Comparison of Random Walk based techniques for estimating network averages

Jithin K. Sreedharan
INRIA, France

Konstantin Avrachenkov
INRIA, France

Vivek S. Borkar
IIT Bombay, India

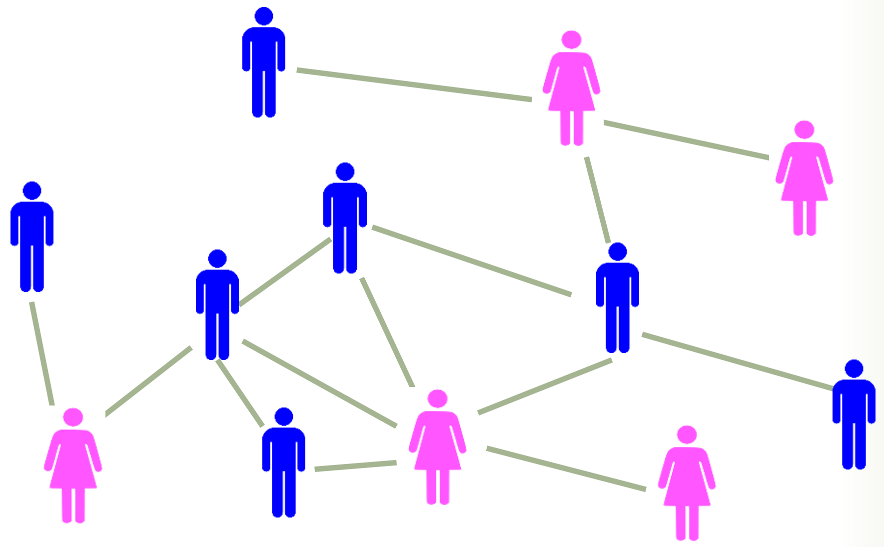
Arun Kadavankandy
INRIA, France

CSoNet 2016, August 2

Motivation

- Estimation in Online Social Network (OSN)
- Example:
  What proportion of a population supports a given political party?

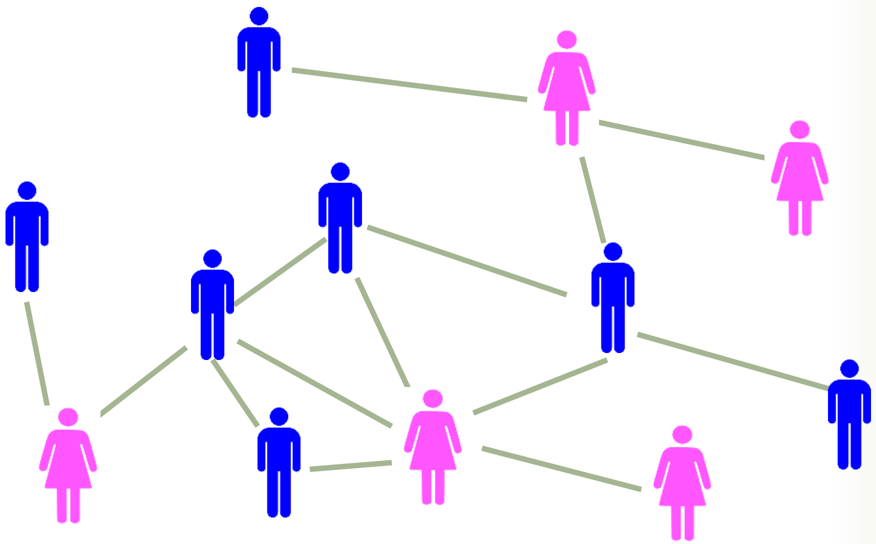
  How young a given social network is?



Motivation

- Estimation in Online Social Network (OSN)
- Example:
  What proportion of a population supports a given political party?

  How young a given social network is?



Easy to answer if the graph is fully known beforehand

What if the network is not known?

- Can only crawl network
- Few queries

Problem definition

Problem definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Problem definition

- Undirected graph

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

Problem definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Undirected graph
- Nodes have labels

Problem definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Undirected graph
- Nodes have labels
- Large graph

Problem definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Undirected graph
- Nodes have labels
- Large graph

Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$

Problem definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Undirected graph
- Nodes have labels
- Large graph

Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$

- Graph is unknown

Problem definition

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- Undirected graph
- Nodes have labels
- Large graph

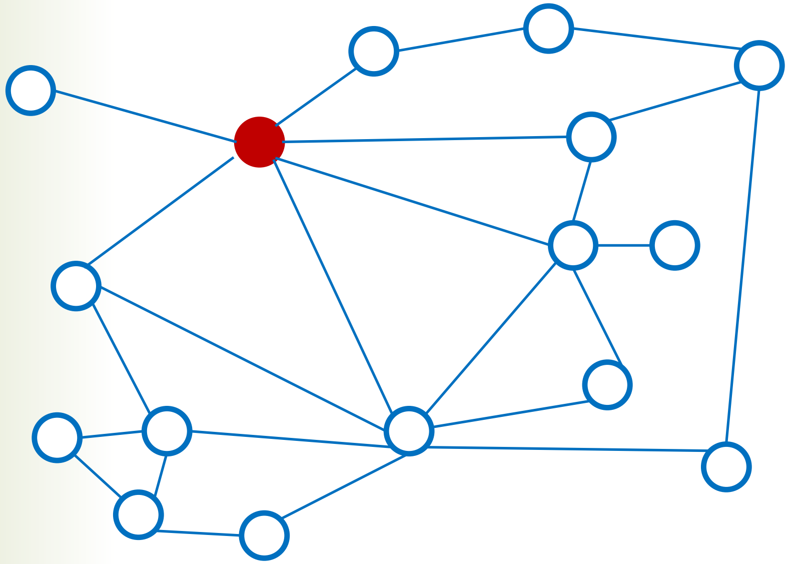
Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$

- Graph is unknown
- Only local information available

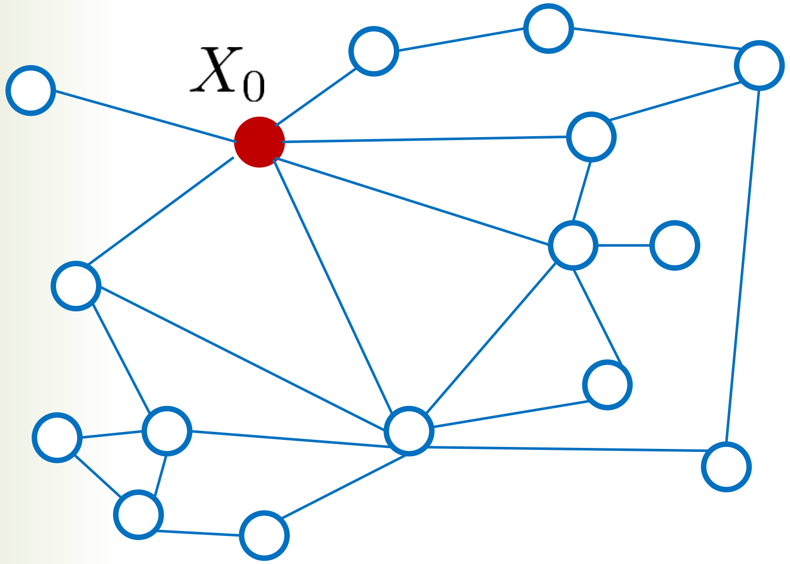
$\left\{ \begin{array}{l} \text{Seed nodes and their neighbor IDs} \\ \text{Query (visit) a neighbor} \\ \text{Visited nodes and their neighbor IDs} \end{array} \right.$

