INTRODUCTION TO MCMC AND PAGERANK

Eric Vigoda
Georgia Tech

Lecture for CS 6505
1 Markov Chain Basics

2 Ergodicity

3 What is the Stationary Distribution?

4 PageRank

5 Mixing Time

6 Preview of Further Topics
What is a Markov chain?

Example: Life in CS 6210, discrete time $t = 0, 1, 2, \ldots$:

- Each vertex is a state of the Markov chain.
- Directed graph, possibly with self-loops.
- Edge weights represent probability of a transition, so:
  - Non-negative
  - Sum of weights of outgoing edges = 1.
What is a Markov chain?

Example: Life in CS 6210, discrete time $t = 0, 1, 2, \ldots$:

Each vertex is a state of the Markov chain.

Directed graph, possibly with self-loops.

Edge weights represent probability of a transition, so:
non-negative and sum of weights of outgoing edges $= 1$. 
In general: \( N \) states \( \Omega = \{1, 2, \ldots, N\} \).

\( N \times N \) transition matrix \( P \) where:
\[
P(i, j) = \text{weight of edge } i \rightarrow j = \Pr \text{ (going from } i \text{ to } j)\]

For earlier example:

\[
P = \begin{bmatrix}
0.5 & 0.5 & 0 & 0 \\
0.2 & 0 & 0.5 & 0.3 \\
0 & 0.3 & 0.7 & 0 \\
0.7 & 0 & 0 & 0.3
\end{bmatrix}
\]

\( P \) is a stochastic matrix = rows sum to 1.
Time: $t = 0, 1, 2, \ldots$.

Let $X_t$ denote the state at time $t$.

$X_t$ is a random variable.

Process is memoryless—only current state matters, previous states do not matter.

Known as Markov property, hence the term Markov chain.
Time: $t = 0, 1, 2, \ldots$
Let $X_t$ denote the state at time $t$.
$X_t$ is a random variable.
For states $k$ and $j$, $\Pr(X_1 = j \mid X_0 = k) = P(k, j)$. 

Process is memoryless – only current state matters, previous states do not matter. 
Known as Markov property, hence the term Markov chain.
One-step transitions

Time: \( t = 0, 1, 2, \ldots \)

Let \( X_t \) denote the state at time \( t \).

\( X_t \) is a random variable.

For states \( k \) and \( j \), \( \Pr(X_1 = j \mid X_0 = k) = P(k, j) \).

In general, for \( t \geq 1 \), given:

- in state \( k_0 \) at time 0, in \( k_1 \) at time 1, \ldots, in \( k_{t-1} \) at time \( t - 1 \),

what’s the probability of being in state \( j \) at time \( t \)?
Time: \( t = 0, 1, 2, \ldots \)
Let \( X_t \) denote the state at time \( t \).
\( X_t \) is a random variable.
For states \( k \) and \( j \), \( \Pr(X_1 = j \mid X_0 = k) = P(k, j) \).
In general, for \( t \geq 1 \), given:
\begin{align*}
\text{in state } k_0 \text{ at time } 0, \text{ in } k_1 \text{ at time } 1, \ldots, \text{ in } k_{t-1} \text{ at time } t-1, \\
\text{what’s the probability of being in state } j \text{ at time } t? \\
\Pr(X_t = j \mid X_0 = k_0, X_1 = k_1, \ldots, X_{t-1} = k_{t-1}) \\
= \Pr(X_t = j \mid X_{t-1} = k_{t-1}) \\
= P(k_{t-1}, j).
\end{align*}
One-step transitions

Time: \( t = 0, 1, 2, \ldots \). Let \( X_t \) denote the state at time \( t \). \( X_t \) is a random variable.

For states \( k \) and \( j \), \( \Pr (X_{1} = j \mid X_{0} = k) = P(k, j) \).

In general, for \( t \geq 1 \), given:
- in state \( k_0 \) at time 0,
- in \( k_1 \) at time 1, \ldots,
- in \( k_{t-1} \) at time \( t - 1 \),
what’s the probability of being in state \( j \) at time \( t \)?

\[
\Pr (X_{t} = j \mid X_{0} = k_0, X_{1} = k_1, \ldots, X_{t-1} = k_{t-1}) = \Pr (X_{t} = j \mid X_{t-1} = k_{t-1}) = P(k_{t-1}, j).
\]

Process is memoryless – only current state matters, previous states do not matter. Known as Markov property, hence the term Markov chain.
What’s probability *Listen* at time 2 given *Email* at time 0?
Try all possibilities for state at time 1.
2-step transitions

What’s probability *Listen* at time 2 given *Email* at time 0? Try all possibilities for state at time 1.

\[
\Pr(X_2 = \text{Listen} \mid X_0 = \text{Email}) = \Pr(X_2 = \text{Listen} \mid X_1 = \text{Listen}) \times \Pr(X_1 = \text{Listen} \mid X_0 = \text{Email}) \\
+ \Pr(X_2 = \text{Listen} \mid X_1 = \text{Email}) \times \Pr(X_1 = \text{Email} \mid X_0 = \text{Email}) \\
+ \Pr(X_2 = \text{Listen} \mid X_1 = \text{StarCraft}) \times \Pr(X_1 = \text{StarCraft} \mid X_0 = \text{Email}) \\
+ \Pr(X_2 = \text{Listen} \mid X_1 = \text{Sleep}) \times \Pr(X_1 = \text{Sleep} \mid X_0 = \text{Email})
\]
What’s probability *Listen* at time 2 given *Email* at time 0?
Try all possibilities for state at time 1.

