Self-Repellent Random Walks on General Graphs – Achieving Minimal Sampling Variance via Nonlinear Markov Chains

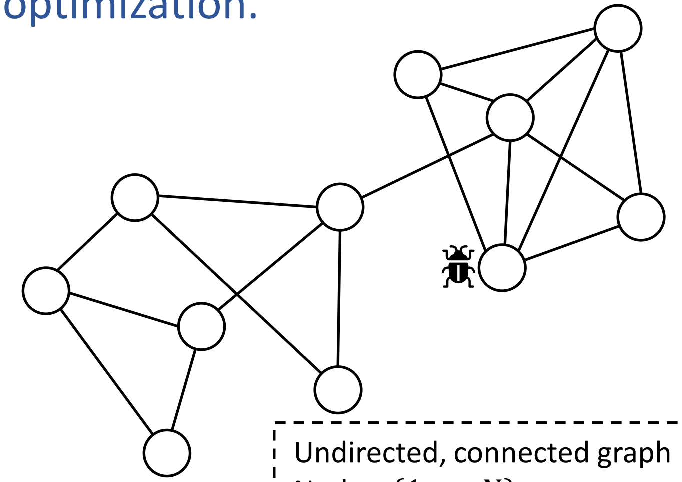
by Vishwaraj Doshi[¶], Jie Hu[†], and Do Young Eun[†]

[¶] Advanced Analytics, IQVIA. Previously at Operations Research Graduate Program, North Carolina State University.
[†] Dept. of Electrical and Computer Engineering, North Carolina State University

International Conference on Machine Learning 2023

Markov Chains on General Graphs

- ***Markov chains - Ubiquitous in statistics and learning.***
  - Markov Chain Monte Carlo (MCMC) for sampling.
  - Stochastic Gradient Descent (SGD) for distributed optimization.
- **Examples of Markov chains used on Graphs**
  - Simple Random Walk
  - Metropolis Hastings Random Walk

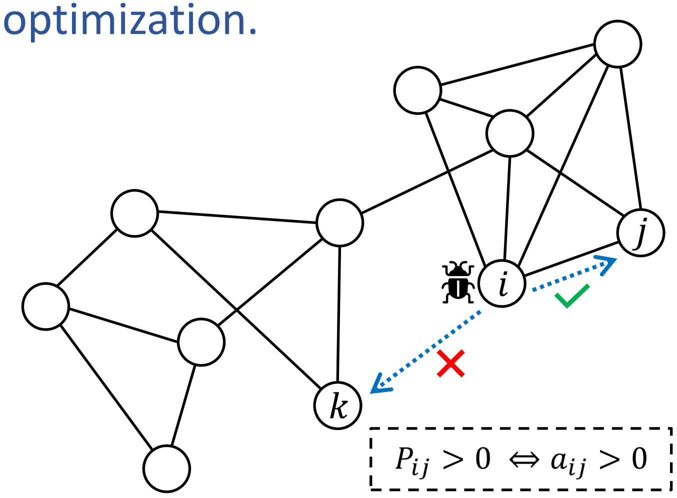


Undirected, connected graph
Nodes: $\{1,\cdots,N\}$
Adj. matrix $\boldsymbol{A} = \left[a_{ij}\right]_{i,j\in\{1,\cdots,N\}}$
where:
$a_{ij} > 0 \Leftrightarrow (i,j)$ is edge,
$a_{ij} = 0$ o/w

[1] Pierre Brémaud. *Markov chains Gibbs fields, Monte Carlo simulation, and Queues*. 2020.
[2] Sun, Tao, Yuejiao Sun, and Wotao Yin. "On Markov Chain Gradient Descent." *NeurIPS* (2018).

Markov Chains on General Graphs

- ***Markov chains - Ubiquitous in statistics and learning.***
  - Markov Chain Monte Carlo (MCMC) for sampling.
  - Stochastic Gradient Descent (SGD) for distributed optimization.
- **Examples of Markov chains used on Graphs**
  - Simple Random Walk
  - Metropolis Hastings Random Walk
- **Input parameters**
  - ‘Target’ probability measure $\boldsymbol{\mu} = [\mu_i]_{i \in \{1,\cdots,N\}}$
  - Transition probabilities $\boldsymbol{P} = [P_{ij}]_{i,j \in \{1,\cdots,N\}}$
  - Satisfy $\boldsymbol{\mu}^T \boldsymbol{P} = \boldsymbol{\mu}^T$ (Balance Equation)
- **Are usually time-reversible**
  - Satisfy $\mu_i P_{ij} = \mu_j P_{ji}$ for all $i, j \in \{1, \cdots, N\}$ (Detailed Balance Equation)



$P_{ij} > 0 \Leftrightarrow a_{ij} > 0$

[1] Pierre Brémaud. *Markov chains Gibbs fields, Monte Carlo simulation, and Queues*. 2020.
[2] Sun, Tao, Yuejiao Sun, and Wotao Yin. "On Markov Chain Gradient Descent." *NeurIPS* (2018).