Random Walk based estimation

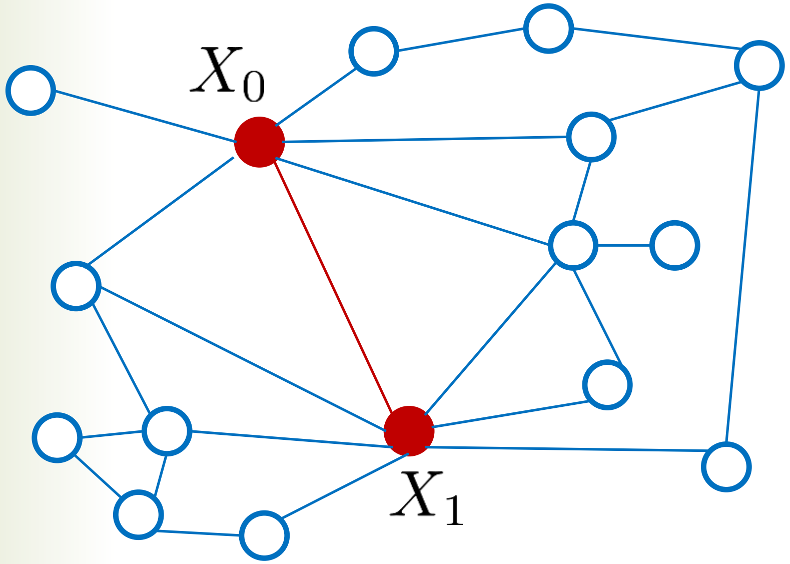
Random Walk based estimation



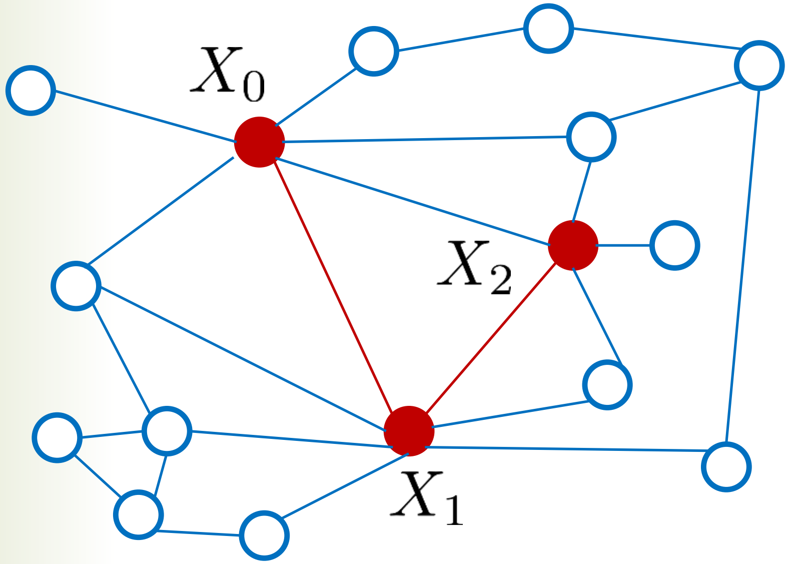
Random Walk based estimation



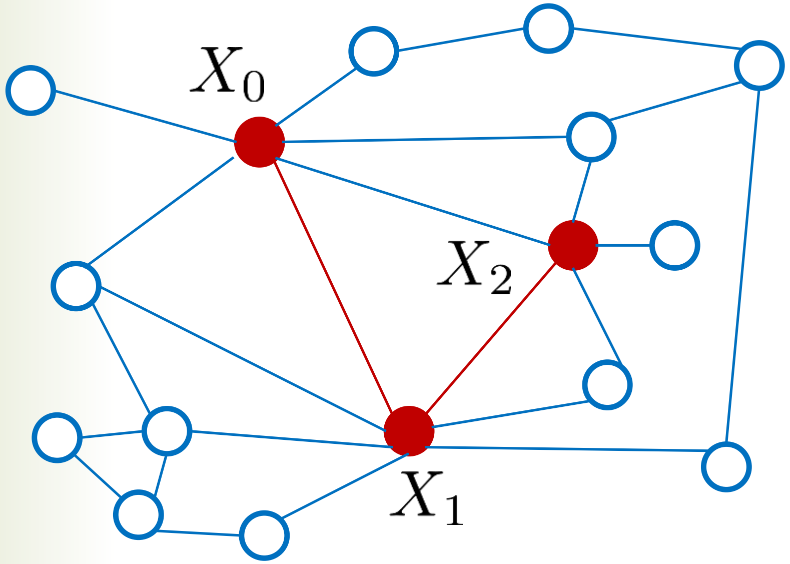
Random Walk based estimation



Random Walk based estimation

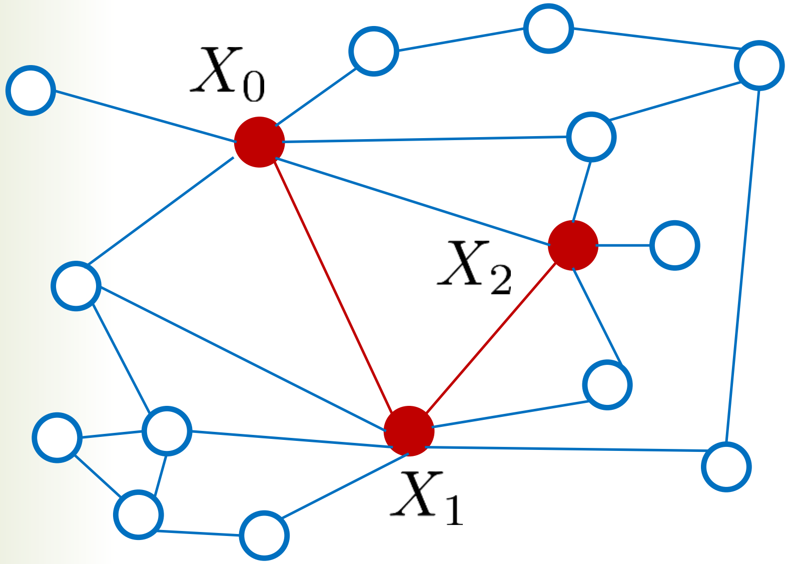


Random Walk based estimation



Random walk $\{X_k\}_{k\geq 0}$ has unique stationary distribution $\{\pi_i\}_{i=1}^{n}$ if graph G is connected and non-bipartite

Random Walk based estimation

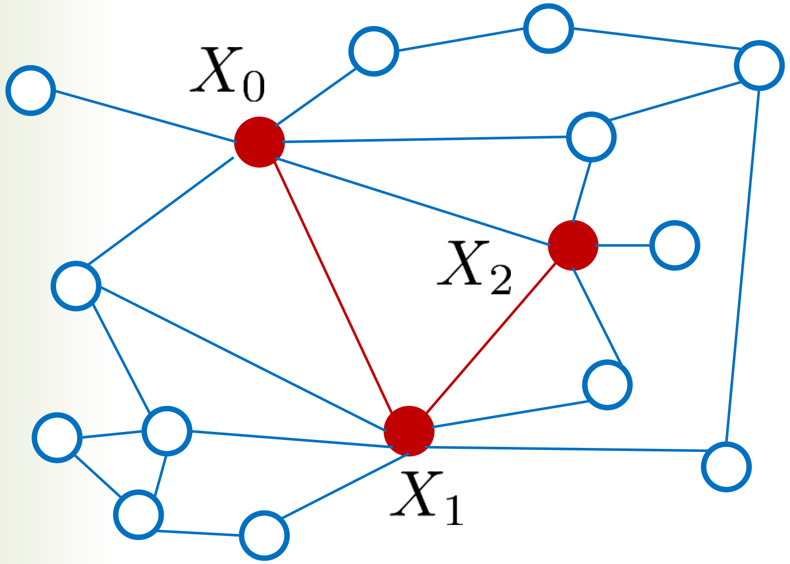


Random walk $\{X_k\}_{k\geq 0}$ has unique stationary distribution $\{\pi_i\}_{i=1}^n$ if graph G is connected and non-bipartite

- Goal:

  Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u)$

Random Walk based estimation



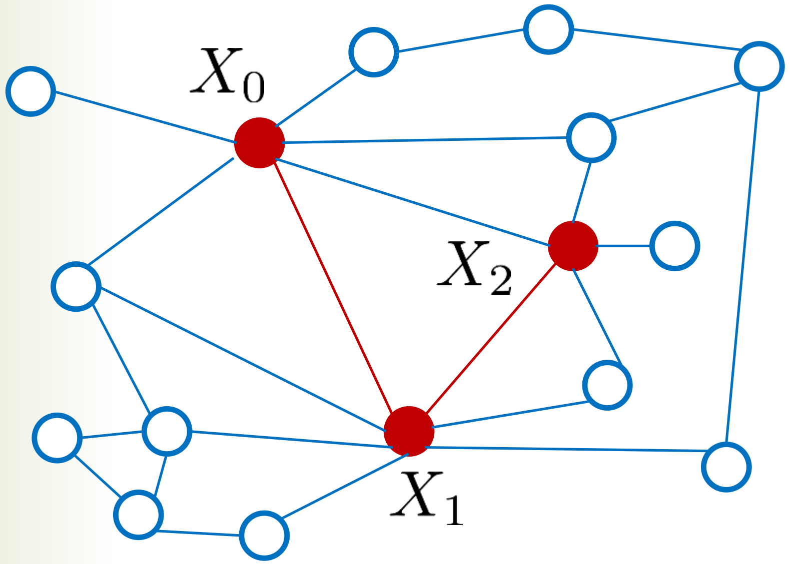
Random walk $\{X_k\}_{k\geq 0}$ has unique stationary distribution $\{\pi_i\}_{i=1}^{n}$ if graph G is connected and non-bipartite

- Goal:

  Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u)$

- How: Ergodic theorem

Random Walk based estimation



Random walk $\{X_k\}_{k \geq 0}$ has unique stationary distribution $\{\pi_i\}_{i=1}^{n}$ if graph G is connected and non-bipartite

- Goal:

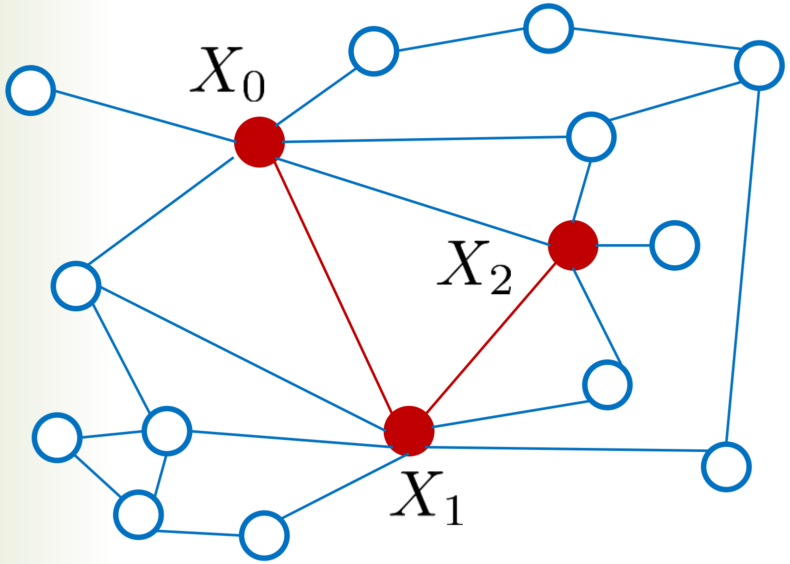
  Estimate $\mu(\mathcal{G}) = \dfrac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$

- How: Ergodic theorem

  For any initial distribution,

$$\frac{1}{n} \sum_{k=0}^{n} f(X_k) \to \sum_{u \in \mathcal{V}} \pi_u f(u) \quad \text{a.s.}$$

Random Walk based estimation



Random walk $\{X_k\}_{k\geq 0}$ has unique stationary distribution $\{\pi_i\}_{i=1}^{n}$ if graph G is connected and non-bipartite

- Goal:

  Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u)$

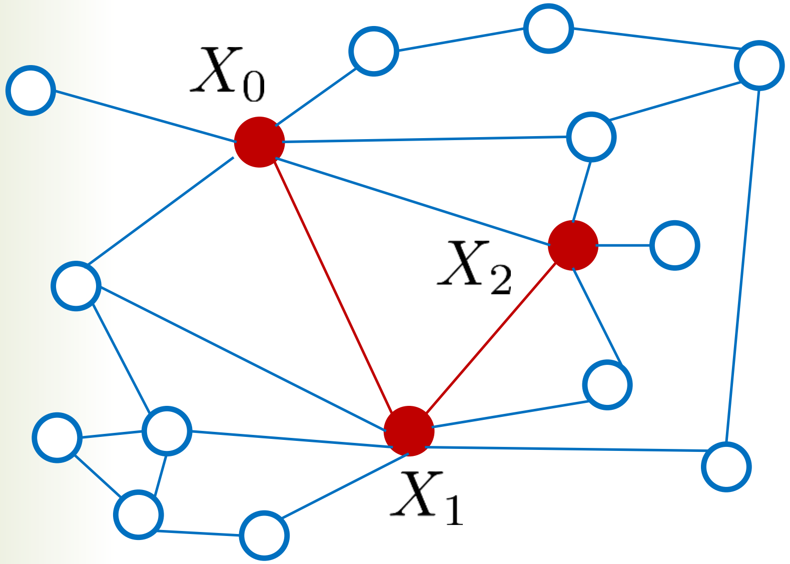
- How: Ergodic theorem

  For any initial distribution,

$$\frac{1}{n}\sum_{k=0}^{n} f(X_k) \to \sum_{u\in\mathcal{V}} \pi_u f(u) \quad \text{a.s.}$$

How to make $\pi_u = \frac{1}{|\mathcal{V}|}$?

Random Walk based estimation



Random walk $\{X_k\}_{k\geq 0}$ has unique stationary distribution $\{\pi_i\}_{i=1}^n$ if graph G is connected and non-bipartite

- Goal:

  Estimate $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u)$

- How: Ergodic theorem

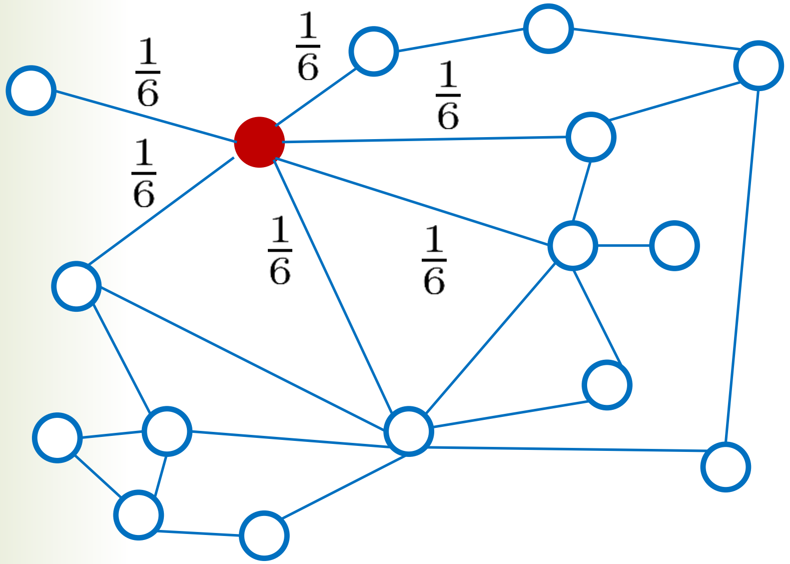
  For any initial distribution,

$$\frac{1}{n}\sum_{k=0}^{n} f(X_k) \to \sum_{u\in\mathcal{V}} \pi_u f(u) \quad \text{a.s.}$$

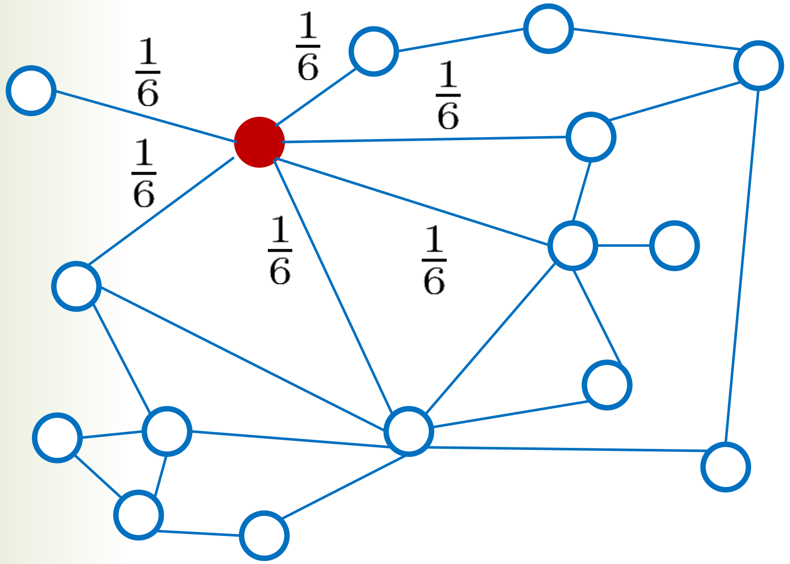
How to make $\pi_u = \frac{1}{|\mathcal{V}|}$? How to compare different random walks?