\[
\Pr(X_2 = \text{Listen} \mid X_0 = \text{Email})
\]
\[
= \Pr(X_2 = \text{Listen} \mid X_1 = \text{Listen}) \times \Pr(X_1 = \text{Listen} \mid X_0 = \text{Email})
+ \Pr(X_2 = \text{Listen} \mid X_1 = \text{Email}) \times \Pr(X_1 = \text{Email} \mid X_0 = \text{Email})
+ \Pr(X_2 = \text{Listen} \mid X_1 = \text{StarCraft}) \times \Pr(X_1 = \text{StarCraft} \mid X_0 = \text{Email})
+ \Pr(X_2 = \text{Listen} \mid X_1 = \text{Sleep}) \times \Pr(X_1 = \text{Sleep} \mid X_0 = \text{Email})
\]
\[
= (.5)(.2) + 0 + 0 + (.7)(.3) = .31
\]

\[
P = \begin{bmatrix}
.5 & .5 & 0 & 0 \\
.2 & 0 & .5 & .3 \\
0 & .3 & .7 & 0 \\
.7 & 0 & 0 & .3
\end{bmatrix}
\]

\[
P^2 = \begin{bmatrix}
.35 & .25 & .25 & .15 \\
.31 & .25 & .35 & .09 \\
.06 & .21 & .64 & .09 \\
.56 & .35 & 0 & .09
\end{bmatrix}
\]

States: 1=Listen, 2=Email, 3=StarCraft, 4=Sleep.
2-step transition probabilities: use $P^2$.

In general, for states $i$ and $j$:

$$
\Pr(X_{t+2} = j \mid X_t = i)
= \sum_{k=1}^{N} \Pr(X_{t+2} = j \mid X_{t+1} = k) \times \Pr(X_{t+1} = k \mid X_t = i)
= \sum_k P(k, j)P(i, k) = \sum_k P(i, k)P(k, j) = P^2(i,j)
$$
2-step transition probabilities: use $P^2$.
In general, for states $i$ and $j$:

$$
\Pr (X_{t+2} = j \mid X_t = i) = \sum_{k=1}^{N} \Pr (X_{t+2} = j \mid X_{t+1} = k) \times \Pr (X_{t+1} = k \mid X_t = i) = \sum_{k} P(k,j)P(i,k) = \sum_{k} P(i,k)P(k,j) = P^2(i,j)
$$

$\ell$-step transition probabilities: use $P^\ell$.
For states $i$ and $j$ and integer $\ell \geq 1$,

$$
\Pr (X_{t+\ell} = j \mid X_t = i) = P^\ell(i,j),
$$
Suppose the state at time 0 is not fixed but is chosen from a probability distribution $\mu_0$. Notation: $X_0 \sim \mu_0$.

What is the distribution for $X_1$?
Suppose the state at time 0 is not fixed but is chosen from a probability distribution $\mu_0$. Notation: $X_0 \sim \mu_0$.

What is the distribution for $X_1$?

For state $j$,

$$\Pr(X_1 = j) = \sum_{i=1}^{N} \Pr(X_0 = i) \times \Pr(X_1 = j \mid X_0 = i)$$

$$= \sum_{i} \mu_0(i) P(i, j) = (\mu_0 P)(j)$$

So $X_1 \sim \mu_1$ where $\mu_1 = \mu_0 P$. 
Random Initial State

Suppose the state at time 0 is not fixed but is chosen from a probability distribution \( \mu_0 \).

Notation: \( X_0 \sim \mu_0 \).

What is the distribution for \( X_1 \)?

For state \( j \),

\[
\Pr(X_1 = j) = \sum_{i=1}^{N} \Pr(X_0 = i) \times \Pr(X_1 = j \mid X_0 = i)
\]

\[
= \sum_{i} \mu_0(i)P(i, j) = (\mu_0P)(j)
\]

So \( X_1 \sim \mu_1 \) where \( \mu_1 = \mu_0P \).

And \( X_t \sim \mu_t \) where \( \mu_t = \mu_0P^t \).
Let's look again at our CS 6210 example:

\[
P = \begin{bmatrix}
.5 & .5 & 0 & 0 \\
.2 & 0 & .5 & .3 \\
0 & .3 & .7 & 0 \\
.7 & 0 & 0 & .3 \\
\end{bmatrix}
\]
Let's look again at our CS 6210 example:

\[ P = \begin{bmatrix}
  .5 & .5 & 0 & 0 \\
  .2 & 0 & .5 & .3 \\
  0 & .3 & .7 & 0 \\
  .7 & 0 & 0 & .3
\end{bmatrix} \]

\[ P^2 = \begin{bmatrix}
  .35 & .25 & .25 & .15 \\
  .31 & .25 & .35 & .09 \\
  .06 & .21 & .64 & .09 \\
  .56 & .35 & 0 & .09
\end{bmatrix} \]
Let's look again at our CS 6210 example:

\[
P = \begin{bmatrix}
0.5 & 0.5 & 0 & 0 \\
0.2 & 0 & 0.5 & 0.3 \\
0 & 0.3 & 0.7 & 0 \\
0.7 & 0 & 0 & 0.3
\end{bmatrix}
\]

\[
P^2 = \begin{bmatrix}
0.35 & 0.25 & 0.25 & 0.15 \\
0.31 & 0.25 & 0.35 & 0.09 \\
0.06 & 0.21 & 0.64 & 0.09 \\
0.56 & 0.35 & 0 & 0.09
\end{bmatrix}
\]

\[
P^{10} = \begin{bmatrix}
0.247770 & 0.244781 & 0.402267 & 0.105181 \\
0.245167 & 0.244349 & 0.405688 & 0.104796 \\
0.239532 & 0.243413 & 0.413093 & 0.103963 \\
0.251635 & 0.245423 & 0.397189 & 0.105754
\end{bmatrix}
\]

