Bounding Mixing Times of Markov Chains

Michael Rabbat

27 February 2014

A Quick Review of Markov Chains

A discrete-time Markov chain is a random process

$$X_1, X_2, X_3, \ldots X_{k-1}, X_k, \ldots,$$

satisfying the Markov property

$$\Pr(X_k | X_{k-1}, X_{k-2}, \ldots, X_1) = \Pr(X_k | X_{k-1}) \; .$$

We'll focus on finite-state Markov chains, where $X_k \in \Omega$ for all $k \geq 1$ and $|\Omega| < \infty$.

Probability Transition Matrix

Suppose $|\Omega| = N$. A discrete-time finite-state Markov chain is completely described by its $N \times N$ *probability transition matrix* P, which has entries

$$P(x, y) = \Pr(X_k = y | X_{k-1} = x) \ .$$

The transition matrix P is *row-stochastic*:

$$\sum_{y \in \Omega} P(x, y) = 1 \ .$$

Let $\boldsymbol{\pi}_0 \in [0,1]^N$ be a distribution over Ω; i.e., $\sum_{x\in\Omega} \pi_0(x) = 1$. Define

$$\begin{aligned}
\boldsymbol{\pi}_k^T &= \boldsymbol{\pi}_{k-1}^T P \\
&= \boldsymbol{\pi}_{k-2}^T PP \\
&\vdots \\
&= \boldsymbol{\pi}_0^T P^k \ .
\end{aligned}$$

Think of $\boldsymbol{\pi}_k$ as the distribution of a particle's location after k steps according to P, given its initial distribution is $\boldsymbol{\pi}_0$.

Stationary Distribution

If, for any $\boldsymbol{\pi}_0$, the distribution $\boldsymbol{\pi}_k$ converges to a limit,

$$\lim_{k \to \infty} \boldsymbol{\pi}_k = \boldsymbol{\pi} \ ,$$

then

- we say the Markov chain P is *ergodic*, and
- we call $\boldsymbol{\pi}$ its *stationary distribution*.

Reversible Markov Chains

A Markov chain P with stationary distribution $\boldsymbol{\pi}$ is *reversible* if

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \text{for all } x, y \in \Omega \ .$$

Random Walk on a Graph

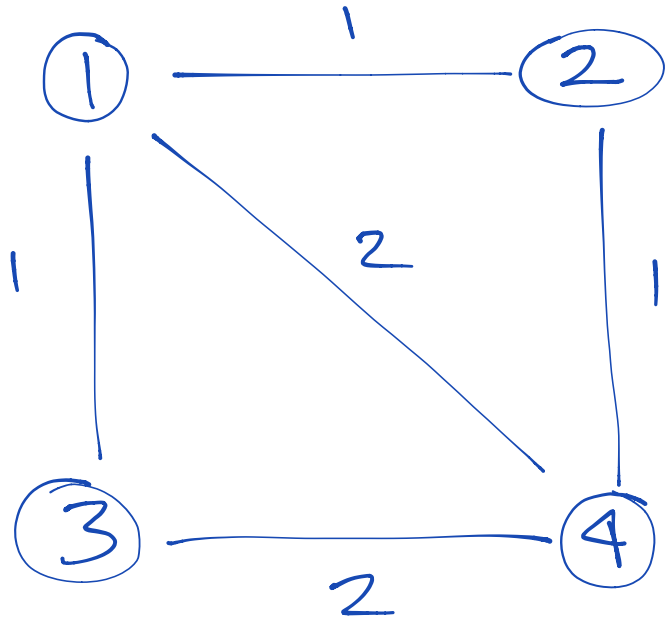
Every weighted undirected graph $G = (\mathcal{N}, \mathcal{A}, w)$ with symmetric positive edge weights $w(x, y)$ satisfying