Respondent Driven Sampling

Respondent Driven Sampling



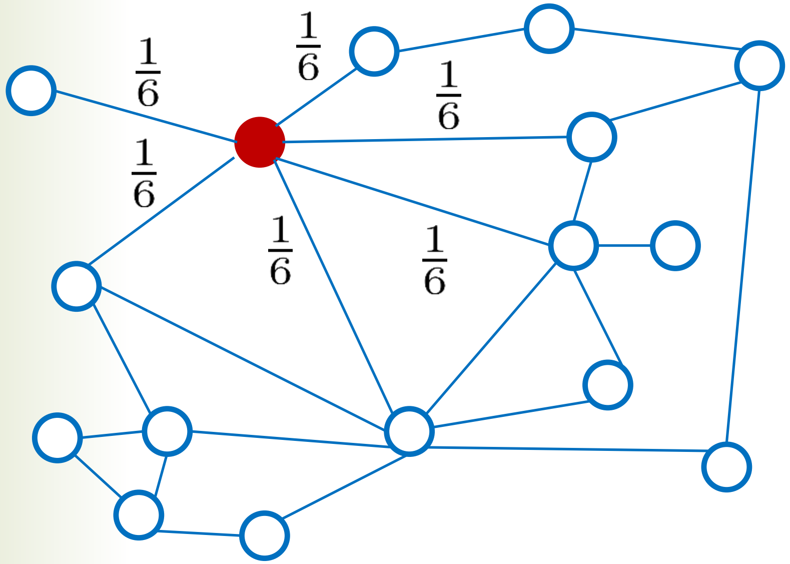
Respondent Driven Sampling



- With re-weighting the function f

$$f'(u) = \frac{f(u)}{|\mathcal{V}|\,\pi_u}$$

Respondent Driven Sampling

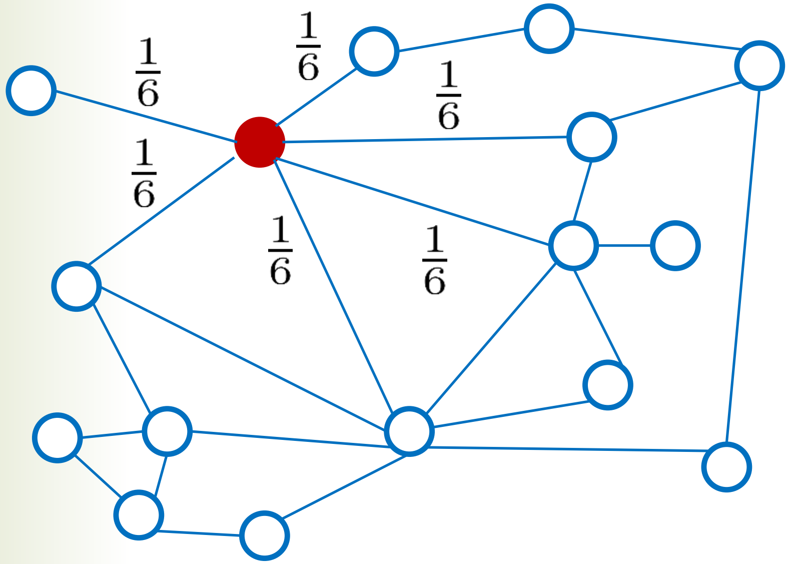


- With re-weighting the function f

$$f'(u) = \frac{f(u)}{|\mathcal{V}|\,\pi_u}$$

Requires knowledge of no. of nodes and no. of edges

Respondent Driven Sampling



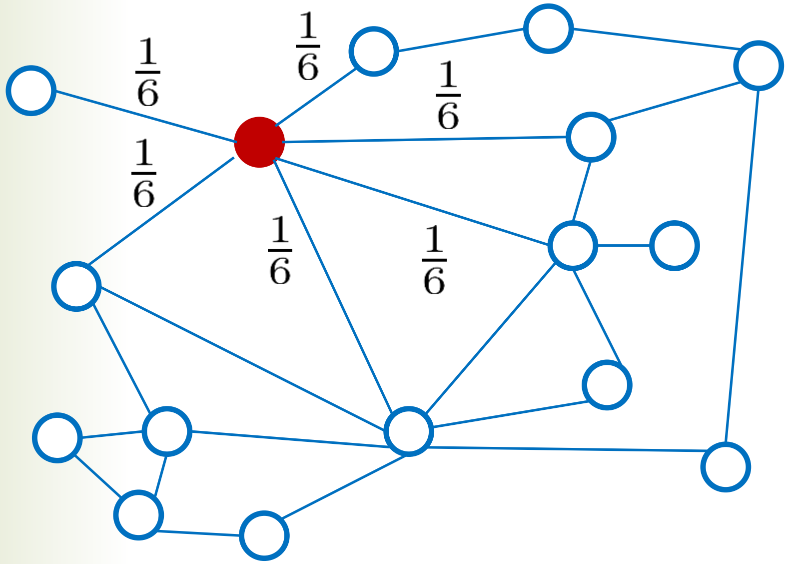
- With re-weighting the function f

$$f'(u) = \frac{f(u)}{|\mathcal{V}|\,\pi_u}$$

Requires knowledge of no. of nodes and no. of edges

Estimator: $$\frac{1}{\sum_{k=1}^{n} \frac{1}{\deg(X_k)}} \sum_{k=1}^{n} \frac{f(X_k)}{\deg(X_k)}$$

Respondent Driven Sampling



- With re-weighting the function f

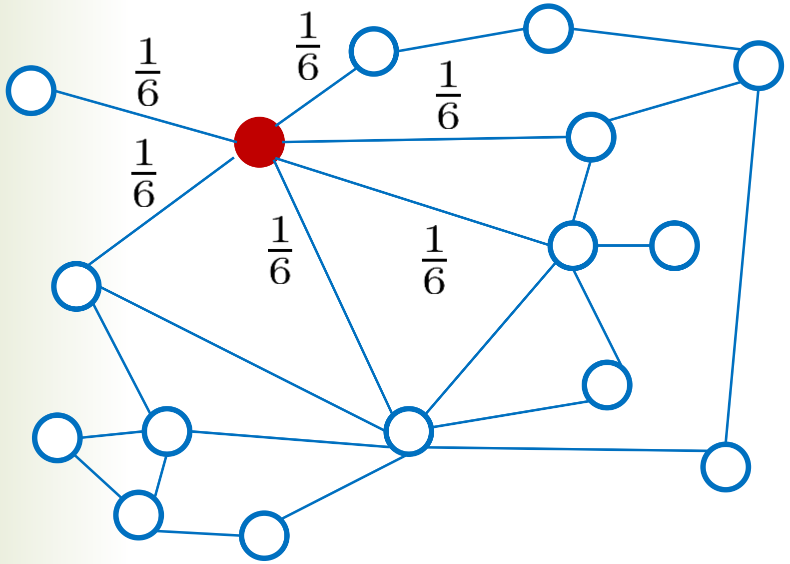
$$f'(u) = \frac{f(u)}{|\mathcal{V}|\, \pi_u}$$