Columns are converging to \( \pi = [0.244186, 0.244186, 0.406977, 0.104651] \).
Let's look again at our CS 6210 example:

\[
P = \begin{bmatrix}
.5 & .5 & 0 & 0 \\
.2 & 0 & .5 & .3 \\
0 & .3 & .7 & 0 \\
.7 & 0 & 0 & .3 \\
\end{bmatrix}
\]

\[
P^2 = \begin{bmatrix}
.35 & .25 & .25 & .15 \\
.31 & .25 & .35 & .09 \\
.06 & .21 & .64 & .09 \\
.56 & .35 & 0 & .09 \\
\end{bmatrix}
\]

\[
P^{10} = \begin{bmatrix}
.247770 & .244781 & .402267 & .105181 \\
.245167 & .244349 & .405688 & .104796 \\
.239532 & .243413 & .413093 & .103963 \\
.251635 & .245423 & .397189 & .105754 \\
\end{bmatrix}
\]

\[
P^{20} = \begin{bmatrix}
.244190 & .244187 & .406971 & .104652 \\
.244187 & .244186 & .406975 & .104651 \\
.244181 & .244185 & .406984 & .104650 \\
.244195 & .244188 & .406966 & .104652 \\
\end{bmatrix}
\]

Columns are converging to \( \pi = [ .244186, .244186, .406977, .104651 ] \).
Let's look again at our CS 6210 example:

\[
P = \begin{bmatrix}
.5 & .5 & 0 & 0 \\
.2 & 0 & .5 & .3 \\
0 & .3 & .7 & 0 \\
.7 & 0 & 0 & .3 \\
\end{bmatrix}
\]

\[
P^2 = \begin{bmatrix}
.35 & .25 & .25 & .15 \\
.31 & .25 & .35 & .09 \\
.06 & .21 & .64 & .09 \\
.56 & .35 & 0 & .09 \\
\end{bmatrix}
\]

\[
P^{10} = \begin{bmatrix}
.247770 & .244781 & .402267 & .105181 \\
.245167 & .244349 & .405688 & .104796 \\
.239532 & .243413 & .413093 & .103963 \\
.251635 & .245423 & .397189 & .105754 \\
\end{bmatrix}
\]

\[
P^{20} = \begin{bmatrix}
.244190 & .244187 & .406971 & .104652 \\
.244187 & .244186 & .406975 & .104651 \\
.244181 & .244185 & .406984 & .104650 \\
.244195 & .244188 & .406966 & .104652 \\
\end{bmatrix}
\]

Columns are converging to

\[
\pi = [ .244186, .244186, .406977, .104651 ]
\]
For big $t$, 

$$P^t \approx \begin{bmatrix} .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \end{bmatrix}$$

Regardless of where it starts $X_0$, for big $t$:

$$\Pr(X_t = 1) = .244186$$
$$\Pr(X_t = 2) = .244186$$
$$\Pr(X_t = 3) = .406977$$
$$\Pr(X_t = 4) = .104651$$

Let $\pi = [.244186, .244186, .406977, .104651]$. In other words, for big $t$, $X_t \sim \pi$. $\pi$ is called a stationary distribution.
For big $t$,\n\[
\mathbf{P}^t \approx \begin{bmatrix}
0.244186 & 0.244186 & 0.406977 & 0.104651 \\
0.244186 & 0.244186 & 0.406977 & 0.104651 \\
0.244186 & 0.244186 & 0.406977 & 0.104651 \\
0.244186 & 0.244186 & 0.406977 & 0.104651 \\
\end{bmatrix}
\]

Regardless of where it starts $X_0$, for big $t$:

\[
\Pr(X_t = 1) = 0.244186 \\
\Pr(X_t = 2) = 0.244186 \\
\Pr(X_t = 3) = 0.406977 \\
\Pr(X_t = 4) = 0.104651
\]
For big $t$, 

$$P^t \approx \begin{bmatrix} .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \\ .244186 & .244186 & .406977 & .104651 \end{bmatrix}$$

Regardless of where it starts $X_0$, for big $t$: 

$$\Pr (X_t = 1) = .244186$$
$$\Pr (X_t = 2) = .244186$$
$$\Pr (X_t = 3) = .406977$$
$$\Pr (X_t = 4) = .104651$$

Let $\pi = [ .244186, .244186, .406977, .104651 ]$. 

In other words, for big $t$, $X_t \sim \pi$.

$\pi$ is called a *stationary distribution*. 
Let $\pi = [0.244186, 0.244186, 0.406977, 0.104651]$. $\pi$ is called a *stationary distribution*.
Let $\pi = [0.244186, 0.244186, 0.406977, 0.104651]$. 
$\pi$ is called a *stationary distribution*.

Once we reach $\pi$ we stay in $\pi$: if $X_t \sim \pi$ then $X_{t+1} \sim \pi$, in other words, $\pi P = \pi$. 
Let \( \pi = [0.244186, 0.244186, 0.406977, 0.104651] \). 
\( \pi \) is called a stationary distribution.

Once we reach \( \pi \) we stay in \( \pi \): if \( X_t \sim \pi \) then \( X_{t+1} \sim \pi \), in other words, \( \pi P = \pi \).

