Markov chains, mixing time and reconfiguration

Tuesday, 10 May 2022 8:59 PM

Banff-GreenhillT...

Markov chains, mixing time and connections with reconfiguration

Catherine Greenhill

School of Mathematics and Statistics
UNSW Sydney

Combinatorial Reconfiguration workshop, Banff

Reconfiguration:
(my current understanding, based on Nishimura's 2018 survey)

We have

- a (finite?) set Ω of solutions to a problem instance
- a notion of adjacency of solutions
- the reconfiguration graph $\mathcal{G}$ with vertex set Ω and edges given by adjacency

Usually (I imagine) Ω is exponentially large as a function of the "size" n of the problem instance.

Reconfiguration: many structural questions

- reachability:
  Given $X, Y \in \Omega$, is there a path from X to Y in $\mathcal{G}$?
- connectivity: Is $\mathcal{G}$ connected?
- shortest paths:
  Given $X, Y \in \Omega$, what is length of
  shortest path from X to Y in $\mathcal{G}$?
- diameter: What is the diameter of $\mathcal{G}$?

Reconfiguration: many algorithmic questions, e.g.

- algorithms to decide reachability, connectivity,
- algorithms to find shortest paths or calculate diameter
- find "best" solution reachable from a given starting point

Related questions of computational complexity, parameterised complexity etc. Even reachability might be intractable!

Markov chains: (discrete time, finite state space)

We have

- a finite set Ω, called a state space
- some allowed transitions, with the next transition chosen randomly according to some rule
- a (directed) graph $\mathcal{G}$ stores the set of all possible transitions

(We'll be more precise soon.)

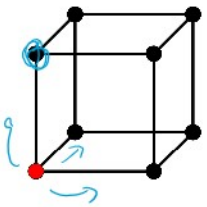
Usually Ω is exponentially large as a function of some parameter n.

A Markov chain $\mathcal{M}$ on state space Ω is a stochastic process $X_0, X_1, X_2, \ldots$ which is memoryless:

$$\Pr(X_{t+1} = y \mid X_0 = x_0, \ldots, X_t = x_t) = \Pr(X_{t+1} = y \mid X_t = x_t)$$

for all $t \in \mathbb{N}$ and $x_0, \ldots, x_t, y \in \Omega$.

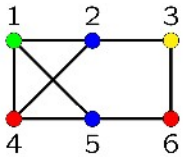
Here each $X_t = X_t(n)$ is a random element of a set $\Omega = \Omega_n$ which is usually finite but exponentially large with respect to some parameter n.



Example: A Markov chain for graph colourings.

Let $k \geq 3$ be a fixed integer and let $G = (V, E)$ be a graph. Write $[k] = \{1, 2, \ldots, k\}$.

A k-colouring of G is a function $\sigma : V \to [k]$ such that if $\{x, y\} \in E$ then $\sigma(x) \neq \sigma(y)$.

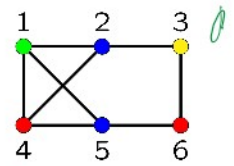


Let $\Omega_k(G)$ be the set of all k-colourings of G.

A simple Markov chain on $\Omega_k(G)$ has the following transition procedure: from the current state $X \in \Omega_k(G)$ do

- Choose $v \in V$ uniformly at random,
- Choose $c \in [k]$ uniformly at random,
- Recolour v with c to give a new state X', if possible. Otherwise, let $X' = X$.

This is the Glauber dynamics or single-site update chain.



$(v, c) = (4, \square)$? ✓

$(v, c) = (3, \square)$? ✗

We say that the Markov chain $\mathcal{M}$ is time-homogeneous if the transition probabilities from a given state x are independent of t.

For a time-homogeneous chain, the transition probabilities can be stored in the transition matrix P of $\mathcal{M}$, with rows and columns indexed by Ω and entries

$$P(x, y) = \Pr(X_{t+1} = y \mid X_t = x).$$

Matrix P is stochastic: all rows sum to 1.