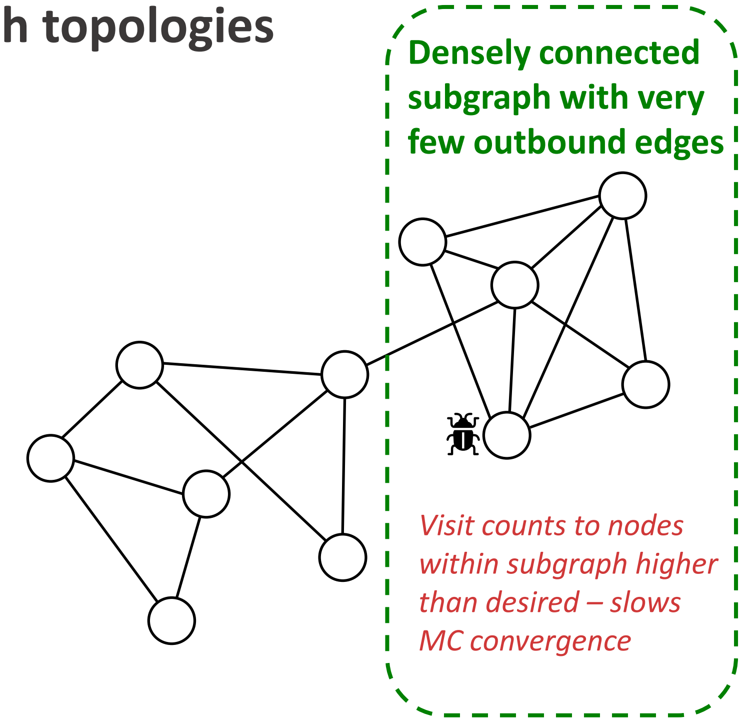
Markov Chains on General Graphs

- **MCMCs designed to be *Scale Invariance (S.I.)* and *Distributed***
  - Do not need to know exact probabilities μ_i's to compute P_{ij}'s
  - At most, only require knowing μ_i's up to a constant multiple, and only for neighbors of the current node (local information only) at any time step

- **Robust implementation with convergence guarantees**
  - S.I. allows graph to be explored on-the-fly; ergodicity guarantees convergence
  - Lead to widespread adoption of MC (e.g. MHRW) for sampling and optimization

Markov Chains on General Graphs

- **Classical Markov chains are victims of ‘bad’ graph topologies**
  - Can get ‘trapped’ within some subgraphs
  - Highly correlated samples

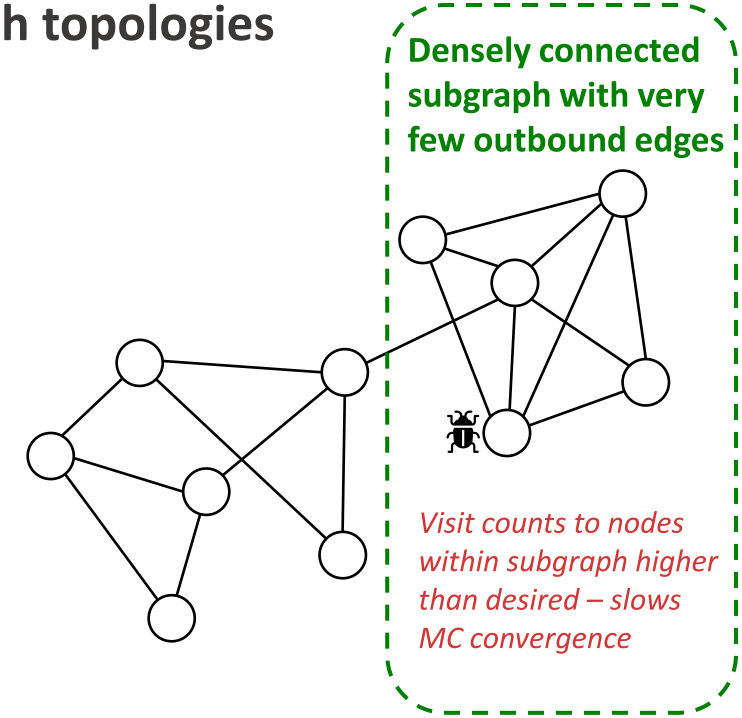


Densely connected subgraph with very few outbound edges

Visit counts to nodes within subgraph higher than desired – slows MC convergence

Markov Chains on General Graphs

- **Classical Markov chains are victims of ‘bad’ graph topologies**
  - Can get ‘trapped’ within some subgraphs
  - Highly correlated samples
- **Time-reversible Markov chains are slow**
  - Slower convergence to (target) stationary dist. $\boldsymbol{\mu}$
  - Non-reversible versions of the original Markov chains are known give better results



Densely connected subgraph with very few outbound edges

Visit counts to nodes within subgraph higher than desired – slows MC convergence

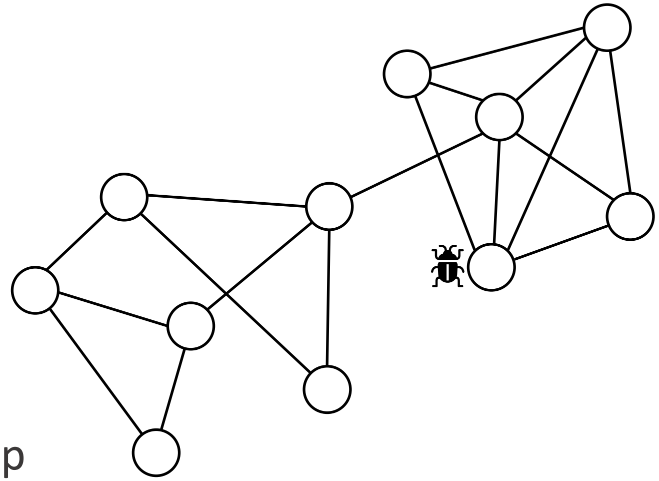
[1] Konstantin S Turitsyn, Michael Chertkov, and Marija Vucelja. Irreversible monte carlo algorithms for efficient sampling. Physica D: Nonlinear Phenomena, 240(4- 5):410–414, 2011.

[2] Andrieu, C. and Livingstone, S. Peskun–tierney ordering for markovian monte carlo: Beyond the reversible scenario. The Annals of Statistics, 49(4):1958–1981, 2021.