Any distribution \( \pi \) where \( \pi P = \pi \) is called a stationary distribution of the Markov chain.
Key questions:
- When is there a stationary distribution?
- If there is at least one, is it unique or more than one?
- Assuming there’s a unique stationary distribution:
  - Do we always reach it?
  - What is it?
  - *Mixing time* = Time to reach unique stationary distribution

Algorithmic Goal:
- If we have a distribution $\pi$ that we want to sample from, can we design a Markov chain that has:
  - Unique stationary distribution $\pi$,
  - From every $X_0$ we always reach $\pi$,
  - Fast mixing time.
1. Markov Chain Basics
2. Ergodicity
3. What is the Stationary Distribution?
4. PageRank
5. Mixing Time
6. Preview of Further Topics
Want a unique stationary distribution $\pi$ and that get to it from every starting state $X_0$. But if multiple strongly connected components (SCCs) then can't go from one to the other:

Starting at 1 gets to different distribution than starting at 5. State $i$ communicates with state $j$ if starting at $i$ can reach $j$: there exists $t$, $P^t(i, j) > 0$. Markov chain is irreducible if all pairs of states communicate.
Want a unique stationary distribution $\pi$ and that get to it from every starting state $X_0$. But if multiple strongly connected components (SCCs) then can’t go from one to the other:

Starting at 1 gets to different distribution than starting at 5.
Want a unique stationary distribution $\pi$ and that get to it from every starting state $X_0$. But if multiple strongly connected components (SCCs) then can’t go from one to the other:

Starting at 1 gets to different distribution than starting at 5.

State $i$ communicates with state $j$ if starting at $i$ can reach $j$:

there exists $t$, $P^t(i,j) > 0$.

Markov chain is irreducible if all pairs of states communicate.
Example of bipartite Markov chain:

Starting at 1 gets to different distribution than starting at 3.
Example of bipartite Markov chain:

Starting at 1 gets to different distribution than starting at 3.

Need that no periodicity.
Return times for state $i$ are times $R_i = \{ t : P^t(i, i) > 0 \}$.

Above example: $R_1 = \{3, 5, 6, 8, 9, \ldots \}$.

Let $r = \gcd(R_i)$ be the period for state $i$. 
Return times for state $i$ are times $R_i = \{ t : P^t(i, i) > 0 \}$. Above example: $R_1 = \{3, 5, 6, 8, 9, \ldots \}$.

Let $r = \gcd(R_i)$ be the period for state $i$.

If $P$ is irreducible then all states have the same period. If $r = 2$ then the Markov chain is bipartite. 

A Markov chain is aperiodic if $r = 1$. 

\[ \text{Aperiodic} \]
Ergodic = Irreducible and aperiodic.
Ergodic = Irreducible and aperiodic.

*Fundamental Theorem for Markov Chains:*
Ergodic Markov chain has a unique stationary distribution $\pi$.
And for all initial $X_0 \sim \mu_0$ then:

$$\lim_{t \to \infty} \mu_t = \pi.$$

In other words, for big enough $t$, all rows of $P^t$ are $\pi$. 
Ergodic = Irreducible and aperiodic.

**Fundamental Theorem for Markov Chains:**
Ergodic Markov chain has a unique stationary distribution $\pi$. And for all initial $X_0 \sim \mu_0$ then:

$$\lim_{t \to \infty} \mu_t = \pi.$$ 

In other words, for big enough $t$, all rows of $P^t$ are $\pi$.

How big does $t$ need to be?

What is $\pi$?
What is a $\pi$?
Proof idea: Ergodic MC has Unique Stationary Distribution

What is a $\pi$?

Fix a state $i$ and set $X_0 = i$. Let $T$ be the first time we visit state $i$ again. $T$ is a random variable.

For every state $j$,

let $\rho(j) =$ expected number of visits to $j$ up to time $T$.

(Note, $\rho(i) = 1$.)

Let $\pi(j) = \rho(j)/Z$ where $Z = \sum_k \rho(k)$.

Can verify that this $\pi$ is a stationary distribution.
What is a $\pi$?

Fix a state $i$ and set $X_0 = i$.
Let $T$ be the first time we visit state $i$ again.
$T$ is a random variable.
For every state $j$,

let $\rho(j) =$ expected number of visits to $j$ up to time $T$.
(Note, $\rho(i) = 1$.)

Let $\pi(j) = \rho(j)/Z$ where $Z = \sum_k \rho(k)$.
Can verify that this $\pi$ is a stationary distribution.

Why is it unique and we always reach it?
What is a $\pi$?

Fix a state $i$ and set $X_0 = i$.
Let $T$ be the first time we visit state $i$ again.
$T$ is a random variable.
For every state $j$,
let $\rho(j) =$ expected number of visits to $j$ up to time $T$.
(Note, $\rho(i) = 1$.)

Let $\pi(j) = \rho(j)/Z$ where $Z = \sum_k \rho(k)$.
Can verify that this $\pi$ is a stationary distribution.

Why is it unique and we always reach it?
Make 2 chains $(X_t)$ and $(Y_t)$:
- $X_0$ is arbitrary, and
- $Y_0$ is chosen from $\pi$ so that $Y_t \sim \pi$ for all $t$.