The matrix P is too large to work with directly. (e.g. can't find eigenvalues in polynomial time).

AIM: sampling Ω

We can also define the directed graph $\mathcal{G}$ underlying the Markov chain, with vertex set Ω and (x, y) a directed edge if and only if $P(x, y) > 0$, for all $x, y \in \Omega$.

A self-loop is a directed edge (x, x) in $\mathcal{G}$.

[Similar to "adjacency", "reconfiguration graph".]

If the state space Ω is connected under moves of $\mathcal{M}$ then we say that $\mathcal{M}$ is irreducible.

[This is (strong) "connectivity of the reconfiguration graph".]

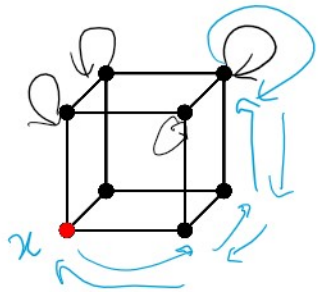
We don't usually worry about shortest paths, diameter.

Write $P_x^t(\cdot)$ for the distribution of state X_t of the Markov chain $\mathcal{M}$, after t steps from initial state $X_0 = x$.

Say $\mathcal{M}$ is aperiodic if for all $x \in \Omega$,

$$\gcd\{t \in \mathbb{Z}^+ \mid P_x^t(x) > 0\} = 1.$$

One self-loop is sufficient to guarantee aperiodicity, if $\mathcal{M}$ is irreducible.



Classical Markov chain theory

A row-vector π with nonnegative entries which add up to 1 is a stationary distribution of Markov chain $\mathcal{M}$ if

$$\pi P = \pi.$$

We say that Markov chain $\mathcal{M}$ is ergodic if it is irreducible and aperiodic.

If $\mathcal{M}$ is ergodic then $\mathcal{M}$ has a unique stationary distribution π such that $\pi(x) > 0$ for all $x \in \Omega$, and

$$\lim_{t \to \infty} P_x^t(y) = \pi(y)$$

for all $x, y \in \Omega$.

See for example: book by Levin, Peres & Wilmer (2009).

How to find π?

We say that Markov chain $\mathcal{M}$ satisfies the detailed balance equations with respect to row vector ψ if

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

We also say that $\mathcal{M}$ is reversible, or time-reversible, with respect to ψ.

If $\mathcal{M}$ is ergodic and satisfies detailed balanced with respect to some vector $\psi \neq 0$, then the unique stationary distribution π is given by normalising ψ.

In this case, the underlying graph $\mathcal{G}$ is really undirected.

eg reconfiguration graph.

Example: A Markov chain for graph colourings.

Fact: The Glauber dynamics $\mathcal{M}_k(G)$ is irreducible on $\Omega_k(G)$ when $k \geq \Delta(G)+2$, where $\Delta(G)$ is the maximum degree of G.
Also $\mathcal{M}_k(G)$ is aperiodic as $P(x,x) \geq 1/k$ for any $x \in \Omega_k(G)$.
So $\mathcal{M}_k(G)$ is ergodic.

The transition probabilities satisfy

$$P(x,y) = \frac{1}{kn} = P(y,x)$$

whenever $x,y \in \Omega_k(G)$ differ on a single vertex,
and $P(x,y) = 0$ for all other $x \neq y$.

The chain satisfies detailed balance with respect to $(1,1,\ldots,1)$, so its stationary distribution is uniform.

We use the detailed balanced equations to design our Markov chains, so they have the desired stationary distribution.

In particular, if π is uniform then P must be symmetric.

We also need to prove irreducibility; that is, the underlying graph $\mathcal{G}$ is connected.

Aperiodicity is easy: just ensure $P(x,x) > 0$ for all $x \in \Omega$.

Now we have a very well behaved Markov chain which converges to its stationary distribution.... eventually.

Q: How quickly?