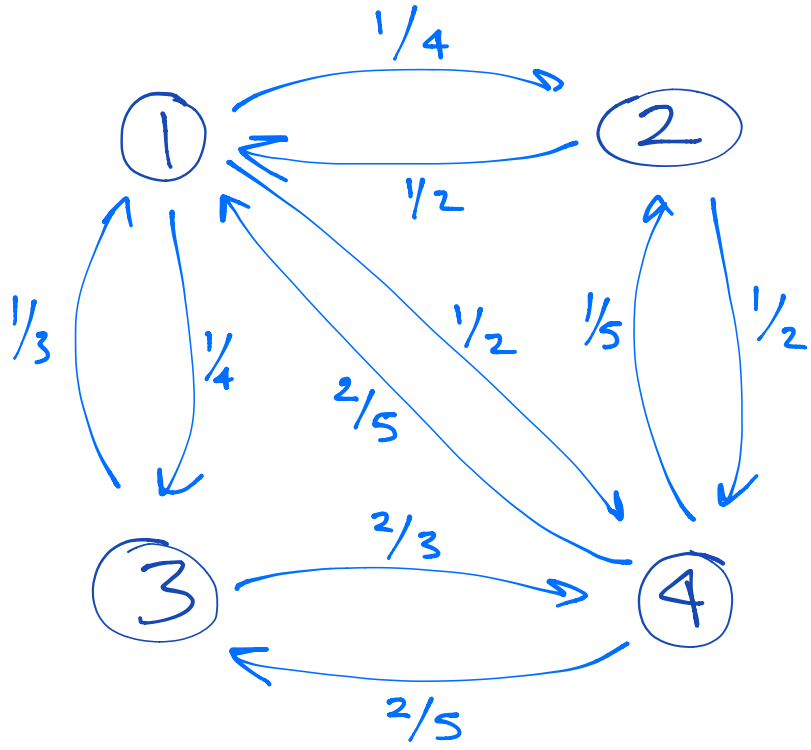
$$\begin{cases} w(x,y) = w(y,x) > 0 & \text{if } (x,y) \in \mathcal{A} \\ w(x,y) = 0 & \text{if } (x,y) \notin \mathcal{A} \end{cases}$$

corresponds to a reversible Markov chain on $\Omega = \mathcal{N}$ with

$$P(x,y) = \frac{w(x,y)}{\sum_{z\in\Omega} w(x,z)} \quad \text{and} \quad \pi(x) = \frac{\sum_{z\in\Omega} w(x,z)}{\sum_{y\in\Omega}\sum_{z\in\Omega} w(y,z)} \,.$$

Also, to every reversible P corresponds a symmetric weighted graph $G = (\mathcal{N}, \mathcal{A}, w)$!


1
2
1
2
1
3
4
2


1/4
1
2
1/2
1/3
1/4
1/2
1/5
1/2
2/5
2/3
3
4
2/5

For many applications involving reversible Markov chains, we would like to know:

How quickly does $\boldsymbol{\pi}_k$ converge to $\boldsymbol{\pi}$?

Example: Estimating the Facebook Degree Distribution

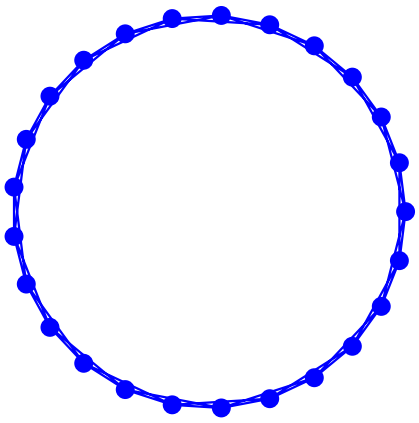
- How many people have d friends, $d = 1, 2, 3, \ldots$?
- Need an iid sample of people
- Facebook API allows queries of the form
  - Give me a random friend of user X
  - How many friends does X have?
- Take a random walk on the Facebook graph!
- *How many steps until arriving at an independent node? (with distribution $\pi(x) \propto d(x)$)*