Using irreducibility, can “couple” the transitions of these chains:
for big $t$ we have $X_t = Y_t$ and thus $X_t \sim \pi$. 
1 Markov Chain Basics

2 Ergodicity

3 What is the Stationary Distribution?

4 PageRank

5 Mixing Time

6 Preview of Further Topics
Symmetric if for all pairs $i,j$: $P(i,j) = P(j,i)$.

Then $\pi$ is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$
Symmetric if for all pairs $i, j$: $P(i, j) = P(j, i)$. Then $\pi$ is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$ 

**Proof:** We’ll verify that $\pi P = \pi$ for this $\pi$. Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$. 
Symmetric if for all pairs \( i, j \):  \( P(i, j) = P(j, i) \).

Then \( \pi \) is uniformly distributed over all of the states \( \{1, \ldots, N\} \):

\[
\pi(j) = \frac{1}{N} \quad \text{for all states } j.
\]

**Proof:** We’ll verify that \( \pi P = \pi \) for this \( \pi \).

Need to check that for all states \( j \): \( (\pi P)(j) = \pi(j) \).

\[
(\pi P)(j) = \sum_{i=1}^{N} \pi(i)P(i,j)
\]
Symmetric if for all pairs $i, j$: $P(i, j) = P(j, i)$.

Then $\pi$ is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$ 

**Proof:** We’ll verify that $\pi P = \pi$ for this $\pi$.

Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$(\pi P)(j) = \sum_{i=1}^{N} \pi(i) P(i, j)$$

$$= \frac{1}{N} \sum_{i=1}^{N} P(i, j)$$
**Symmetric** if for all pairs $i, j$: $P(i, j) = P(j, i)$.

Then $\pi$ is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$

**Proof:** We’ll verify that $\pi P = \pi$ for this $\pi$.

Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$(\pi P)(j) = \sum_{i=1}^{N} \pi(i) P(i, j)$$

$$= \frac{1}{N} \sum_{i=1}^{N} P(i, j)$$

$$= \frac{1}{N} \sum_{i=1}^{N} P(j, i) \text{ since } P \text{ is symmetric}$$
Symmetric if for all pairs $i, j$: $P(i, j) = P(j, i)$.

Then $\pi$ is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \text{ for all states } j.$$ 

**Proof:** We'll verify that $\pi P = \pi$ for this $\pi$.

Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$(\pi P)(j) = \sum_{i=1}^{N} \pi(i)P(i, j)$$

$$= \frac{1}{N} \sum_{i=1}^{N} P(i, j)$$

$$= \frac{1}{N} \sum_{i=1}^{N} P(j, i) \text{ since } P \text{ is symmetric}$$

$$= \frac{1}{N} \text{ since rows of } P \text{ always sum to 1}$$
Symmetric if for all pairs $i, j$: $P(i, j) = P(j, i)$.

Then $\pi$ is uniformly distributed over all of the states $\{1, \ldots, N\}$:

$$\pi(j) = \frac{1}{N} \quad \text{for all states } j.$$ 

**Proof:** We’ll verify that $\pi P = \pi$ for this $\pi$.

Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$
(\pi P)(j) = \sum_{i=1}^{N} \pi(i) P(i, j) \\
= \frac{1}{N} \sum_{i=1}^{N} P(i, j) \\
= \frac{1}{N} \sum_{i=1}^{N} P(j, i) \quad \text{since } P \text{ is symmetric} \\
= \frac{1}{N} \quad \text{since rows of } P \text{ always sum to 1} \\
= \pi(j)
$$
Reversible with respect to $\pi$ if for all pairs $i,j$:

$$\pi(i)P(i,j) = \pi(j)P(j,i).$$

If can find such a $\pi$ then it is the stationary distribution.
Reversible with respect to $\pi$ if for all pairs $i,j$:

$$\pi(i)P(i,j) = \pi(j)P(j,i).$$

If can find such a $\pi$ then it is the stationary distribution.

Proof: Similar to the symmetric case.
Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$. 
Reversible with respect to $\pi$ if for all pairs $i,j$:

$$\pi(i)P(i,j) = \pi(j)P(j,i).$$

If can find such a $\pi$ then it is the stationary distribution.

Proof: Similar to the symmetric case. Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$(\pi P)(j) = \sum_{i=1}^{N} \pi(i)P(i,j)$$
Reversible with respect to $\pi$ if for all pairs $i, j$:

$$\pi(i)P(i, j) = \pi(j)P(j, i).$$

If can find such a $\pi$ then it is the stationary distribution.

**Proof:** Similar to the symmetric case. Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$
\begin{align*}
(\pi P)(j) &= \sum_{i=1}^{N} \pi(i)P(i, j) \\
&= \sum_{i=1}^{N} \pi(j)P(j, i) \quad \text{since } P \text{ is reversible}
\end{align*}
$$
Reversible with respect to $\pi$ if for all pairs $i, j$:

$$\pi(i)P(i, j) = \pi(j)P(j, i).$$

If can find such a $\pi$ then it is the stationary distribution.

**Proof:** Similar to the symmetric case.

Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$(\pi P)(j) = \sum_{i=1}^{N} \pi(i)P(i, j)$$

$$= \sum_{i=1}^{N} \pi(j)P(j, i) \text{ since } P \text{ is reversible}$$

$$= \pi(j) \sum_{i=1}^{N} P(j, i)$$
Determining $\pi$: Reversible Markov Chain

Reversible with respect to $\pi$ if for all pairs $i, j$:

$$\pi(i)P(i, j) = \pi(j)P(j, i).$$

If can find such a $\pi$ then it is the stationary distribution.