For probability distributions σ, μ on Ω, the total variation distance between σ and μ is

$$d_{\mathrm{TV}}(\sigma,\mu) = \frac{1}{2}\sum_{x\in\Omega} |\sigma(x) - \mu(x)|. \quad = \max_{A\subseteq\Omega} \sigma(A) - \mu(A).$$

Define the mixing time $\tau(\varepsilon)$ of $\mathcal{M}$ by

$$\tau(\varepsilon) = \max_{x\in\Omega} \min\{t \mid d_{\mathrm{TV}}(P_x^t, \pi) < \varepsilon\}.$$

Convergence is monotonic.

Here $\varepsilon > 0$ is a small user-defined tolerance.

We say that the Markov chain $\mathcal{M}$ is rapidly mixing if $\tau(\varepsilon)$ is bounded above by some polynomial in $\log|\Omega|$ and $\log(\varepsilon^{-1})$.

This is a strong condition: we want to get exponentially close to the stationary distribution, over an exponentially large state space, in polynomial time.

Linear algebra $\Rightarrow$ eigenvalues of P control convergence of $\mathcal{M}$. But Ω is too big to allow direct computation of the eigenvalues in polynomial time.

Other methods:

- coupling, path coupling [probabilistic]
- geometric arguments [*]
- functional inequalities,
  e.g. Poincaré inequality, log-Sobolev inequality

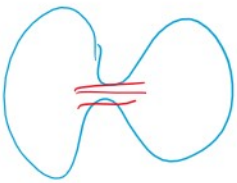
Geometry of the state space

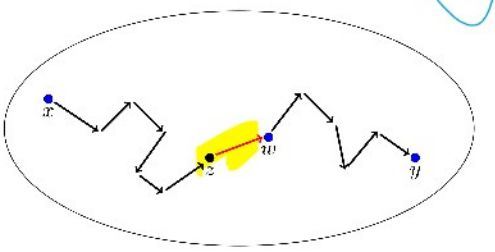
Which state space do you think encourages rapid mixing?

Constrictions in the state space make it difficult for the chain to escape: exponential mixing time!

Lack of constrictions allows chain to mix freely. Results by Jerrum & Sinclair (1987) make this precise: conductance.

Method II: Canonical paths





- For all pairs $(x, y) \in \Omega^2$, define a path γ_{xy} from x to y, where each step is a transition of the Markov chain.
- Analyse the congestion of the set of all paths: are any transitions heavily loaded? Then apply Sinclair (1992).

NB: Canonical paths might not be shortest paths.

Instead, we want to avoid having too many paths going through the same edge of $\mathcal{G}$.

(These conditions almost seem orthogonal!??, somehow??)

Let $\mathcal{M}$ be time-homogeneous, ergodic, reversible Markov chain with $N = |\Omega|$. The eigenvalues of the transition matrix P are real and satisfy

$$1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} > -1$$

and the mixing time of $\mathcal{M}$ is controlled by

$$\lambda_{\max} = \max\{\lambda_1, |\lambda_{N-1}|\}.$$

If $\pi^* = \min\{\pi(x) \mid x \in \Omega\}$ then

$$\tau(\varepsilon) \leq (1 - \lambda_{\max})^{-1} \left(\log(1/\pi^*) + \log(\varepsilon^{-1})\right).$$

See Sinclair (1992).

The quantity $(1 - \lambda_{\max})^{-1}$ is the relaxation time of $\mathcal{M}$.

Typically $\lambda_{\max} = \lambda_1$. This can be guaranteed by making the chain lazy, that is, replacing P with $(I + P)/2$.

However, a method of Diaconis & Saloff-Coste (1993) can be applied directly to bound $(1+\lambda_{N-1})^{-1}$, without resorting to laziness.

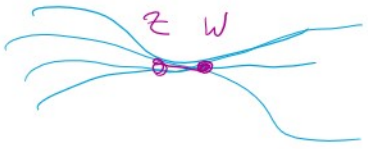
As a special case, if every state has a self-loop then

$$(1 + \lambda_{N-1})^{-1} \leq \tfrac{1}{2} \max_{x \in \Omega} P(x, x)^{-1}.$$

The point is, we can focus on λ_1.