[3] Diaconis, P., Holmes, S., and Neal, R. M. Analysis of a nonreversible markov chain sampler. Annals of Applied Probability, pp. 726–752, 2000.

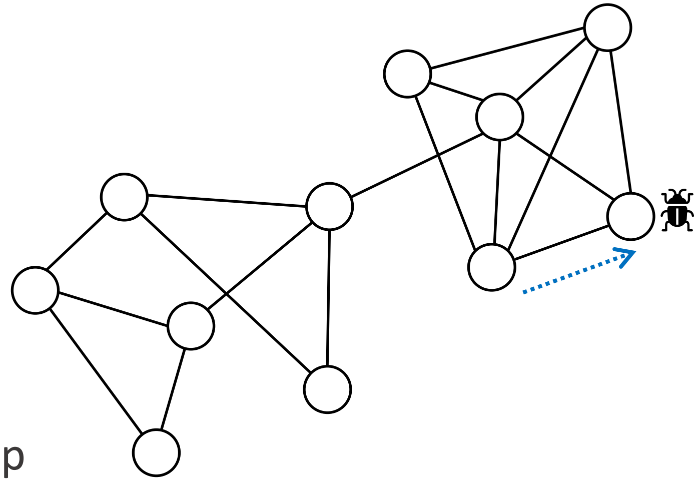
Markov Chains on General Graphs

- **Classical Markov chains are victims of ‘bad’ graph topologies**
  - Can get ‘trapped’ within some subgraphs
  - Highly correlated samples
- **Time-reversible Markov chains are slow**
  - Slower convergence to (target) stationary dist. $\boldsymbol{\mu}$
  - Non-reversible versions of the original Markov chains are known give better results
- **Non-backtracking approaches work better**
  - Avoids transitioning to node visited in previous step



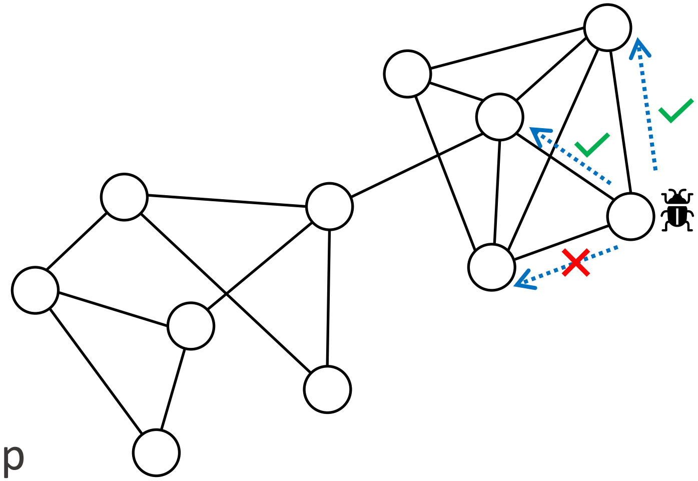
Markov Chains on General Graphs

- **Classical Markov chains are victims of ‘bad’ graph topologies**
  - Can get ‘trapped’ within some subgraphs
  - Highly correlated samples
- **Time-reversible Markov chains are slow**
  - Slower convergence to (target) stationary dist. $\boldsymbol{\mu}$
  - Non-reversible versions of the original Markov chains are known give better results
- **Non-backtracking approaches work better**
  - Avoids transitioning to node visited in previous step



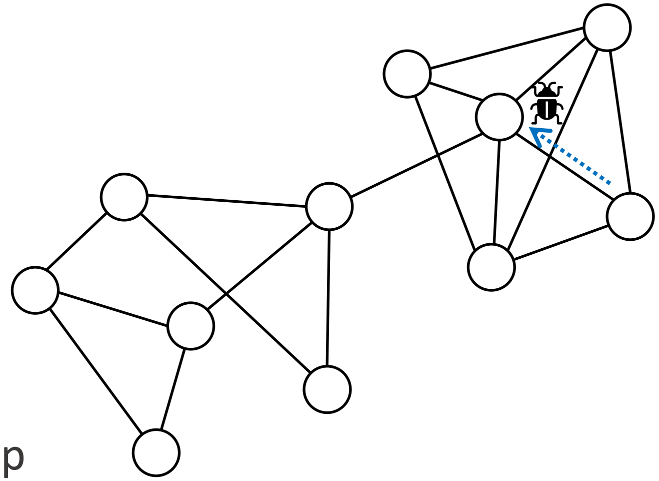
Markov Chains on General Graphs

- **Classical Markov chains are victims of ‘bad’ graph topologies**
  - Can get ‘trapped’ within some subgraphs
  - Highly correlated samples
- **Time-reversible Markov chains are slow**
  - Slower convergence to (target) stationary dist. $\boldsymbol{\mu}$
  - Non-reversible versions of the original Markov chains are known give better results
- **Non-backtracking approaches work better**
  - Avoids transitioning to node visited in previous step



Markov Chains on General Graphs

- **Classical Markov chains are victims of ‘bad’ graph topologies**
  - Can get ‘trapped’ within some subgraphs
  - Highly correlated samples
- **Time-reversible Markov chains are slow**
  - Slower convergence to (target) stationary dist. $\boldsymbol{\mu}$
  - Non-reversible versions of the original Markov chains are known give better results
- **Non-backtracking approaches work better**
  - Avoids transitioning to node visited in previous step
  - Non-reversible in the original state space (although still time-reversible in an augmented state space)
  - Smaller asymptotic variance of the estimator compared to base Markov chain