Requires knowledge of no. of nodes and no. of edges

Estimator: $$\frac{1}{\sum_{k=1}^{n} \frac{1}{\deg(X_k)}} \sum_{k=1}^{n} \frac{f(X_k)}{\deg(X_k)}$$

$$\frac{|\mathcal{V}|}{2\,|\mathcal{E}|}$$

Respondent Driven Sampling



- With re-weighting the function f

$$f'(u) = \frac{f(u)}{|\mathcal{V}|\,\pi_u}$$

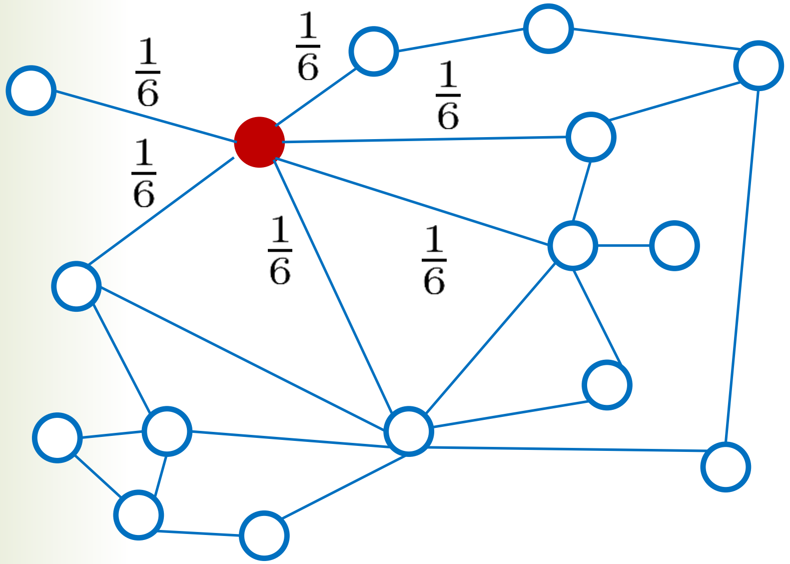
Requires knowledge of no. of nodes and no. of edges

$$\frac{1}{2|\mathcal{E}|}\sum_{u\in\mathcal{V}} f(u)$$

Estimator: $$\frac{1}{\sum_{k=1}^{n}\frac{1}{\deg(X_k)}}\sum_{k=1}^{n}\frac{f(X_k)}{\deg(X_k)}$$

$$\frac{|\mathcal{V}|}{2\,|\mathcal{E}|}$$

Respondent Driven Sampling



- With re-weighting the function f

$$f'(u) = \frac{f(u)}{|\mathcal{V}|\,\pi_u}$$

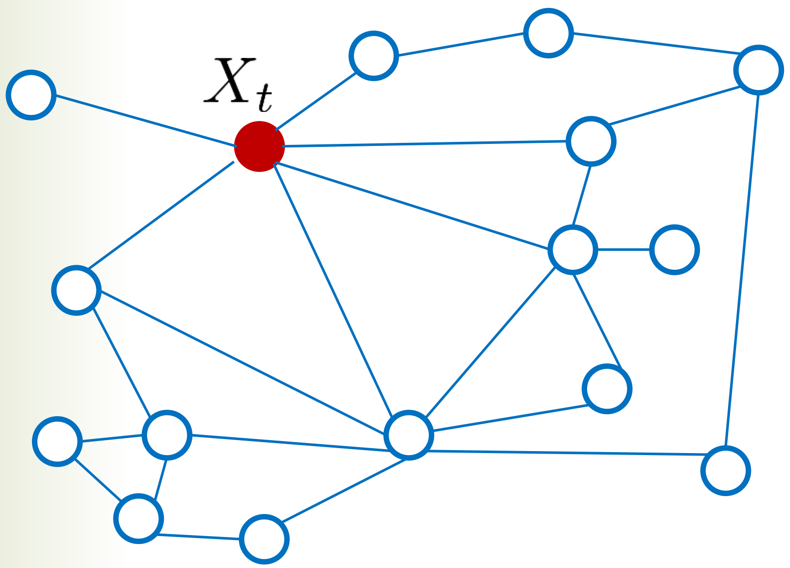
Requires knowledge of no. of nodes and no. of edges

$$\frac{1}{2|\mathcal{E}|}\sum_{u\in\mathcal{V}} f(u)$$

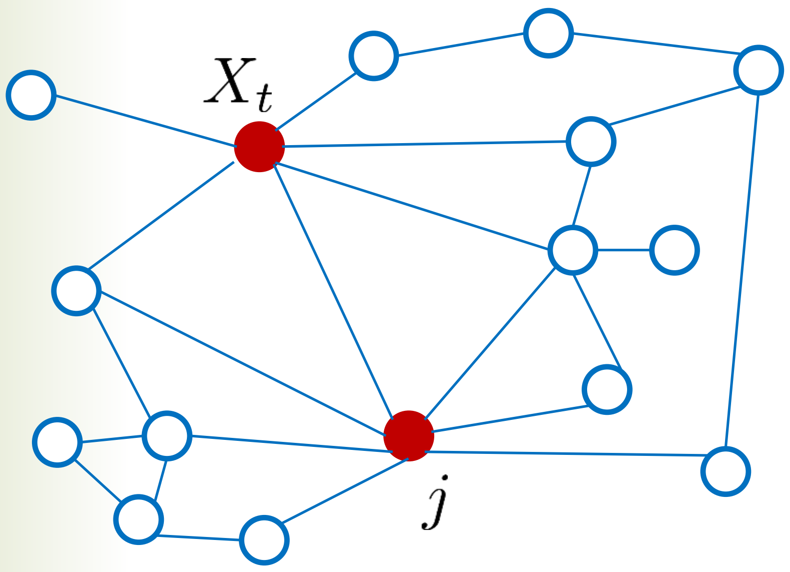
Estimator: $$\frac{1}{\sum_{k=1}^{n}\frac{1}{\deg(X_k)}}\sum_{k=1}^{n}\frac{f(X_k)}{\deg(X_k)} \quad \to \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u) \text{ a.s.}$$