The quantity $1 - \lambda_1$ is called the spectral gap.

Sinclair (1992): Let $\mathcal{M}$ be time-homogenous, ergodic and reversible with stationary distribution π.



Let $\Gamma = \{\gamma_{xy} \mid x, y \in \Omega\}$ be a set of canonical paths for $\mathcal{M}$. Define the congestion

$$\bar{\rho} = \bar{\rho}(\Gamma) = \max_{zw \in E(\mathcal{G})} \frac{1}{\pi(z)P(z,w)} \sum_{\gamma_{xy} \ni zw} \pi(x)\pi(y)\,|\gamma_{xy}|$$

where $|\gamma_{xy}|$ is the length of the path γ_{xy}.

$\pi(z)P(z,w)$
$= \pi(w)P(w,z)$
$= Q(z,w)$

Then

$$(1-\lambda_1)^{-1} \le \bar{\rho}.$$

So we want an upper bound on $\bar{\rho}$ which is polynomial in $\log|\Omega|$.

Theorem (Sinclair, 1992)
Suppose that $\mathcal{M}$ is time-homogenous, ergodic and reversible, and let Γ be a set of canonoical paths for $\mathcal{M}$. Then

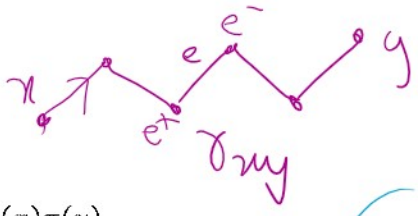
$$(1-\lambda_1)^{-1} \le \bar{\rho}.$$

Proof. Let $L = I - P$, so that the eigenvalues of L are $\mu_i = 1 - \lambda_i$. The variational characterisation of μ_1 is

$$\mu_1 = \inf_{\psi} \frac{\sum_{x,y\in\Omega} \big(\psi(x)-\psi(y)\big)^2 \pi(x)P(x,y)}{\sum_{x,y\in\Omega} \big(\psi(x)-\psi(y)\big)^2 \pi(x)\pi(y)},$$

with the infimum taken over all non-constant functions $\psi : \Omega \to \mathbb{R}$.

smallest positive eigenvalue of L



Now

$$\sum_{x,y} \big(\psi(x)-\psi(y)\big)^2 \pi(x)\pi(y)$$
$$= \sum_{x,y} \pi(x)\pi(y) \left(\sum_{e \in \gamma_{xy}} \big(\psi(e^+)-\psi(e^-)\big) \right)^2$$
$$\le \sum_{x,y} \pi(x)\pi(y)\,|\gamma_{xy}| \sum_{e \in \gamma_{xy}} \big(\psi(e^+)-\psi(e^-)\big)^2,$$

writing $e = e^- e^+$ for each $e \in \gamma_{xy}$. The final line uses the Cauchy–Schwarz inequality.

Exchanging the order of summation gives

$$\sum_{x,y} \big(\psi(x)-\psi(y)\big)^2 \pi(x)\psi(y)$$
$$\le \sum_{e} \big(\psi(e^+)-\psi(e^-)\big)^2 \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y)\,|\gamma_{xy}|$$
$$\le \sum_{e} \big(\psi(e^+)-\psi(e^-)\big)^2 \pi(e^-)\,P(e^-,e^+)\,\rho(\bar{\Gamma})$$
$$= \rho(\bar{\Gamma}) \sum_{x,y} \big(\psi(x)-\psi(y)\big)^2 \pi(x) P(x,y),$$

which implies that

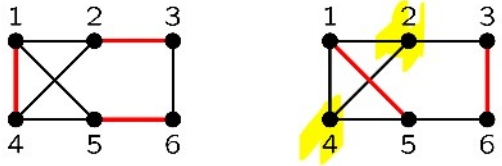
$$1 \le \bar{\rho}\,(1-\lambda_1),$$

completing the proof. □

Example:
A Markov chain for perfect and near-perfect matchings.