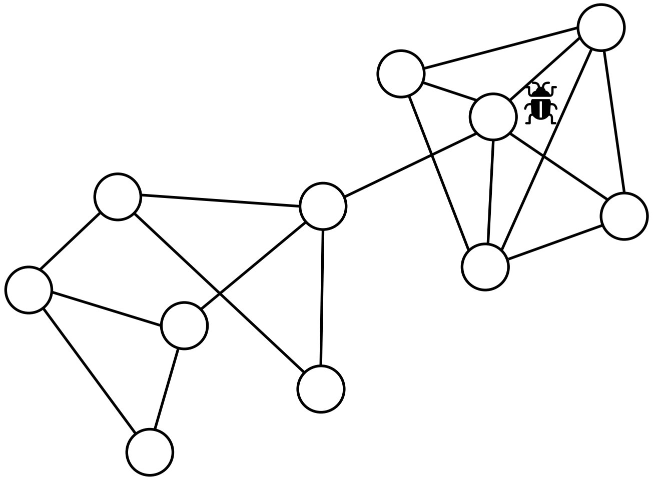


[1] Alon, N., Benjamini, I., Lubetzky, E., and Sodin, S. Nonbacktracking random walks mix faster. Communications in Contemporary Mathematics, 9(04):585–603, 2007

[2] Chul-Ho Lee, Xin Xu, and Do Young Eun. *Beyond Random Walk and Metropolis-Hastings Samplers: Why You Should Not Backtrack for Unbiased Graph Sampling*. In ACM SIGMETRICS 2012.

Random Walks with Self-Repellence

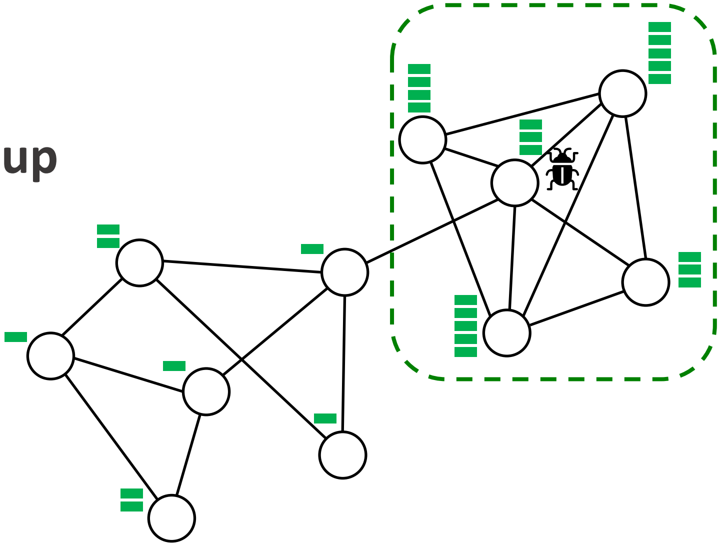
- **Non-backtracking walks are *weakly* *Self-Repellent***
  - Only interacting with their most recent past



Random Walks with Self-Repellence

- **Non-backtracking walks are *weakly* *Self-Repellent***
  - Only interacting with their most recent past

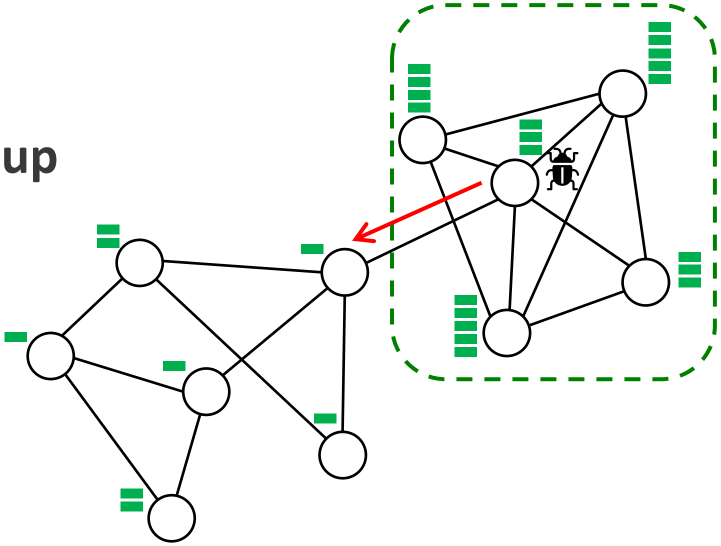
- **Can a *stronger* version of Self-Repellence speed up Markov Chains?**



Random Walks with Self-Repellence

- **Non-backtracking walks are *weakly* *Self-Repellent***
  - Only interacting with their most recent past

- **Can a *stronger* version of Self-Repellence speed up Markov Chains?**
  - Interact with entire history
  - Prioritize transitions to seldom visited nodes
  - Empirical measure still needs to converge to target distribution $\boldsymbol{\mu}$
  - Needs to be provably better than the original Markov chain in some sense



Our Contribution

Input: any Time-Reversible 'base' Markov Chain kernel P and target measure μ

- **We design a Self-Repellent Random Walk (SRRW), such that**
  - Empirical distribution converges almost surely to $\boldsymbol{\mu}$ (SLLN)
  - Achieves smaller asymptotic variance compared to base MC

- **First result for general, finite graphs used for algorithm design**
  - ***Self-repellent dynamics in literature:*** Focus on graphs such as d-dimensional grids; little to no knowledge of stationary probabilities – difficult to use as a basis for real world algorithm design.
  - ***Vertex reinforced Random walks:*** Closely related to our process, but key difference being that it is *self-attractive* (reinforced) instead of repellent; no control over stationary distribution.