**Proof:** Similar to the symmetric case.

Need to check that for all states $j$: $(\pi P)(j) = \pi(j)$.

$$
(\pi P)(j) = \sum_{i=1}^{N} \pi(i)P(i, j)
= \sum_{i=1}^{N} \pi(j)P(j, i) \quad \text{since } P \text{ is reversible}
= \pi(j) \sum_{i=1}^{N} P(j, i)
= \pi(j).
$$
Random walk on a $d$-regular, connected undirected graph $G$:
What is $\pi$?
Random walk on a $d$-regular, connected undirected graph $G$:

What is $\pi$?

Symmetric: for edge $(i,j)$, $P(i,j) = P(j,i) = 1/d$.
So $\pi$ is uniform: $\pi(i) = 1/n$. 
Random walk on a \( d \)-regular, connected undirected graph \( G \):

What is \( \pi \)?

Symmetric: for edge \((i, j)\), \( P(i, j) = P(j, i) = 1/d \).
So \( \pi \) is uniform: \( \pi(i) = 1/n \).

Random walk on a general connected undirected graph \( G \):

What is \( \pi \)?
Random walk on a $d$-regular, connected undirected graph $G$:

What is $\pi$?

Symmetric: for edge $(i,j)$, $P(i,j) = P(j,i) = 1/d$.

So $\pi$ is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph $G$:

What is $\pi$?

Consider $\pi(i) = d(i)/Z$ where

$d(i) = \text{degree of vertex } i$ and

$Z = \sum_{j \in V} d(j)$. (Note, $Z = 2m = 2|E|$.)

Check it’s reversible: $\pi(i)P(i,j) = \frac{d(i)}{Z} \frac{1}{d(i)} = \frac{1}{Z} = \pi(j)P(j,i)$. 

Random walk on a $d$-regular, connected undirected graph $G$:

What is $\pi$?

Symmetric: for edge $(i,j)$, $P(i,j) = P(j,i) = 1/d$.
So $\pi$ is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph $G$:

What is $\pi$?

Consider $\pi(i) = d(i)/Z$ where

$d(i) =$ degree of vertex $i$ and

$Z = \sum_{j \in V} d(j)$. (Note, $Z = 2m = 2|E|$.)

Check it's reversible: $\pi(i)P(i,j) = \frac{d(i)}{Z} \frac{1}{d(i)} = \frac{1}{Z} = \pi(j)P(j,i)$.

What if $G$ is a directed graph?
Random walk on a $d$-regular, connected undirected graph $G$:

What is $\pi$?
Symmetric: for edge $(i,j)$, $P(i,j) = P(j,i) = 1/d$.
So $\pi$ is uniform: $\pi(i) = 1/n$.

Random walk on a general connected undirected graph $G$:

What is $\pi$?
Consider $\pi(i) = d(i)/Z$ where
$d(i) =$ degree of vertex $i$ and
$Z = \sum_{j \in V} d(j)$. (Note, $Z = 2m = 2|E|$.)

Check it’s reversible: $\pi(i)P(i,j) = \frac{d(i)}{Z} \frac{1}{d(i)} = \frac{1}{Z} = \pi(j)P(j,i)$.

What if $G$ is a directed graph?

Then it may not be reversible, and if it’s not reversible:
then usually we can’t figure out the stationary distribution since typically $N$ is HUGE.
1. **Markov Chain Basics**

2. **Ergodicity**

3. **What is the Stationary Distribution?**

4. **PageRank**

5. **Mixing Time**

6. **Preview of Further Topics**
PageRank is an algorithm devised by Brin and Page 1998: determine the “importance” of webpages.
PageRank is an algorithm devised by Brin and Page 1998: determine the “importance” of webpages.

Webgraph:
- \( V \) = webpages
- \( E \) = directed edges for hyperlinks

Let \( \pi(x) \) = “rank” of page \( x \).
We are trying to define \( \pi(x) \) in a sensible way.
PageRank is an algorithm devised by Brin and Page 1998: determine the “importance” of webpages.

Webgraph:
- \( V = \text{webpages} \)
- \( E = \text{directed edges for hyperlinks} \)

Notation:
For page \( x \in V \), let:

\[
\begin{align*}
\text{Out}(x) &= \{ y : x \rightarrow y \in E \} = \text{outgoing edges from } x \\
\text{In}(x) &= \{ w : w \rightarrow x \in E \} = \text{incoming edges to } x
\end{align*}
\]

Let \( \pi(x) = \text{“rank” of page } x \).
We are trying to define \( \pi(x) \) in a sensible way.
First idea for ranking pages: like academic papers

use citation counts

Here, citation = link to a page.

So set $\pi(x) = |\text{In}(x)| = \text{number of links to } x$. 
What if:
   a webpage has 500 links and one is to Eric’s page.
   another webpage has only 5 links and one is to Santosh’s page.

Which link is more valuable?
Refining the Ranking Idea

What if:
   a webpage has 500 links and one is to Eric’s page.
   another webpage has only 5 links and one is to Santosh’s page.

Which link is more valuable?

*Academic papers*: If a paper cites 50 other papers, then each reference gets $1/50$ of a citation.
What if:
   a webpage has 500 links and one is to Eric’s page.
   another webpage has only 5 links and one is to Santosh’s page.

Which link is more valuable?

*Academic papers*: If a paper cites 50 other papers, then each reference gets $1/50$ of a citation.

*Webpages*: If a page $y$ has $|\text{Out}(y)|$ outgoing links, then:
   each linked page gets $1/|\text{Out}(y)|$.