Let $G = (V, E)$ be a graph. A matching in G is a set of edges $M \subseteq E$ such that no vertex is incident with more than one edge of M.

If $|V| = 2n$ and $|M| = n$ then M is a perfect matching. Matchings with $n - 1$ edges are called near-perfect.



Let $\mathcal{P}$ be the set of all perfect matchings of G and let $\mathcal{N}$ be the set of all near-perfect matchings of G. Broder (1986) introduced a Markov chain with state space $\mathcal{P} \cup \mathcal{N}$.

From current state $M \in \mathcal{P} \cup \mathcal{N}$,

- with probability $\frac{1}{2}$ let $M' = M$; otherwise: lazy
  - Choose $e = uv \in E(G)$ uniformly at random.
  - If $M \in \mathcal{P}$ and $e \in M$ then let $M' = M \setminus e$.

- If $M \in \mathcal{N}$ and both u, v are unmatched in M then let $M' = M + e$.

- If $M \in \mathcal{N}$ and exactly one of u, v are matched in M, with matching edge e', then let $M' = (M \setminus e') + e$.

The Broder chain is irreducible (honest!), and aperiodic (in fact it is lazy).

If $P(M, M')$ is nonzero for distinct M, M' then

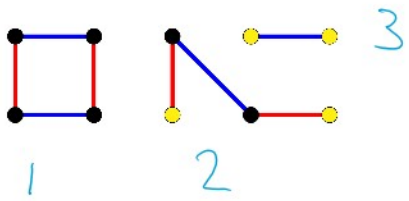
$$P(M, M') = \frac{1}{2m} = P(M', M),$$

where m is the number of edges of G.
So the stationary distribution is uniform.

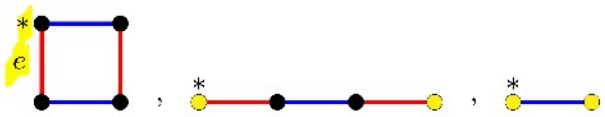
Broder (1986) applied coupling to try to analyse this chain, but his proof was incorrect. Jerrum & Sinclair (1989) used canonical paths.

Let M, M' be two elements of $\mathcal{P} \cup \mathcal{N}$.

The symmetric difference $M \ominus M'$ is the disjoint union of cycles and at most two paths.



We process the components of $M \ominus M'$ in some canonical order, and let the start vertex of each cycle (respectively, path) be the least labelled vertex (respectively, endvertex) in the cycle/path.

