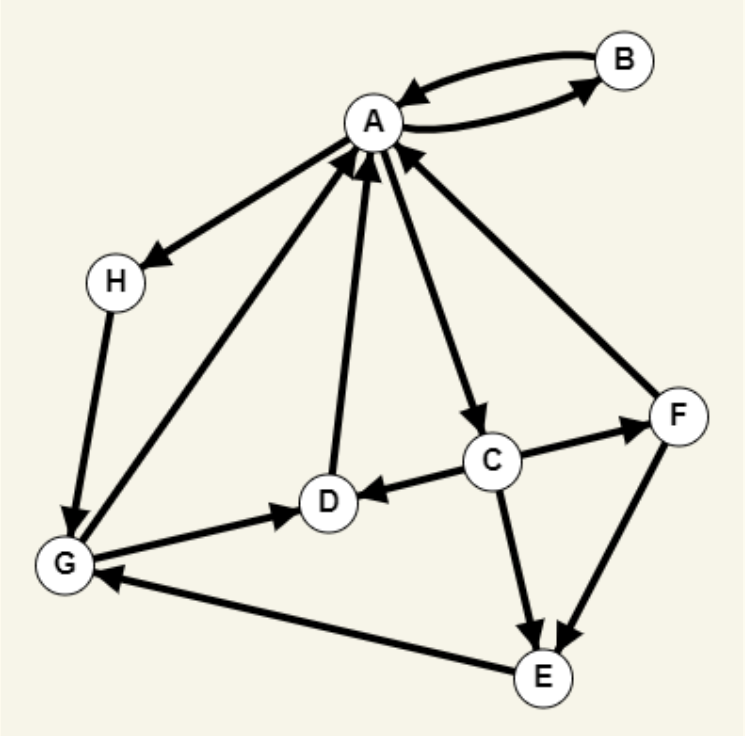
Use **hyperlink analysis** to compute what pages are high quality or have high authority.

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



PageRank - Idea

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



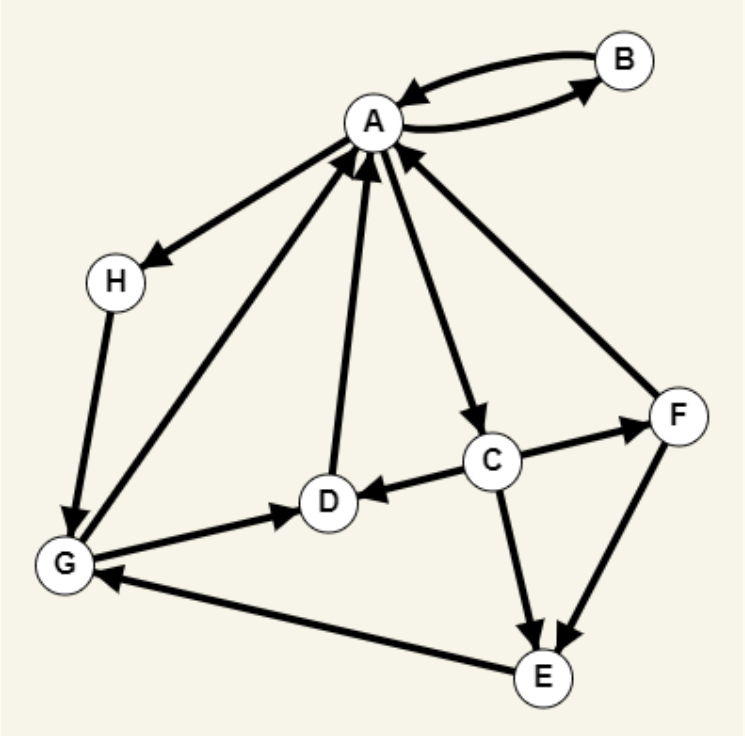
Rank pages by in-degree?

Two problems:

- Susceptible to spamming
- Doesn't take into account quality of link creator.