[1] Balint Toth. *The "True" Self-Avoiding Walk with Bond Repulsion on Z: Limit Theorems*. The Annals of Probability, 23(4), 1995

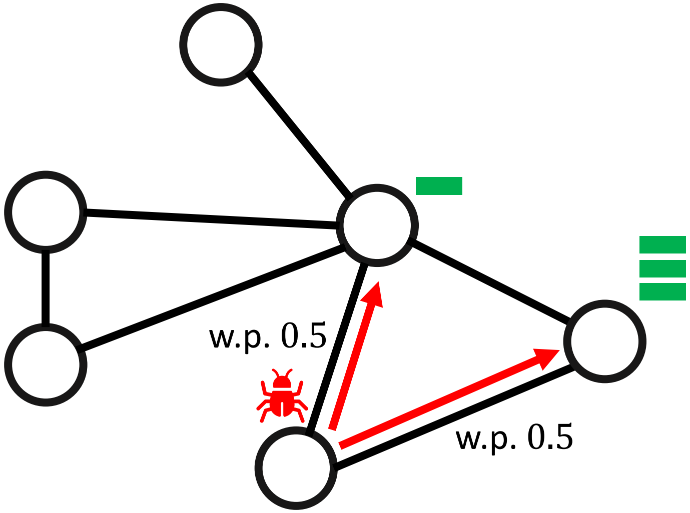
[2] Balint Veto and Balint Toth. *Self-repelling random walk with directed edges on Z*. Electronic Journal of Probability, 13(none), 2008.

[3] Robin Pemantle. *Vertex-reinforced random walk*. Probability Theory and Related Fields, 92(1), 1992.

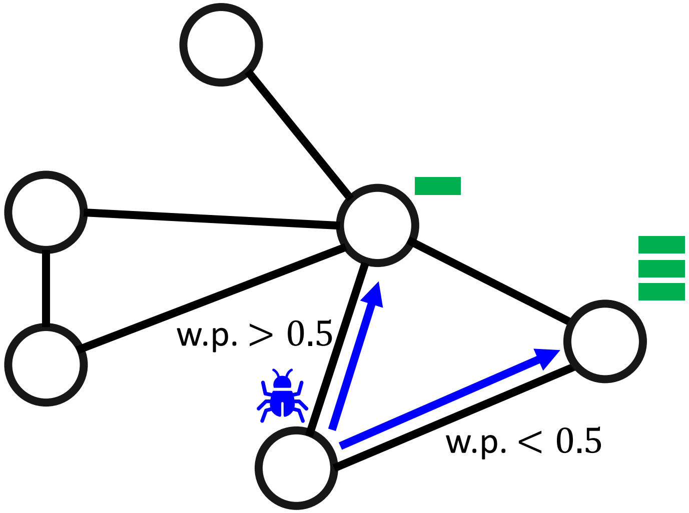
[4] Michel Benaimï, Olivier Raimond, and Bruno Schapira. *Strongly vertex-reinforced random-walk on the complete graph*. arXiv preprint arXiv:1208.6375, 2012.

Simple Random Walk → Self-Repellent Random Walk

- **Simple Random Walk (SRW):**
  - Equally likely to visit neighbouring nodes (unweighted graph).
- **Self Repellent Random Walk (SRRW):**
  - Needs a ‘base’ Markov chain as input (e.g. SRW)
  - Transition probability is a decreasing function of the visit count of a node.



Transition Probabilities for SRW



Transition probabilities for SRRW with SRW base chain

Simple Random Walk → Self-Repellent Random Walk

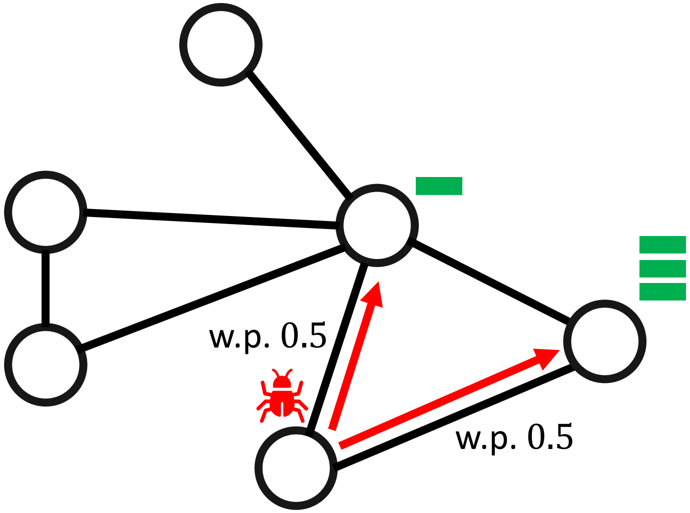
We say $\deg(i)$ = # neighbours of i. For all neighbours j of node i.

- **SRW**

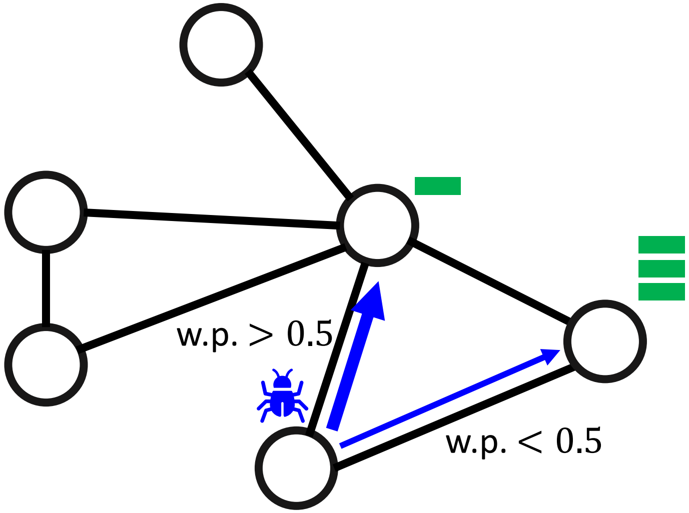
$$P(X_{n+1} = j \mid X_n = i) = \frac{1}{\deg(i)}$$