$$\frac{|\mathcal{V}|}{2\,|\mathcal{E}|}$$

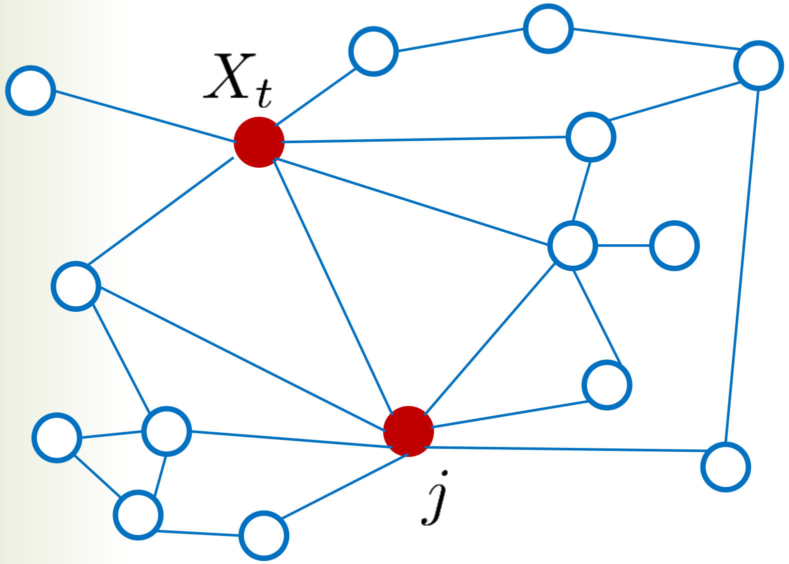
Metropolis Hastings Sampling



Metropolis Hastings Sampling

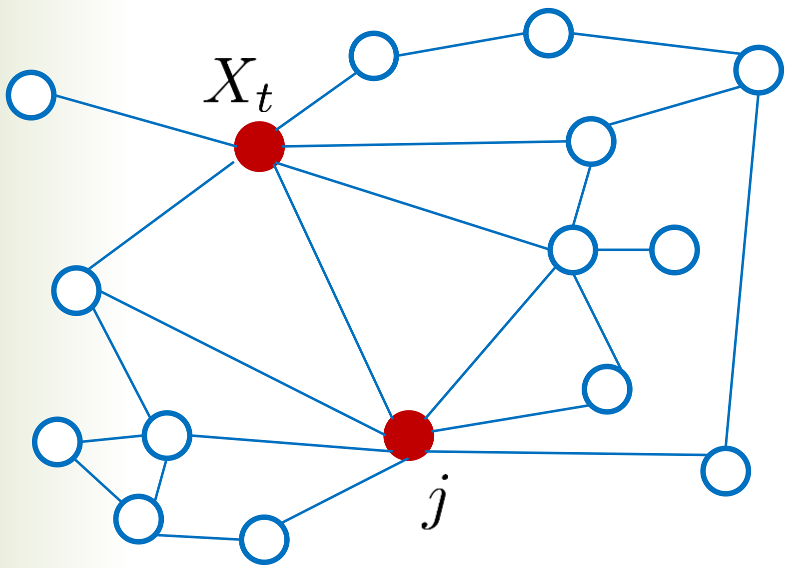


Metropolis Hastings Sampling



$$\Pr(\text{head}) := \min\left\{1, \frac{\deg(X_t)}{\deg(j)}\right\}$$

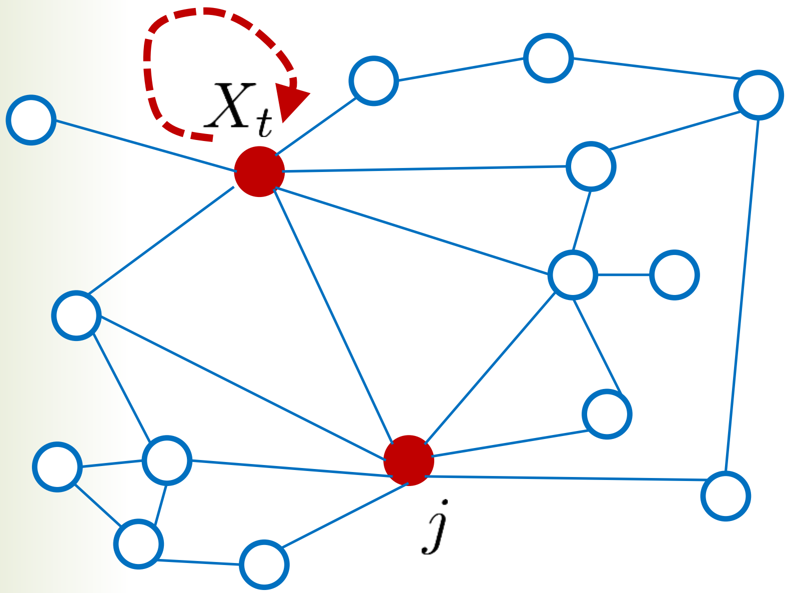
Metropolis Hastings Sampling



$$\Pr(\text{head}) := \min\left\{1, \frac{\deg(X_t)}{\deg(j)}\right\}$$

If head appears: move to *j*

Metropolis Hastings Sampling

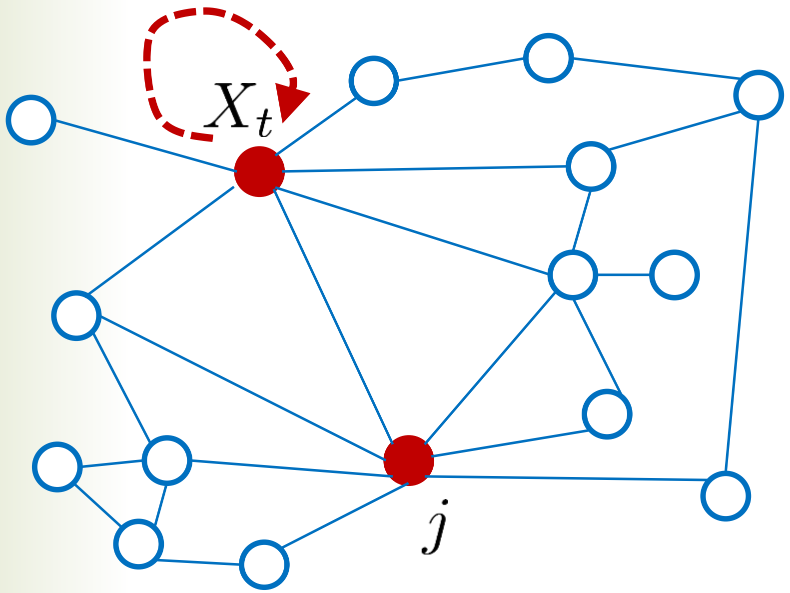


$$\Pr(\text{head}) := \min\left\{1, \frac{\deg(X_t)}{\deg(j)}\right\}$$

If head appears: move to j

If tail appears: stays at X_t

Metropolis Hastings Sampling



$$\Pr(\text{head}) := \min\left\{1, \frac{\deg(X_t)}{\deg(j)}\right\}$$

If head appears: move to j

If tail appears: stays at X_t

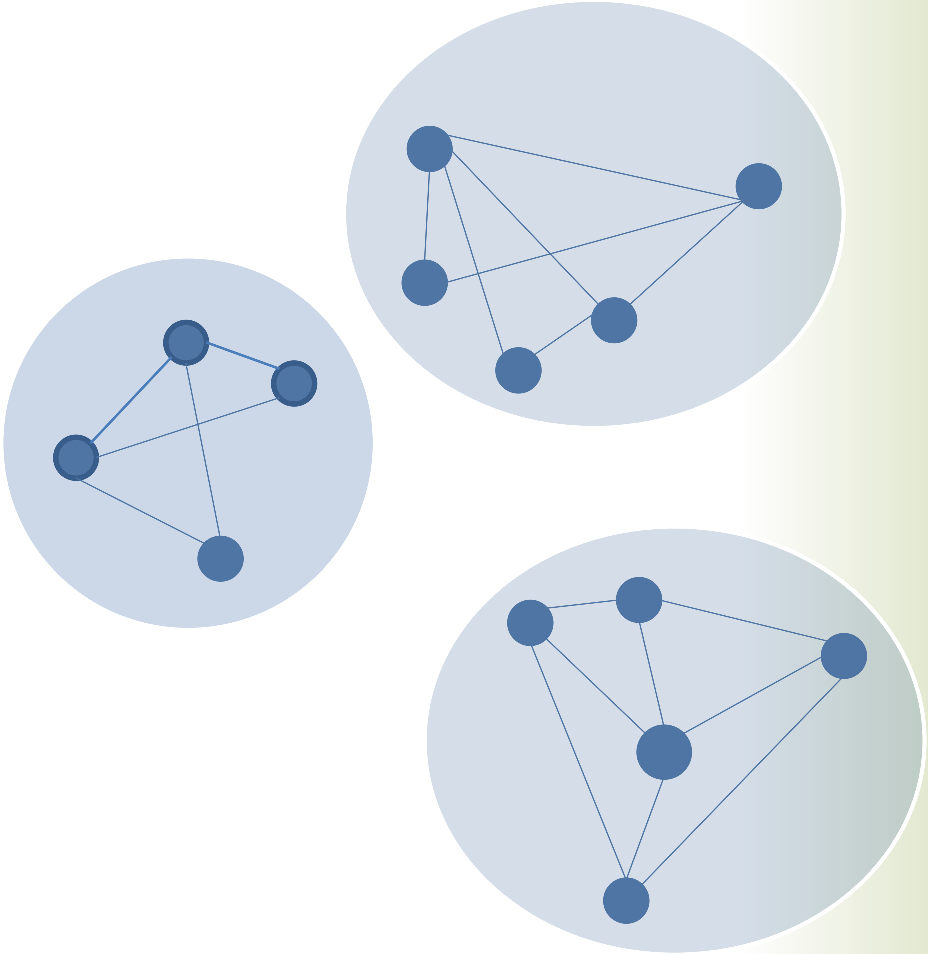
For any initial distribution,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \rightarrow \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u) \text{ a.s.}$$