Other Applications

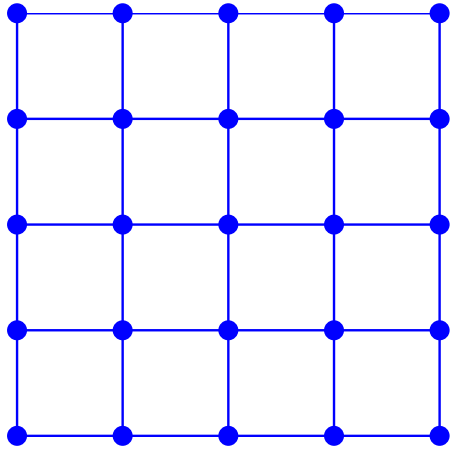
- Counting configurations that satisfy a given property
- Markov Chain Monte Carlo methods
  - Bayesian statistics
  - Simulation
- Convergence rates of randomized algorithms

Developing Intuition

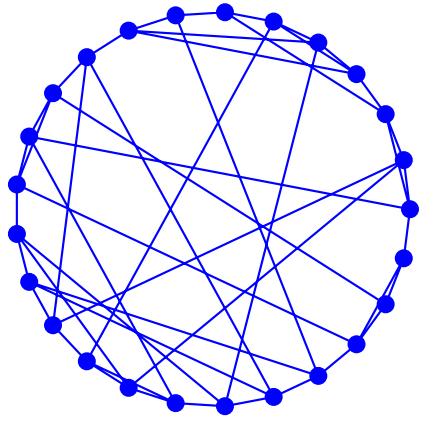
- How quickly does π_k converge to π?
- Convergence rate related to properties of network structure?



4-connected Ring



2-d Grid



Random 4-regular Graph

Measuring How Close are Two Distributions

Variation distance between two probability vectors $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$

$$\|\boldsymbol{\mu} - \boldsymbol{\pi}\|_{\mathsf{Var}} = \max_{A \subset \Omega} |\mu(A) - \pi(A)| = \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \pi(x)| \ .$$

Fact: $0 \leq \|\boldsymbol{\mu} - \boldsymbol{\pi}\|_{\mathsf{Var}} \leq 1$

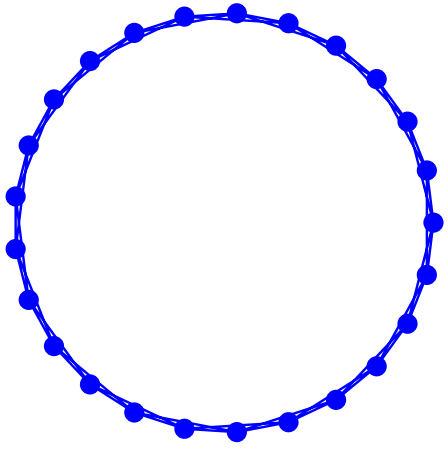
Quantifying the Rate of Convergence

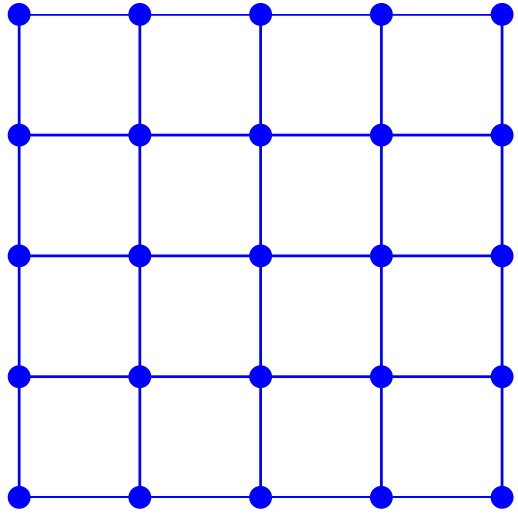
The ϵ-mixing time of the Markov chain P with stationary distribution $\boldsymbol{\pi}$ started from state x (i.e., $\boldsymbol{\pi}_0 = \boldsymbol{e}_x$) is