- **SRRW with SRW as 'base chain'**

$$P(X_{n+1} = j \mid X_n = i, X_{n-1}, \cdots, X_0) \propto \left(\frac{1 + \#visits\ to\ j}{\deg(j)}\right)^{-\alpha}$$



Transition Probabilities for SRW



Transition probabilities for SRRW with SRW base chain

Time-Reversible MC → Self-Repellent Random Walk

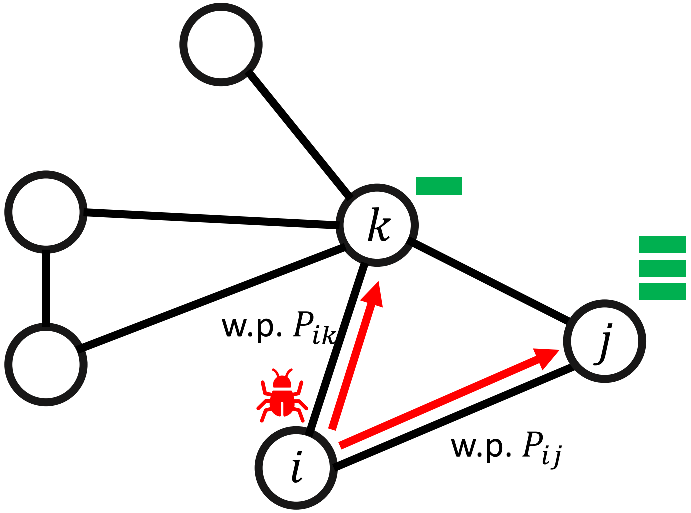
SRRW can be adapted for any time-reversible Markov chain also inheriting the S.I. property

- **Any Time-reversible Markov chain**

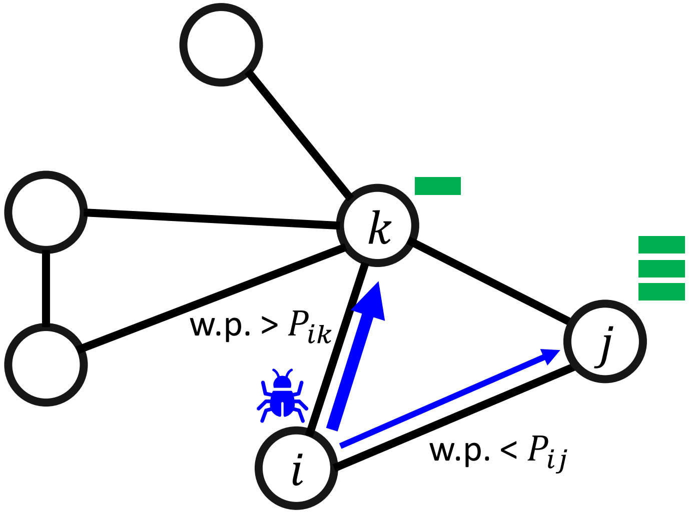
$$P(X_{n+1} = j \mid X_n = i) = P_{ij}$$

- **SRRW version**

$$P(X_{n+1} = j \mid X_n = i, X_{n-1}, \cdots, X_0) \propto P_{ij} \left(\frac{1 + \#visits\ to\ j}{\mu_j} \right)^{-\alpha}$$



Markov chain with transition kernel $\boldsymbol{P}$



SRRW version of Markov chain with kernel $\boldsymbol{P}$

Time-Reversible MC → Self-Repellent Random Walk

SRRW can be adapted for any time-reversible Markov chain also inheriting the S.I. property

- **Any Time-reversible Markov chain**

$$P(X_{n+1} = j \,|\, X_n = i) = P_{ij}$$

- **SRRW version**

$$P(X_{n+1} = j \,|\, X_n = i, X_{n-1}, \cdots, X_0) \propto P_{ij} \boxed{\left(\frac{1 + \#visits\ to\ j}{\mu_j}\right)^{-\alpha}}$$

Larger $\alpha > 0$ implies stronger self-repellence

- **Why polynomial form as shown?**
  - Only form for which the S.I. property of time-reversible chains is inherited
  - Key to robust implementation for any general graph

Self-Repellent Random Walk

- **Consider a stochastic process** $\{X_n, \mathbf{x}_n\}$ **taking values in** $[N] \times \Sigma$, which satisfy:

Set: $X_0 \in [N]$, and $\mathbf{x}_0 \in \mathrm{Int}(\Sigma)$ $(\text{e.g. } \mathbf{x}_0 = [1/N, \cdots, 1/N]^T)$

Draw: $X_{n+1} \sim K[\mathbf{x}_n]_{(X_n,\cdot)}$ (transition to $X_{n+1} \in \mathcal{N}(X_n)$)

Iterate: $\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{n+2}(\boldsymbol{\delta}_{X_{n+1}} - \mathbf{x}_n)$ (update empirical measure)

where for any $\mathbf{x} \in \mathrm{Int}(\Sigma)$,

$$K\boxed{[\mathbf{x}]}_{ij} \triangleq P(X_{n+1} = j \mid X_n = i, \mathbf{x}) = P_{ij}\left(\frac{x_j}{\mu_j}\right)^{-\alpha} \Bigg/ \sum_{k \in [N]} P_{ik}\left(\frac{x_k}{\mu_k}\right)^{-\alpha}$$

Transition probabilities are ***functions of probability distributions***.
In this case, a function of X_n's own historical empirical measure.