PageRank - Idea

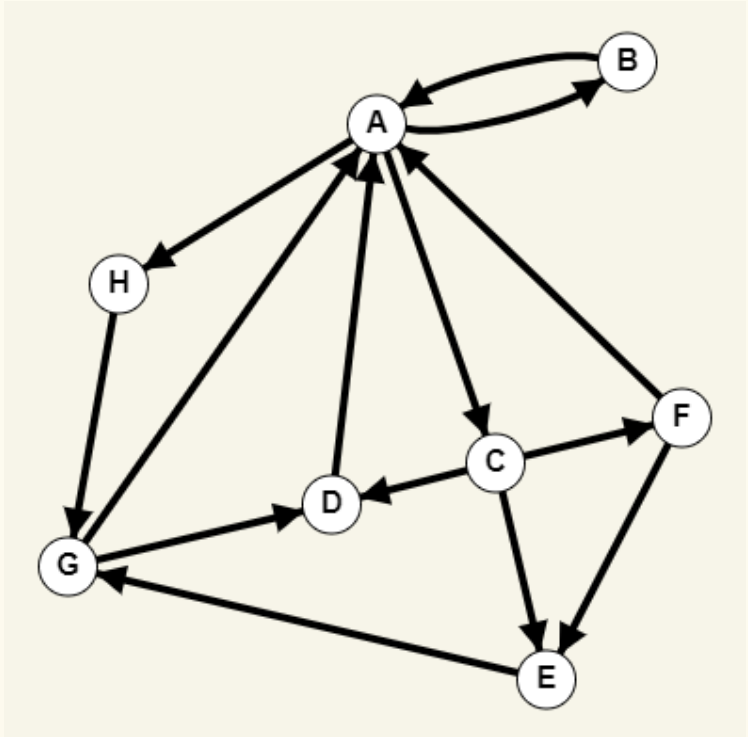
Idea 2 : Perhaps we should weight the links somehow and then use the weights of the in-links to rank pages



Inching towards PageRank

1. A web page has high quality if it's linked to by lots of high quality pages

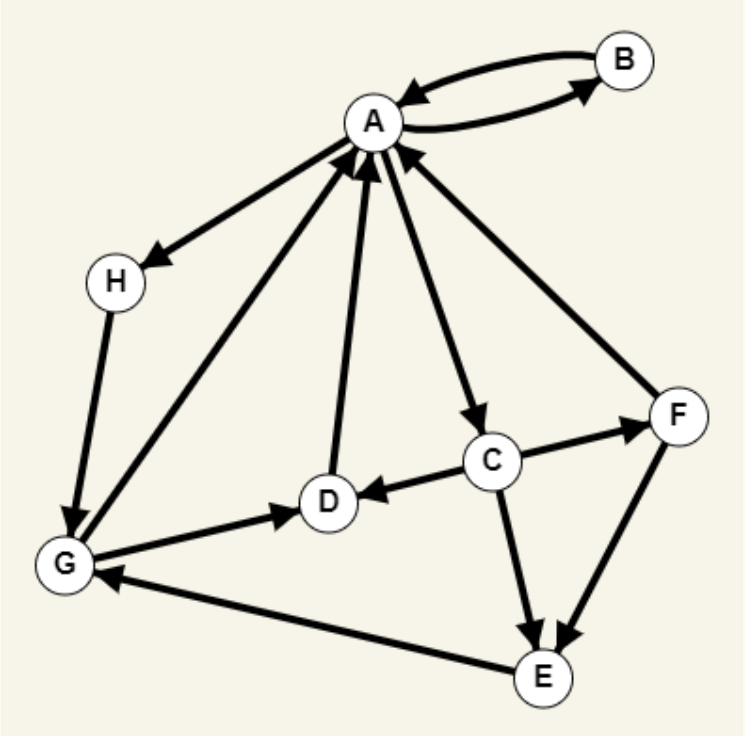
2. A web page is high quality if it links to lots of high quality pages



That's a recursive definition!

Inching towards PageRank

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

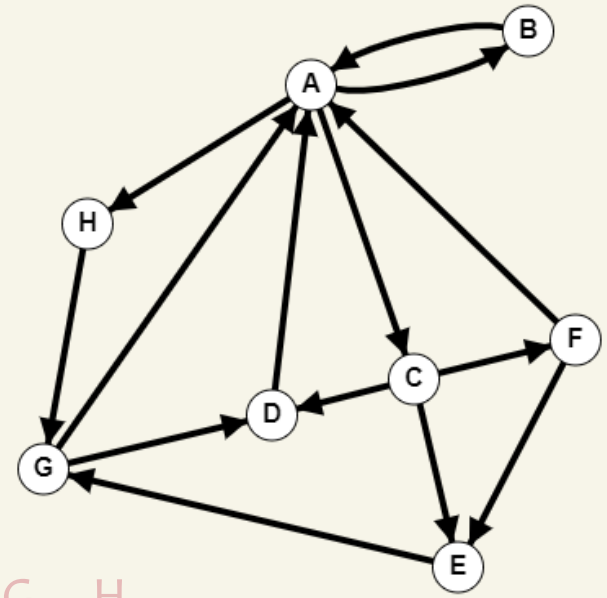


- But $1/d$ of what?

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

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

Define q_x to be the quality of page x



$$q_A = q_B \cdot 1 + q_D \cdot 1 + q_F \cdot \frac{1}{2} + q_G \cdot \frac{1}{2}$$

$$(q_A, q_B, \dots, q_F, q_G, q_H) = (q_A, q_B, \dots, q_F, q_G, q_H)$$

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{3} \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Do you recognize this matrix as representing a Markov Chain we mentioned?