New solution:

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{1}{|\text{Out}(y)|}.$$
Further Refining the Ranking Idea

Previous:

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{1}{|\text{Out}(y)|}.$$ 

But if *Eric’s children’s webpage* has a link to a Eric’s page and *CNN* has a link to Santosh’s page, which is more important?
Further Refining the Ranking Idea

Previous:

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{1}{|\text{Out}(y)|}.$$ 

But if Eric’s children’s webpage has a link to a Eric’s page and CNN has a link to Santosh’s page, which is more important?

Solution: define $\pi(x)$ recursively.

Page $y$ has importance $\pi(y)$.

A link from $y$ gets $\pi(y)/|\text{Out}(y)|$ of a citation.

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{\pi(y)}{|\text{Out}(y)|}.$$
Importance of page $x$:

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{\pi(y)}{|\text{Out}(y)|}.$$
Importance of page $x$:

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{\pi(y)}{|\text{Out}(y)|}.$$  

Recursive definition of $\pi$, how do we find it?
Importance of page $x$:

$$\pi(x) = \sum_{y \in \text{In}(x)} \frac{\pi(y)}{|\text{Out}(y)|}.$$ 

Recursive definition of $\pi$, how do we find it?

Look at the random walk on the webgraph $G = (V, E)$. From a page $y \in V$, choose a random link and follow it. This is a Markov chain.

For $y \to x \in E$ then:

$$P(y, x) = \frac{1}{|\text{Out}(y)|}$$

What is the stationary distribution of this Markov chain?
Random walk on the webgraph $G = (V, E)$.
For $y \rightarrow x \in E$ then:

$$P(y, x) = \frac{1}{|\text{Out}(y)|}$$

What is the stationary distribution of this Markov chain?
Random walk on the webgraph $G = (V, E)$. For $y \to x \in E$ then:

$$P(y, x) = \frac{1}{|\text{Out}(y)|}$$

What is the stationary distribution of this Markov chain? Need to find $\pi$ where $\pi = \pi P$. Thus,

$$\pi(x) = \sum_{y \in V} \pi(y) P(y, x) = \sum_{y \in \text{In}(x)} \frac{\pi(y)}{|\text{Out}(y)|}.$$ 

This is identical to the definition of the importance vector $\pi$.

**Summary:** the stationary distribution of the random walk on the webgraph gives the importance $\pi(x)$ of a page $x$. 
Random walk on the webgraph $G = (V, E)$.

Is $\pi$ the \textbf{only} stationary distribution?
In other words, is the Markov chain \textit{ergodic}?
Random walk on the webgraph $G = (V, E)$.

Is $\pi$ the only stationary distribution? In other words, is the Markov chain ergodic?

Need that $G$ is strongly connected – it probably is not. And some pages have no outgoing links...

then hit the “random” button!
Random walk on the webgraph $G = (V, E)$.

Is $\pi$ the only stationary distribution? In other words, is the Markov chain ergodic?

Need that $G$ is strongly connected – it probably is not. And some pages have no outgoing links... then hit the “random” button!

**Solution to make it ergodic:**
Introduce “damping factor” $\alpha$ where $0 < \alpha \leq 1$.
(in practice apparently use $\alpha \approx .85$)

From page $y$, 
with prob. $\alpha$ follow a random outgoing link from page $y$. 
with prob. $1 - \alpha$ go to a completely random page (uniformly chosen from all pages $V$).
Let $N = |V|$ denote number of webpages.

Transition matrix of new Random Surfer chain:

$$P(y, x) = \begin{cases} 
\frac{1-\alpha}{N} & \text{if } y \to x \not\in E \\
\frac{1-\alpha}{N} + \frac{\alpha}{|\text{Out}(y)|} & \text{if } y \to x \in E
\end{cases}$$

This new Random Surfer Markov chain is ergodic. Thus, unique stationary distribution is the desired $\pi$. 

How to find $\pi$?

Take last week's $\pi$, and compute $\pi P_t$ for big $t$. 

What's a big enough $t$?
Let $N = |V|$ denote number of webpages.

Transition matrix of new Random Surfer chain:

$$P(y, x) = \begin{cases} 
\frac{1-\alpha}{N} & \text{if } y \rightarrow x \notin E \\
\frac{1-\alpha}{N} + \frac{\alpha}{|Out(y)|} & \text{if } y \rightarrow x \in E 
\end{cases}$$

This new Random Surfer Markov chain is ergodic. Thus, unique stationary distribution is the desired $\pi$.

How to find $\pi$?

Take last week’s $\pi$, and compute $\pi P^t$ for big $t$.

What’s a big enough $t$?
1. Markov Chain Basics
2. Ergodicity
3. What is the Stationary Distribution?
4. PageRank
5. Mixing Time
6. Preview of Further Topics
How fast does an ergodic MC reach its unique stationary $\pi$?
How fast does an ergodic MC reach its unique stationary $\pi$?

Need to measure distance from $\pi$, use total variation distance. For distributions $\mu$ and $\nu$ on set $\Omega$:

$$d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.$$
How fast does an ergodic MC reach its unique stationary \( \pi \)?

Need to measure distance from \( \pi \), use **total variation distance**.

For distributions \( \mu \) and \( \nu \) on set \( \Omega \):

\[
d_{TV}(\mu, \nu) = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)|.
\]

*Example*: \( \Omega = \{1, 2, 3, 4\} \).

\( \mu \) is uniform: \( \mu(1) = \mu(2) = \mu(3) = \mu(4) = .25 \).