Reinforcement Learning technique

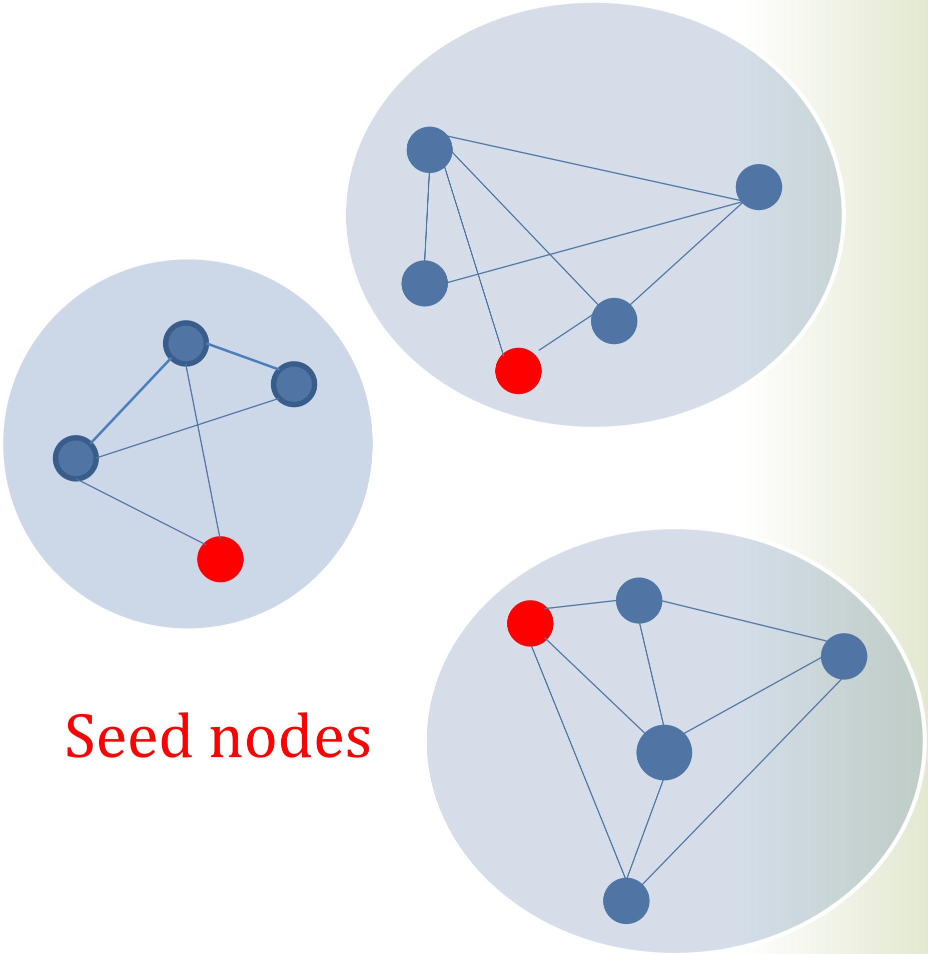
Reinforcement Learning technique

- Graph not necessarily connected or
  has included connected components of interest



Reinforcement Learning technique

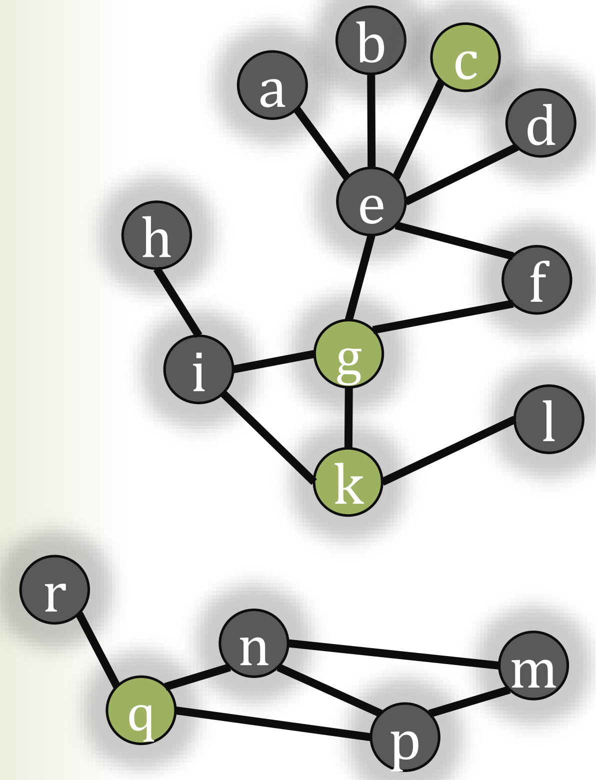
- Graph not necessarily connected or has included connected components of interest
- Few seed nodes



Idea of tours

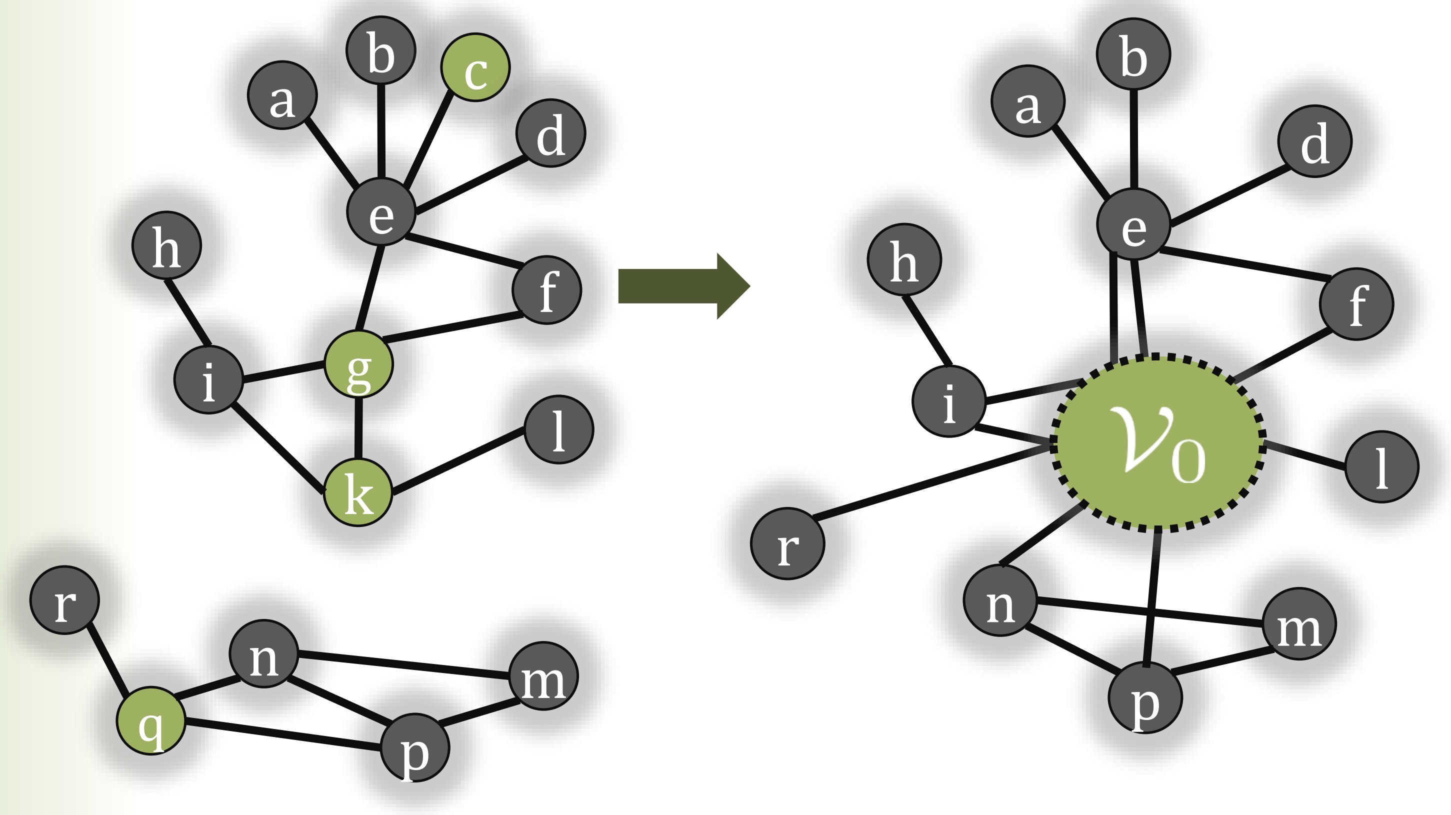
Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