$[N] = \{1, \cdots, N\}$, and $\Sigma = \{\mathrm{z} \in [0,1]^N \mid 1^T \mathrm{z} = 1\}$ (probability simplex)

SRRW: Stochastic Dynamics

- **The matrix $\boldsymbol{K}[\mathbf{x}] = [K[\mathbf{x}]_{ij}] \in [0,1]^{N\times N}$ is a *nonlinear Markov* kernel. Ergodic for all $\mathbf{x} \in \mathrm{Int}(\Sigma)$.**
- **Can show there exists a unique stationary dist. $\boldsymbol{\pi}(\mathbf{x}) \in \mathrm{Int}(\Sigma)$ satisfying $\pi_i(\mathbf{x})K[\mathbf{x}]_{ij} = \pi_j(\mathbf{x})K[\mathbf{x}]_{ji}$ (detailed balance eqn.).**
- **Can decompose the iteration as**

$$\boxed{\mathbf{x}_{n+1} = \mathbf{x}_n - \frac{1}{n+2}\big(\boldsymbol{f}(\mathbf{x}_n) + \boldsymbol{\epsilon}(X_{n+1}, \mathbf{x}_n)\big)}$$

where $\boldsymbol{f}(\mathbf{x}_n) = \boldsymbol{\pi}(\mathbf{x}_n) - \mathbf{x}_n$ (mean field)

and $\boldsymbol{\epsilon}(X_{n+1}, \mathbf{x}_n) = \boldsymbol{\delta}_{X_{n+1}} - \boldsymbol{\pi}(\mathbf{x}_n)$ (noise)

Stochastic approximation (SA) with state dependent noise. Related to ODE:

$$\dot{\mathbf{x}}(t) = \boldsymbol{\pi}\big(\mathbf{x}(t)\big) - \mathbf{x}(t)$$

SRRW: Deterministic analysis

- **Can derive closed form of $\boldsymbol{\pi}(\mathbf{x}) = [\pi_i(\mathbf{x})]$, given $\forall i \in [n]$ by**

$$\pi_i(\mathbf{x}) = \frac{\sum_j \mu_j P_{ij} \left(\frac{x_i}{\boldsymbol{\mu}_i}\right)^{-\alpha} \left(\frac{x_j}{d_j}\right)^{-\alpha}}{\sum_k \sum_l \mu_k P_{kl} \left(\frac{x_k}{d_k}\right)^{-\alpha} \left(\frac{x_l}{d_l}\right)^{-\alpha}}$$

Theorem 1 (Global stability of ODE) For all $\alpha \geq 0$, $\mathbf{x}(0) \in \mathrm{Int}(\Sigma)$, we have

$$\mathbf{x}(t) \longrightarrow \boldsymbol{\mu} \quad \text{as} \quad t \to \infty,$$

where $\boldsymbol{\mu} = [\mu_i] \in \mathrm{Int}(\Sigma)$ is the target stationary distribution.

SRRW: Deterministic analysis

- **Can derive closed form of $\boldsymbol{\pi}(\mathbf{x}) = [\pi_i(\mathbf{x})]$, given $\forall i \in [n]$ by**

$$\pi_i(\mathbf{x}) = \frac{\sum_j \mu_j P_{ij} \left(\frac{x_i}{\boldsymbol{\mu}_i}\right)^{-\alpha} \left(\frac{x_j}{d_j}\right)^{-\alpha}}{\boxed{\sum_k \sum_l \mu_k P_{kl} \left(\frac{x_k}{d_k}\right)^{-\alpha} \left(\frac{x_l}{d_l}\right)^{-\alpha}} = \omega(\mathbf{x})}$$

Theorem 1 (Global stability of ODE) For all $\alpha \geq 0$, $\mathbf{x}(0) \in \mathrm{Int}(\Sigma)$, we have

$$\mathbf{x}(t) \longrightarrow \boldsymbol{\mu} \quad \text{as} \quad t \to \infty,$$

where $\boldsymbol{\mu} = [\mu_i] \in \mathrm{Int}(\Sigma)$ is the target stationary distribution.

- **Proof steps:**
  - Show $\boldsymbol{\pi}(\mathbf{x}) = \mathbf{x}$ has a unique solution, given by $\boldsymbol{\mu}$.
  - Show that $\omega(\mathbf{x})$ is a Lyapunov function.
  - Apply LaSalle's Invariance Principle to obtain convergence.

SRRW: Stochastic analysis

- **The ODE global stability via Lyapunov function help provide convergence results for the stochastic seq. of empirical measures $\{\mathbf{x}_n\}_{n\geq 0}$.**

Theorem 2 (SLLN and CLT) For all $\alpha \geq 0$, any $\mathbf{x}_0 \in \mathrm{Int}(\Sigma)$, and any $X_0 \in [N]$, we have

$$\mathbf{x}_n \longrightarrow \boldsymbol{\mu} \quad \text{as} \quad t \to \infty, \qquad\qquad \textit{almost surely}$$

$$\sqrt{n}(\mathbf{x}_n - \boldsymbol{\mu}) \longrightarrow N\big(\mathbf{0}, \boldsymbol{V}(\alpha)\big) \quad as\ t \to \infty, \qquad in\ dist.$$

where $N\big(\mathbf{0}, \boldsymbol{V}(\alpha)\big)$ is a normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{V}(\alpha)$, given by

$$\boldsymbol{V}(\alpha) = \sum_{k=1}^{N-1} \boxed{\frac{1}{2\alpha(1+\lambda_k)+1} \cdot \frac{1+\lambda_k}{1-\lambda_k} \mathbf{u}_k \mathbf{u}_k^T.}$$