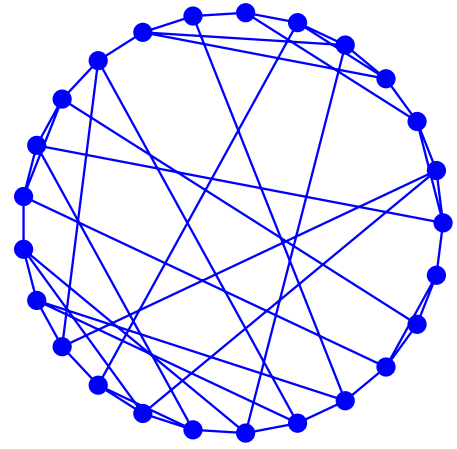
$$\tau_x(\epsilon) = \min\{k\colon\ \|\boldsymbol{\pi}_{k'} - \boldsymbol{\pi}\|_{\mathsf{Var}} \leq \epsilon \text{ for all } k' \geq k\}\ .$$

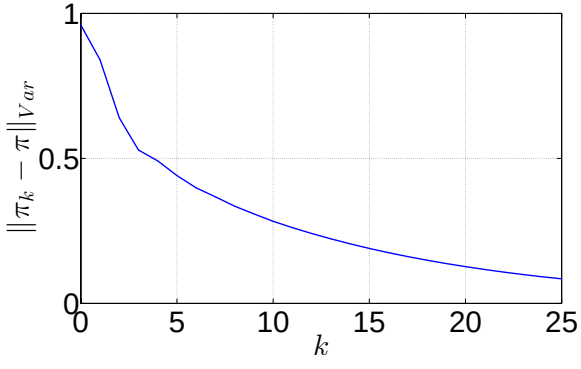
The ϵ-mixing time of the Markov chain P is

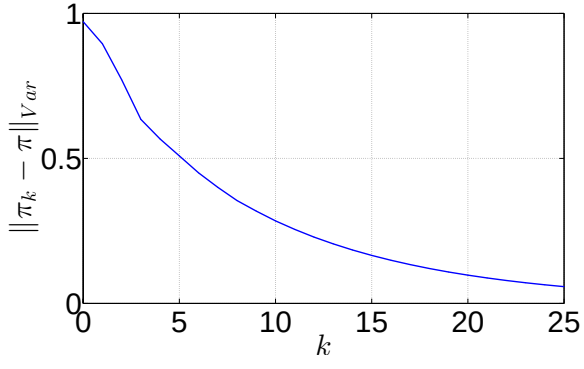
$$\tau(\epsilon) = \max_{x \in \Omega} \tau_x(\epsilon)\ .$$

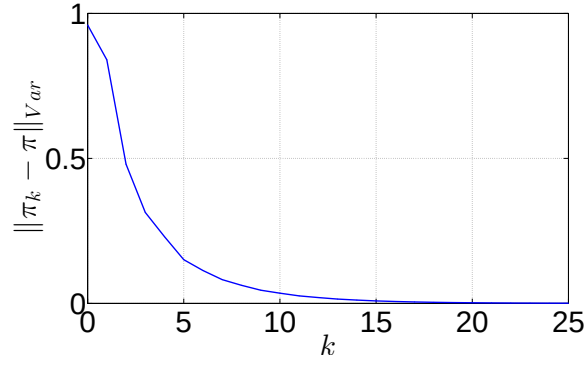







1
0.5
0
0
5
10
15
20
25
k
$\|\pi_k - \pi\|_{Var}$


1
0.5
0
0
5
10
15
20
25
k
$\|\pi_k - \pi\|_{Var}$


1
0.5
0
0
5
10
15
20
25
k
$\|\pi_k - \pi\|_{Var}$

Perron-Fronbenius Theory and Eigenvalues of P

Let P be a reversible Markov chain with stationary distribution π.