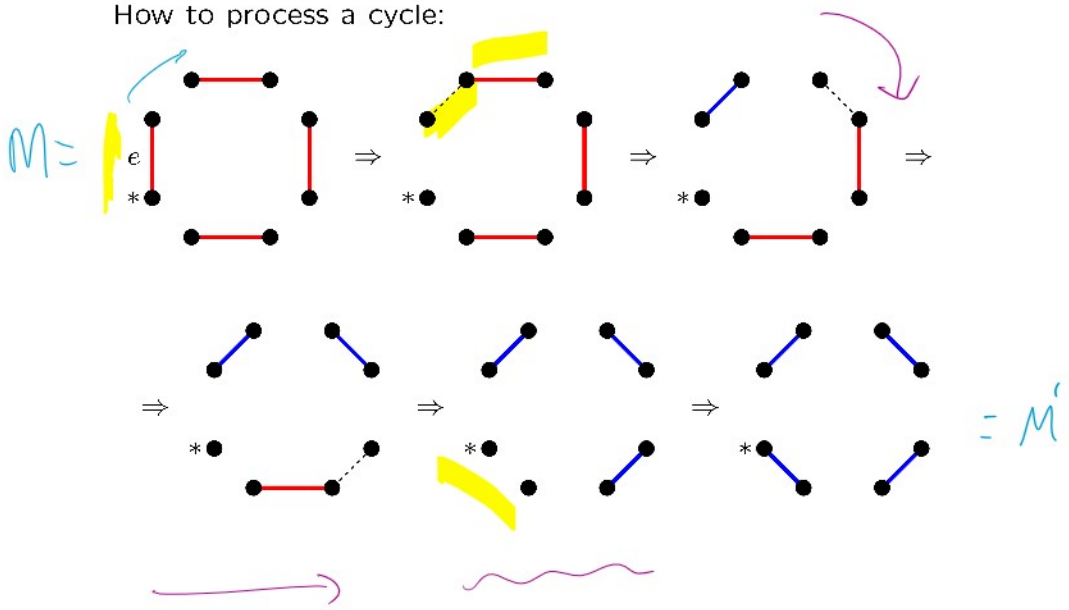


Denote by $e = e(C)$ the edge of M incident with the start-vertex of each cycle C.

With each component we process, we add a few more steps to our canonical path from M to M'.

M is a perfect matching

How to process a cycle:



Paths are processed similarly. This defines a canonical path γ_{XY} between each pair $(X, Y) \in \mathcal{P} \cup \mathcal{N}$.

Now we must analyse the set of canonical paths $\Gamma = \{\gamma_{XY}\}$. What is the load on each transition?

Key tool: define the encoding $\eta_t(X, Y)$ of the transition t on the canonical path γ_{XY}, such that if you know $t = (M, M')$ and η_t then you can uniquely recover (X, Y).

Encodings used in Jerrum & Sinclair (1989):

If $t = (M, M')$ where $M, M' \in \mathcal{N}$ and we are processing a cycle, then

$$\eta_t(X, Y) = (X \oplus Y \ominus (M \cup M')) \setminus \{e\},$$

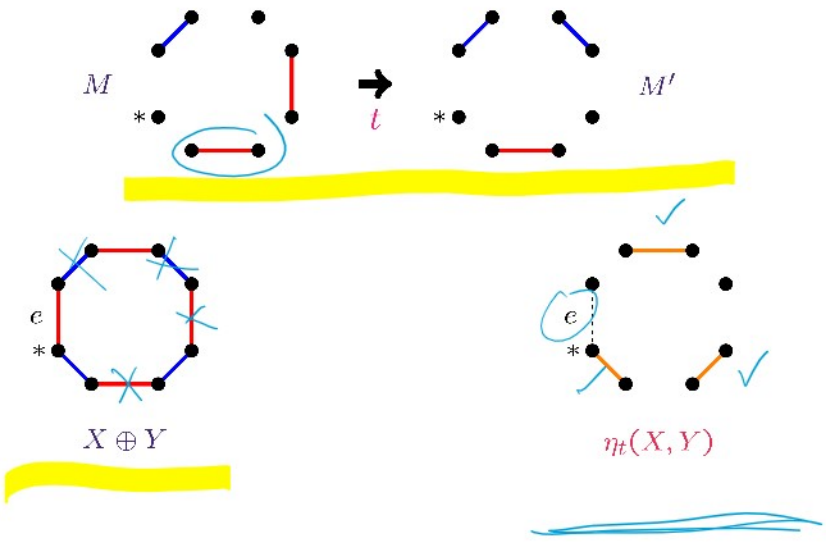
where e is the edge of X adjacent to the start-vertex of the cycle; and otherwise

$$\eta_t(X, Y) = X \oplus Y \oplus (M \cup M').$$

Q: Why remove e?

A: This ensures that $\eta_t(X, Y)$ belongs to $\mathcal{P} \cup \mathcal{N}$.

Example: $\eta_t(X, Y) = (X \oplus Y \ominus (M \cup M')) \setminus \{e\}$



Facts:

- The encoding $\eta_t(X,Y)$ is either a perfect or near-perfect matching of G.
- If you know $t=(M,M')$ and η_t then you can uniquely recover (X,Y).

Hence no transition lies on more than $|\mathcal{P}\cup\mathcal{N}|$ canonical paths.

Also, if G has $2n$ vertices then $|\gamma_{XY}| \le 2n$ for all $X,Y \in \mathcal{P}\cup\mathcal{N}$.