And \( \nu \) has: \( \nu(1) = .5, \nu(2) = .1, \nu(3) = .15, \nu(4) = .25 \).

\[
d_{TV}(\mu, \nu) = \frac{1}{2} (.25 + .15 + .1 + 0) = .25
\]
Consider ergodic MC with states $\Omega$, transition matrix $P$, and unique stationary distribution $\pi$.
For state $x \in \Omega$, time to mix from $x$:

$$T(x) = \min\{ t : d_{TV}(P^t(x, \cdot), \pi) \leq 1/4 \}.$$
Consider ergodic MC with states $\Omega$, transition matrix $P$, and unique stationary distribution $\pi$.
For state $x \in \Omega$, time to mix from $x$:

$$T(x) = \min\{t : d_{TV}(P^t(x, \cdot), \pi) \leq 1/4\}.$$

Then, mixing time $T_{mix} = \max_x T(x)$.

**Summarizing in words:** mixing time is time to get within distance $\leq 1/4$ of $\pi$ from the worst initial state $X_0$. 
Consider ergodic MC with states $Ω$, transition matrix $P$, and unique stationary distribution $π$.
For state $x ∈ Ω$, time to mix from $x$:

$$T(x) = \min\{t : d_{TV}(P^t(x, \cdot), \pi) \leq 1/4\}.$$ 

Then, mixing time $T_{mix} = \max_x T(x)$.

**Summarizing in words:**
mixing time is time to get within distance $\leq 1/4$ of $π$ from the worst initial state $X_0$.

Choice of constant $1/4$ is somewhat arbitrary.
Can get within distance $\leq \epsilon$ in time $O(T_{mix} \log(1/\epsilon))$. 


Coupling proof:
Consider 2 copies of the Random Surfer chain \((X_t)\) and \((Y_t)\).

Choose \(Y_0\) from \(\pi\). Thus, \(Y_t \sim \pi\) for all \(t\).
And \(X_0\) is arbitrary.
Coupling proof:
Consider 2 copies of the Random Surfer chain \((X_t)\) and \((Y_t)\).

Choose \(Y_0\) from \(\pi\). Thus, \(Y_t \sim \pi\) for all \(t\).
And \(X_0\) is arbitrary.

If \(X_{t-1} = Y_{t-1}\) then they choose the same transition at time \(t\).
If \(X_{t-1} \neq Y_{t-1}\) then with prob. \(1 - \alpha\) choose the same random page \(z\) for both chains.

Therefore,

\[
\Pr(X_t \neq Y_t) \leq \alpha^t.
\]
Coupling proof:
Consider 2 copies of the Random Surfer chain \((X_t)\) and \((Y_t)\).

Choose \(Y_0\) from \(\pi\). Thus, \(Y_t \sim \pi\) for all \(t\).
And \(X_0\) is arbitrary.

If \(X_{t-1} = Y_{t-1}\) then they choose the same transition at time \(t\).
If \(X_{t-1} \neq Y_{t-1}\) then with prob. \(1 - \alpha\) choose the same random page \(z\) for both chains.

Therefore,
\[
\Pr(X_t \neq Y_t) \leq \alpha^t.
\]

Setting: \(t \geq -2 / \log(\alpha)\) we have \(\Pr(X_t \neq Y_t) \leq 1/4\).
Therefore, mixing time:
\[
T_{\text{mix}} \leq \frac{-2}{\log \alpha} \approx 8.5 \text{ for } \alpha = .85.
\]
1. **Markov Chain Basics**

2. **Ergodicity**

3. **What is the Stationary Distribution?**

4. **PageRank**

5. **Mixing Time**

6. **Preview of Further Topics**
Undirected graph:

Matching = subset of vertex disjoint edges.

Let $\Omega = \text{collection of all matchings of } G \text{ (of all sizes)}$. 
Example Chain: Random Matching

Undirected graph:

Matching = subset of vertex disjoint edges.

Let $\Omega =$ collection of all matchings of $G$ (of all sizes).

Can we generate a matching uniformly at random from $\Omega$?

in time polynomial in $n = |V|$?
Consider an undirected graph $G = (V, E)$.

From a matching $X_t$ the transition $X_t \rightarrow X_{t+1}$ is defined as follows:

1. Choose an edge $e = (v, w)$ uniformly at random from $E$.
2. If $e \in X_t$ then set $X_{t+1} = X_t \setminus \{e\}$.
3. If $v$ and $w$ are unmatched in $X_t$ then set $X_{t+1} = X_t \cup \{e\}$.
4. Otherwise, set $X_{t+1} = X_t$.

Symmetric and ergodic and thus $\pi$ is uniform over $\Omega$. How fast does it reach $\pi$?

Further topic (in MCMC class): we'll see that it's close to $\pi$ after poly$(n)$ steps and this holds for every $G$. Thus, we can generate a random matching of a graph in polynomial-time.
Markov Chain for Matchings

Consider an undirected graph $G = (V, E)$.

From a matching $X_t$ the transition $X_t \rightarrow X_{t+1}$ is defined as follows:

1. Choose an edge $e = (v, w)$ uniformly at random from $E$.
2. If $e \in X_t$ then set $X_{t+1} = X_t \setminus \{e\}$.
3. If $v$ and $w$ are unmatched in $X_t$ then set $X_{t+1} = X_t \cup \{e\}$.
4. Otherwise, set $X_{t+1} = X_t$.

Symmetric and ergodic and thus $\pi$ is uniform over $\Omega$.

How fast does it reach $\pi$?

Further topic (in MCMC class):

we’ll see that it’s close to $\pi$ after $\text{poly}(n)$ steps and this holds for every $G$.

Thus, we can generate a random matching of a graph in polynomial-time.