P is row-stochastic $\implies P\mathbf{1} = \mathbf{1}$.
Also, since $\boldsymbol{\pi}_{k-1}^T P = \boldsymbol{\pi}_k^T \to \boldsymbol{\pi}$, we must have $\boldsymbol{\pi}^T P = \boldsymbol{\pi}^T$.
$\implies 1$ is an eigenvalue of P

All other eigenvalues of P have magnitude less than 1.

Consider the eigenvalue decomposition $P = U\Lambda U^{-1} = U\Lambda\tilde{U}^T$.
Let $\lambda_1 = 1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$ denote the eigenvalues of P.
Since we know $\boldsymbol{\pi}_k \to \boldsymbol{\pi}$, it must be that

$$\begin{aligned}
\boldsymbol{\pi}_k^T P &= \boldsymbol{\pi}_0^T P^k \\
&= \boldsymbol{\pi}_0^T U \Lambda^k \tilde{U}^T \\
&= \sum_{i=1}^{n} \lambda_i^k \cdot (\boldsymbol{\pi}_0^T \boldsymbol{u}_i) \cdot \tilde{\boldsymbol{u}}_i^T \\
&= 1^k \cdot \underbrace{(\boldsymbol{\pi}_0^T \mathbf{1})}_{=1} \cdot \boldsymbol{\pi}^T + \underbrace{\sum_{i=2}^{n} \lambda_i^k \cdot (\boldsymbol{\pi}_0^T \boldsymbol{u}_i) \cdot \tilde{\boldsymbol{u}}_i^T}_{\to 0 \text{ as } k \to \infty}
\end{aligned}$$

Eigenvalue Bound on the Mixing Time

Let P be a reversible Markov chain with stationary distribution $\boldsymbol{\pi}$.
Let $\lambda_1 = 1 > \lambda_2 \geq \cdots \geq \lambda_N > -1$ denote the eigenvalues of P.

Then

$$\tau(\epsilon) \leq \frac{\ln(1/\pi_{\min}) + \ln(1/\epsilon)}{1 - \lambda_{\max}}$$

where $\pi_{\min} = \min_{x \in \Omega} \pi(x)$ and $\lambda_{\max} = \max\{\lambda_2, |\lambda_N|\}$.

λ_2 and the Lazy Random Walk

Usually focus on λ_2 and forget about $|\lambda_N|$.

This is justifiable since $\widetilde{P} = \frac{1}{2}(I + P)$ has all eigenvalues in the range $(0, 1]$ and it is no more than twice as slow as P.

- $\lambda_i(\widetilde{P}) = \frac{1}{2} + \frac{1}{2}\lambda_i(P)$
- $\widetilde{P}$ is called the lazy version of P
- $1 - \lambda_2$ is called the spectral gap

What does this have to do with Network Optimization?

For a reversible Markov chain P with stationary distribution $\boldsymbol{\pi}$, define a $N \times N$ matrix Q with entries

$$Q(x, y) = \pi(x)P(x, y) = \pi(y)P(y, x) = Q(y, x) \ .$$

We will think of $Q(x, y)$ as the capacity of arc (x, y).

Given $S \subseteq \Omega$, let

- $\pi(S) = \sum_{x \in S} \pi(x)$
- $Q(S, S^c) = \sum_{x \in S} \sum_{y \notin S} Q(x, y)$

Cheeger's Inequality

The Cheeger constant (also called "conductance", "isoperimetric constant", and "bottleneck ratio") is defined as

$$\Phi = \min_{\substack{S \subseteq \Omega \\ 0 < \pi(S) \leq \frac{1}{2}}} \frac{Q(S, S^c)}{\pi(S)}.$$