Thus

$$\bar{\rho} = \max_{zw\in E(\mathcal{G})} \frac{1}{\pi(z)P(z,w)} \sum_{\gamma_{xy}\ni zw} \pi(x)\pi(y)\,|\gamma_{xy}|$$

$$\le |\mathcal{P}\cup\mathcal{N}|\,2m \cdot |\mathcal{P}\cup\mathcal{N}| \cdot \frac{2n}{|\mathcal{P}\cup\mathcal{N}|^2}$$

$$= 4mn.$$

Also

$$1/\pi^* = |\mathcal{P}\cup\mathcal{N}| \le (2n)! \le \exp(2n\log n).$$

Jerrum & Sinclair (1989): The Broder chain has mixing time

$$\tau(\varepsilon) \le 4mn(2n\log n + \log(\varepsilon^{-1}))$$

where G has n vertices and m edges.

The canonical path method is a thing of beauty, when it works. But...

* Usually, it does NOT lead to tight bounds, and

* Finding a good set of canonical paths can be tricky.

Huang, Lu & Zhang (SODA 2016),
"Canonical paths for MCMC: from art to science".

They build on work of McQuillan (2013) to reduce the task of designing canonical paths to solving a system of linear equations.

Sinclair (1992) introduced the multicommodity flow method, where $\pi(x)\pi(y)$ units of flow are split across a set of paths from x to y in $\mathcal{G}$.

This method has been used to analyse the switch chain for sampling graphs (or directed graphs or bipartite graphs) with given degree sequence, under various conditions.

Transitions:
From a given graph, choose a pair of non-incident edges randomly and replace them by a (randomly chosen) pair of edges without changing the degree sequence.

$G \oplus H$

ALL pairings

ψ

$\gamma_{x,y}(\psi)$

References:

- Several chapters by Mark Jerrum
  "Counting, Sampling and Integrating: Algorithms and Complexity" (2003) Monograph
- Levin, Peres & Wilmer,
  "Markov Chains and Mixing Times" (2009) Book

- Sinclair, "Improved bounds for mixing rates of Markov chains and multicommodity flow", CPC **1** (1992), 351–370

The switch chain is ergodic and its stationary distribution is uniform.

The switch chain has been shown to be rapidly mixing in various situations, by several authors, using multicommodity flow. (See for example my BCC 2021 talk.)

But the mixing time bounds are just awful, e.g. Cooper, Dyer, Greenhill (2007) proved that

$$\tau(\varepsilon) \leq d^{23}\, n^8 \big(dn \log(dn) + \log(\varepsilon^{-1})\big)$$

for d-regular graphs on n vertices.

Q: Is it really this bad?

A: Maybe not.

Tikhomirov & Youssef, `arXiv.2007.02729` proved a mixing time bound of

$$C\, dn \big(dn \log dn + \log(2\varepsilon^{-1})\big)$$

for the switch chain on d-regular bipartite graphs, where $3 \leq d \leq n^c$, for some constants $c, C > 0$.

This is a **huge** improvement on any previously-known bound.

Proof is long & technical, involves establishing a Poincaré inequality to bound the eigenvalues of the chain directly.

Connections with approximate counting

If a problem is "self-reducible" then approximate counting can be reduced to approximately uniform sampling, e.g. using Markov chains.

Other approaches to approximate counting using deterministic algorithms:

- correlation decay method (Weitz, 2006)
- polynomial interpolation method (Barvinok, 2016)

Some related computational complexity questions, mostly related to approximate counting:

Q: Is the counting problem #P-complete?

Dyer, Goldberg, Greenhill, Jerrum (2003) defined approximation-preserving reductions (AP-reductions) and identified 3 classes of approximate counting problems:

- solvable in randomized polynomial time (RP),
- AP-interreducible with #SAT,
- AP-interreducible with #BIS

Here #BIS is the problem of counting independent sets in bipartite graphs.

* **Thank you!** *