Function of $\alpha > 0$ eigenvalues and eigenvectors of transition matrix $\boldsymbol{P}$

SRRW: Ordering of Asymptotic Variance

- **Full characterization of asymptotic variance of SRRW in Theorem 2 allows us to derive the following ordering result**
  - The $<_L$ denotes a Loewner ordering of two matrices

Corollary 3 For any $\alpha_1 > \alpha_2 > 0$, we have

$$\boldsymbol{V}(\alpha_1) <_L \boldsymbol{V}(\alpha_2) <_L \boldsymbol{V}(0)$$

SRRW: Ordering of Asymptotic Variance

- **Full characterization of asymptotic variance of SRRW in Theorem 2 allows us to derive the following ordering result**
  - The $<_L$ denotes a Loewner ordering of two matrices

Corollary 3 For any $\alpha_1 > \alpha_2 > 0$, we have

$$\boldsymbol{V}(\alpha_1) <_L \boldsymbol{V}(\alpha_2) <_L \boldsymbol{V}(0)$$

- **Upper bound on asymptotic variance for MCMC sampling**

Corollary 4 (Sampling variance) For any $\alpha > 0$, and any bounded scalar valued function $g: [N] \to \mathrm{R}$ we have

Estimator variance of SRRW

Estimator variance of base MC

$$\frac{\mathbf{g}^T \boldsymbol{V}(\alpha)\mathbf{g}}{\mathbf{g}^T \boldsymbol{V}(0)\mathbf{g}} \leq O(1/\alpha)$$

where $\mathbf{g} = [g(1), \cdots g(N)]^T$.

SRRW: Ordering of Asymptotic Variance

- **SRRW variance goes to zero – large enough α can eventually beat *i.i.d.* sampler**
  - Typical *i.i.d.* sampler achieves smaller variance than random walkers on graph which needs to adhere to graph topology while walking
  - SRRW with sufficiently large $\alpha > 0$ is a rare example of random walker which can beat *i.i.d.* sampler despite the graph constraints

Corollary 4 (Sampling variance) For any $\alpha > 0$, and any bounded scalar valued function $g: [N] \to \mathrm{R}$ we have

Estimator variance of SRRW

Estimator variance of base MC

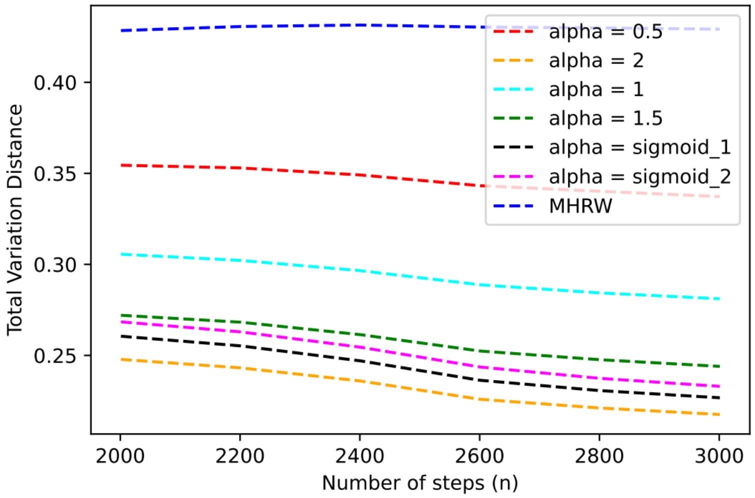
$$\frac{\mathbf{g}^T \boldsymbol{V}(\alpha)\mathbf{g}}{\mathbf{g}^T \boldsymbol{V}(0)\mathbf{g}} \leq \boxed{O(1/\alpha) \to 0, \text{ as } \alpha \to \infty}$$

where $\mathbf{g} = [g(1), \cdots g(N)]^T$.

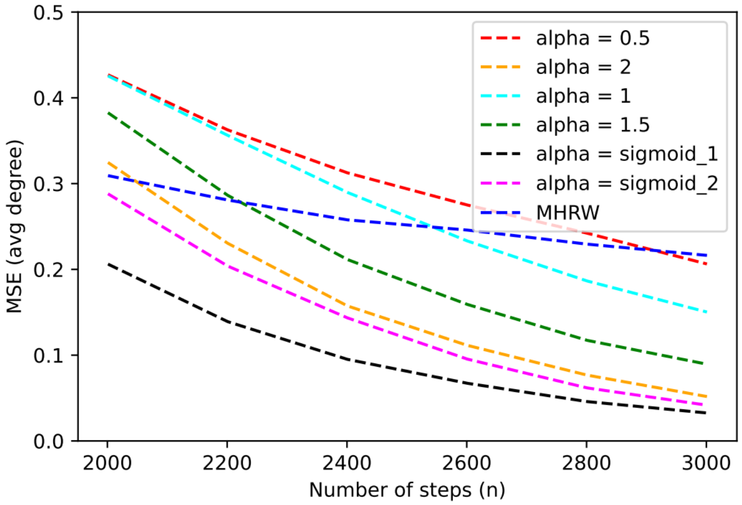
Ending Remarks

- **Nonlinearity of the transition kernel is key**
  - Nonlinearity induced via self-interactions can be used for effective algorithm design
  - Allows us to achieve asymptotically minimal sampling variance

- **Numerical simulations over different combinations of $\alpha > 0$ show its performance benefits and confirm our theoretical findings**



(a) Convergence of $\mathbf{x}_n$ to the uniform distribution.



(b) Convergence of $\psi_n(g)$ to the ground truth $\mathbf{g}^T\mathbf{1}/N$.