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 $\boldsymbol{q} = \boldsymbol{qM}$.

Do you recognize this matrix-vector equation?

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 $\boldsymbol{qM = q}$ **Seem familiar?**

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

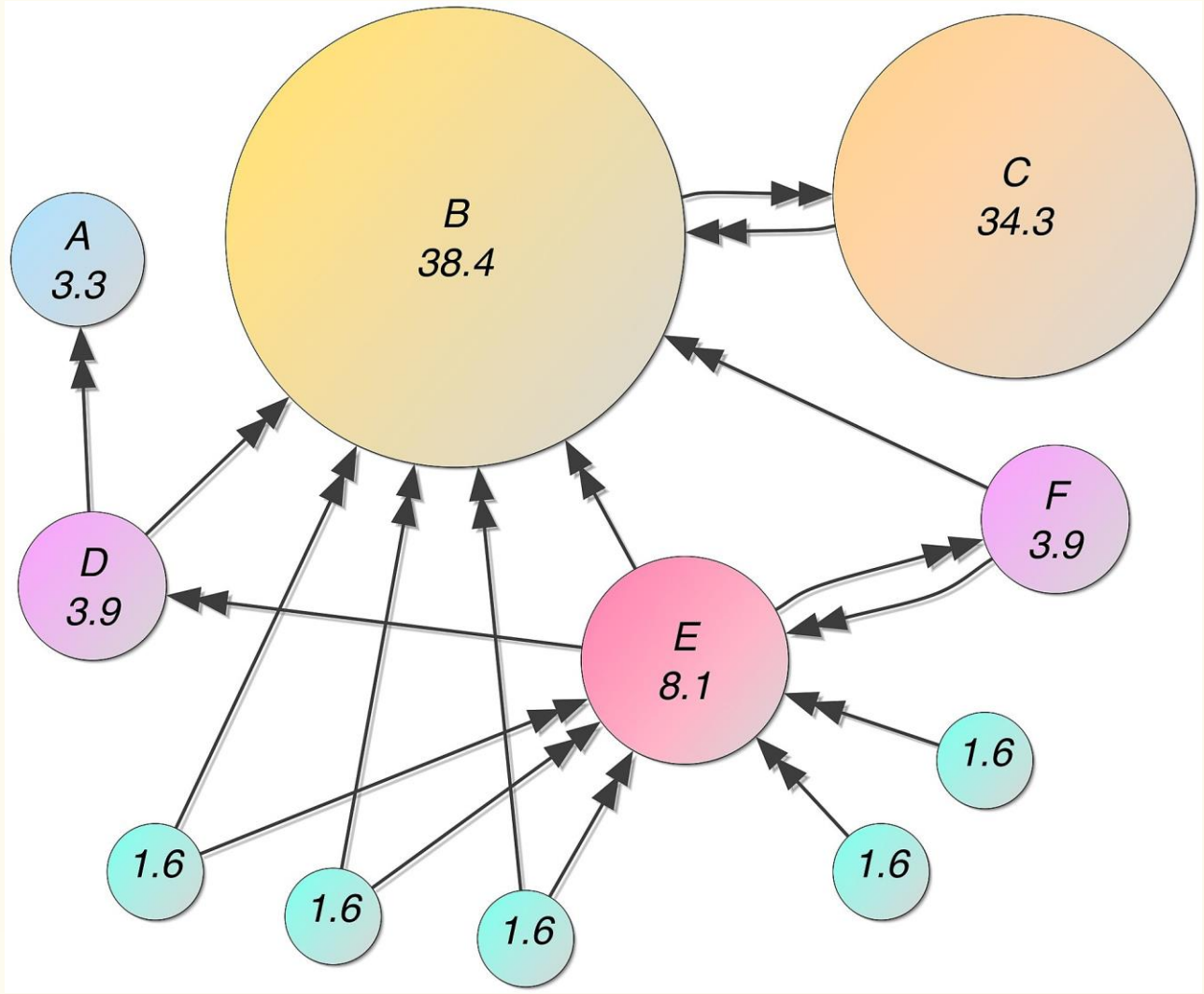
- How to handle dangling nodes (dead ends that don't link to anything) ?
- How to handle Rank sinks – group of pages that only link to each other ?

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 $\boldsymbol{M}_{ij} = \frac{1}{outdeg(i)}$ if there is a link from i to j.
2. Then modify as follows. 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 π_i.
5. Find all pages with decent textual match to search and then order those pages by PageRank!

PageRank - Example



It Gets More Complicated

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.

This slide was written a few years ago. Everything is changing now!

Still, PageRank is what launched Google!!!