Theorem (Cheeger's Inequality): For any reversible Markov chain P,

$$\frac{\Phi^2}{2} \leq 1 - \lambda_2(P) \leq 2\Phi \ .$$

Recall Max-Flow/Min-Cut

Single-Commodity Max Flow Problems

- Single-commodity: maximize the flow from one source s to one sink t
- We saw that the max flow is equal to the minimum cut separating s from t [Ford-Fulkerson]

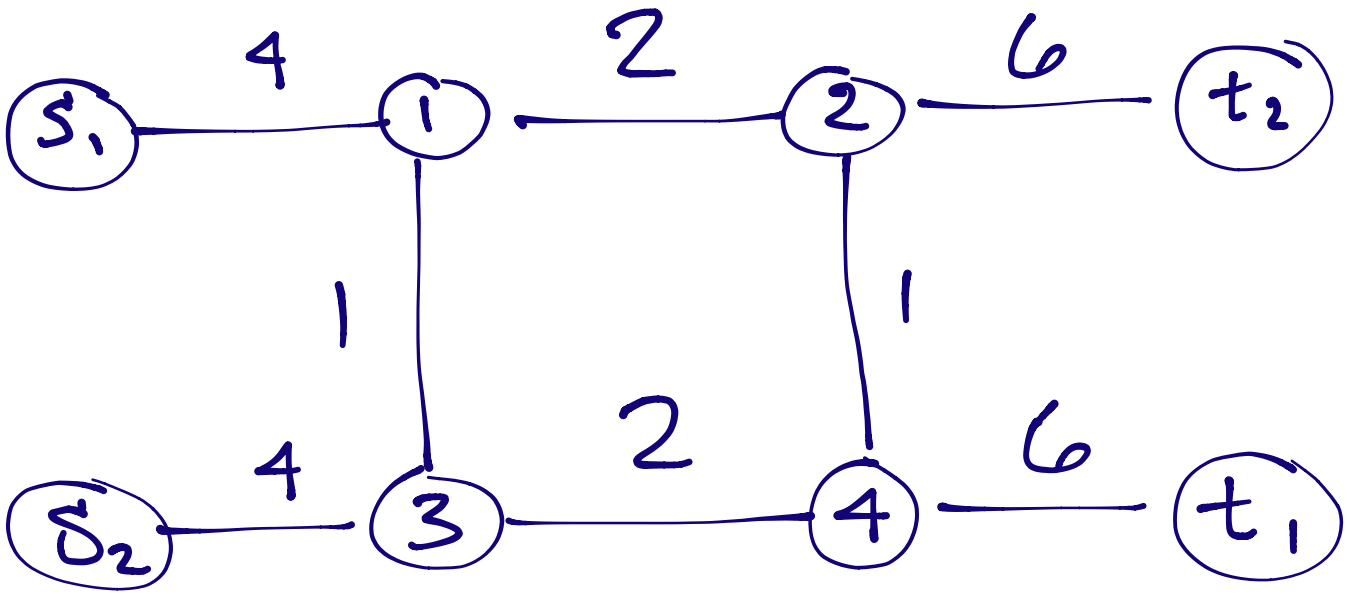
Multi-Commodity Flow Problems

- $m \geq 1$ different commodities
- Source s_i and sink t_i for commodity i
- Coupled via capacity constraints

Multi-commodity Flow

- $m \geq 1$ different commodities
- Source s_i and sink t_i for commodity i
- Also have demand $D_i > 0$ for flow i
- Maximize $f \in \mathbb{R}$ such that fD_i units of commodity i flow from s_i to t_i for $i = 1, \ldots, m$, subject to capacity constraints on each edge

An example with two sources and two sinks:



Multi-commodity Flow

Consider a multi-commodity flow problem on graph $G = (\mathcal{N}, \mathcal{A})$ with capacities $c(x, y) > 0$ for each $(x, y) \in \mathcal{A}$.

For a subset $S \subset \mathcal{N}$,

$$C(S, S^c) = \sum_{x \in S} \sum_{y \notin S} c(x, y)$$

$$D(S, S^c) = \sum_{i:\ s_i \in S \wedge t_i \notin S \text{ or } s_i \notin S \wedge t_i \in S} D_i$$

Multi-commodity Max-Flow / Min-Cut

Define the min-cut for a multi-commodity flow problem to be

$$\mathcal{C} = \min_{S \subseteq \mathcal{N}} \frac{C(S, S^c)}{D(S, S^c)}$$

If $D_i = 1 \; \forall i$ or, more generally, $D_i = \psi(s_i)\psi(t_i)$ for some function $\psi(\cdot) : \mathcal{N} \to \mathbb{R}$, then the max flow f^* satisfies

$$\Omega\left(\frac{\mathcal{C}}{\log N}\right) \leq f^* \leq \mathcal{C}$$

This result is due to Leighton and Rao (1999).

- Cheeger's inequality says $1 - \lambda_2(P) \geq \frac{1}{2}\Phi^2$ where

$$\Phi = \min_{\substack{S \subset \Omega \\ 0 < \pi(S) \leq \frac{1}{2}}} \frac{Q(S, S^c)}{\pi(S)}$$

- The multi-commodity min-cut satisfies $\mathcal{C} \geq f^*$ where

$$\mathcal{C} = \min_{S \subseteq \mathcal{N}} \frac{C(S, S^c)}{D(S, S^c)}$$

- Think of $Q(x, y)$ as capacity of edge (x, y)
- Think of routing commodity with demand $\pi(x)\pi(y)$ from x to y for each pair of states $x, y \in \Omega$

The Canonical Path Method [Jerrum and Sinclair]

Let $\Gamma = \{\gamma_{x,y}\}$ be a set of paths between all pairs $x, y \in \Omega$

Define

$$\rho(\Gamma) = \max_{e \in \mathcal{A}} \frac{1}{Q(e)} \sum_{\gamma_{x,y} \ni e} \pi(x)\pi(y)$$

Theorem: $\Phi \geq \frac{1}{2\rho(\Gamma)}$ for any paths Γ.

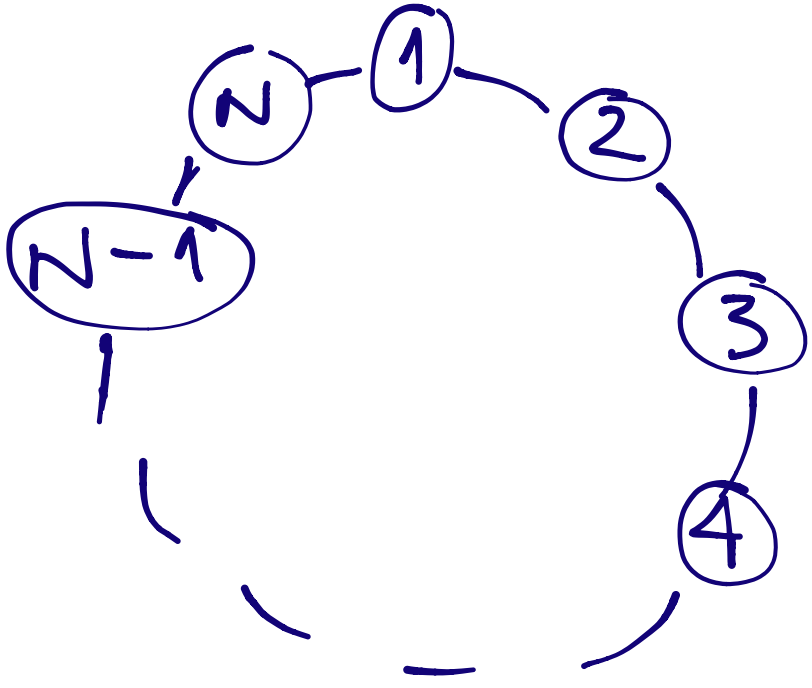
From Cheeger's inequality, this implies $1 - \lambda_2(P) \geq \frac{1}{8\rho(\Gamma)^2}$

Proof:

- Let $S \subseteq \Omega$ have $\pi(S) \leq \frac{1}{2}$
- Let $\partial S = \{(x, y) \colon x \in S \text{ and } y \notin S\}$
- The total demand crossing the cut from S to S^c is
  $\Rightarrow \; D(S, S^c) = \pi(S)\pi(S^c) \geq \frac{1}{2}\pi(S)$
- Also, summing the flow over all cut edges gives

$$\begin{aligned} D(S, S^c) &\leq \sum_{e \in \partial S} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y) \\ &= \sum_{e \in \partial S} Q(e) \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y) \\ &\leq \sum_{e \in \partial S} Q(e)\rho(\Gamma) \\ &= \rho(\Gamma) Q(S, S^c) \end{aligned}$$

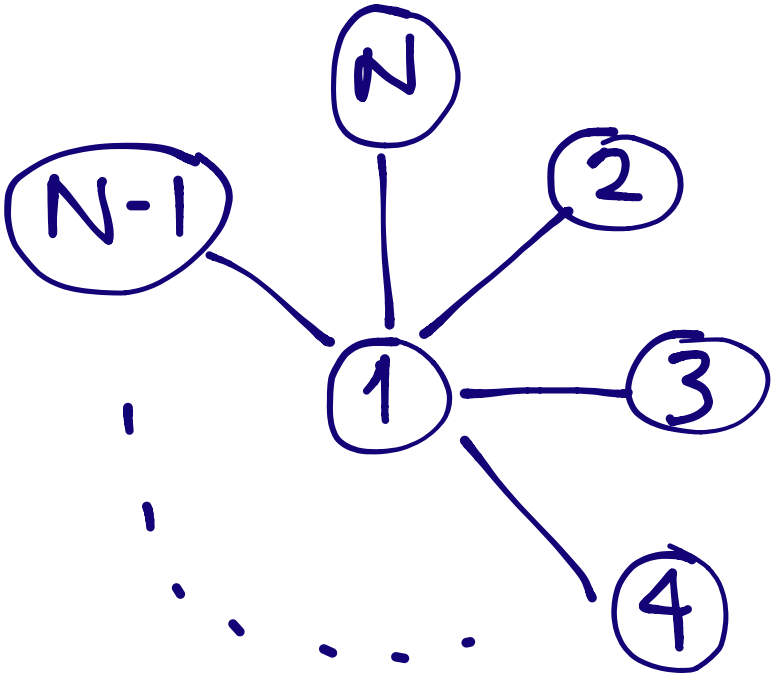
Example: Ring Graph



N nodes, assume N is odd

$$P(x,y) = \begin{cases} \frac{1}{2} & \text{if } (x,y) \in \mathcal{A} \\ 0 & \text{otherwise} \end{cases}$$

Example: Star Graph



N nodes with transition matrix

$$P(x,y) = \begin{cases} \frac{1}{N} & \text{if } x = 1 \\ 1 & \text{if } x \neq 1, y = 1 \\ 0 & \text{otherwise} \end{cases}$$

References and Additional Reading

- A. Sinclair, "Improved bounds for mixing rates of Markov chains and multicommodity flow," *Combinatorics, Probability & Computing*, vol. 1, pp. 351–370, 1992.
- P. Diaconis and D. Stroock, "Geometric bounds for eigenvalues of Markov chains," *The Annals of Applied Probability*, vol. 1, no. 1, pp. 36–61, 1991.
- D. Levin, Y. Peres, and E.L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, 2008.
- T. Leighton and S. Rao, "Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms," *Journal of the ACM*, vol. 46, no. 6, pp. 787–832, Nov. 1999.

Thank You!



Questions?

For more questions:

themisc@kth.se
chatw@kth.se
carlofi@kth.se
michael.rabbat@mcgill.ca