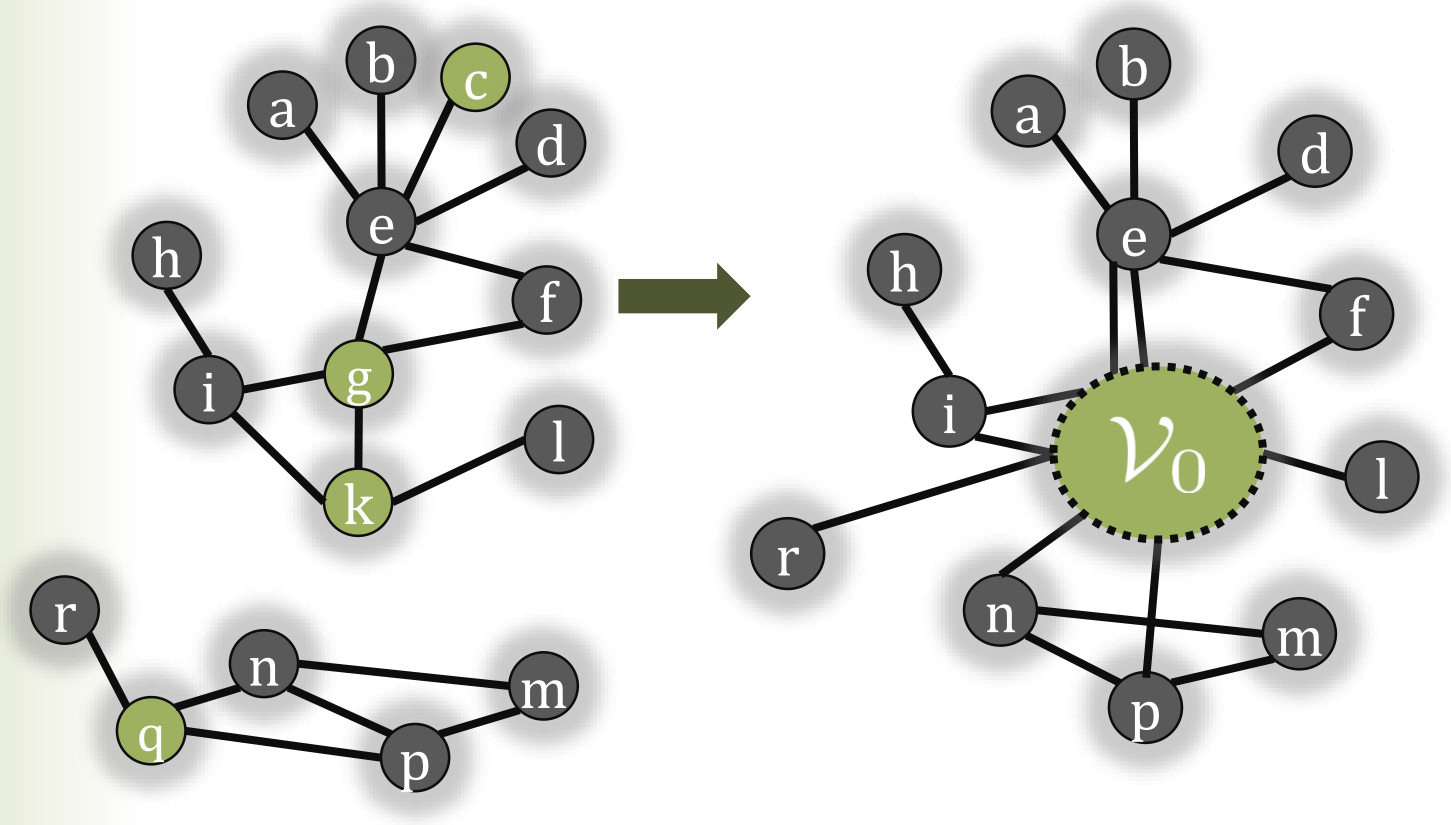
Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



Idea of tours

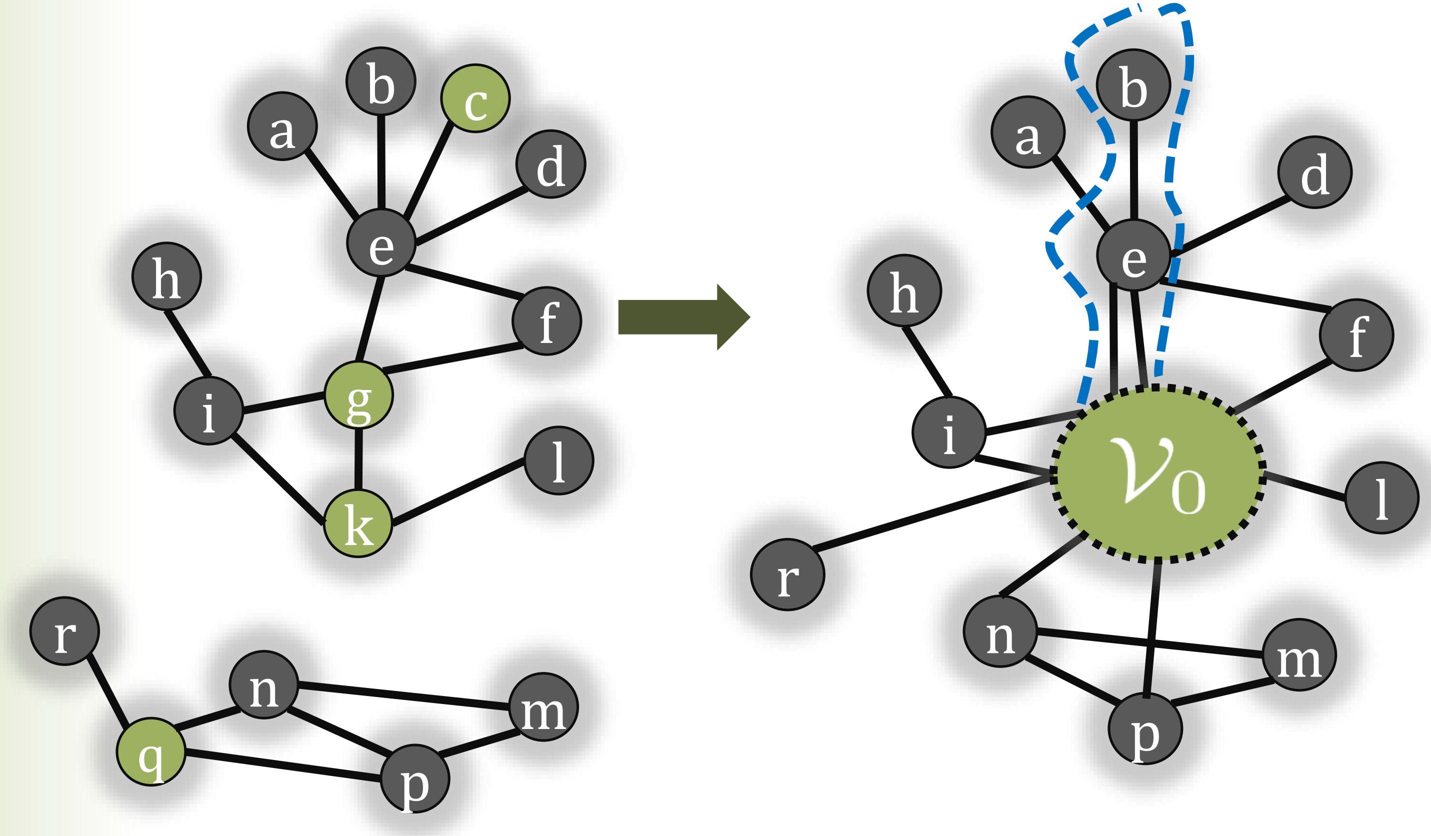
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



Sample
$z \sim \text{Uniform}(\mathcal{V}_0)$

Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$

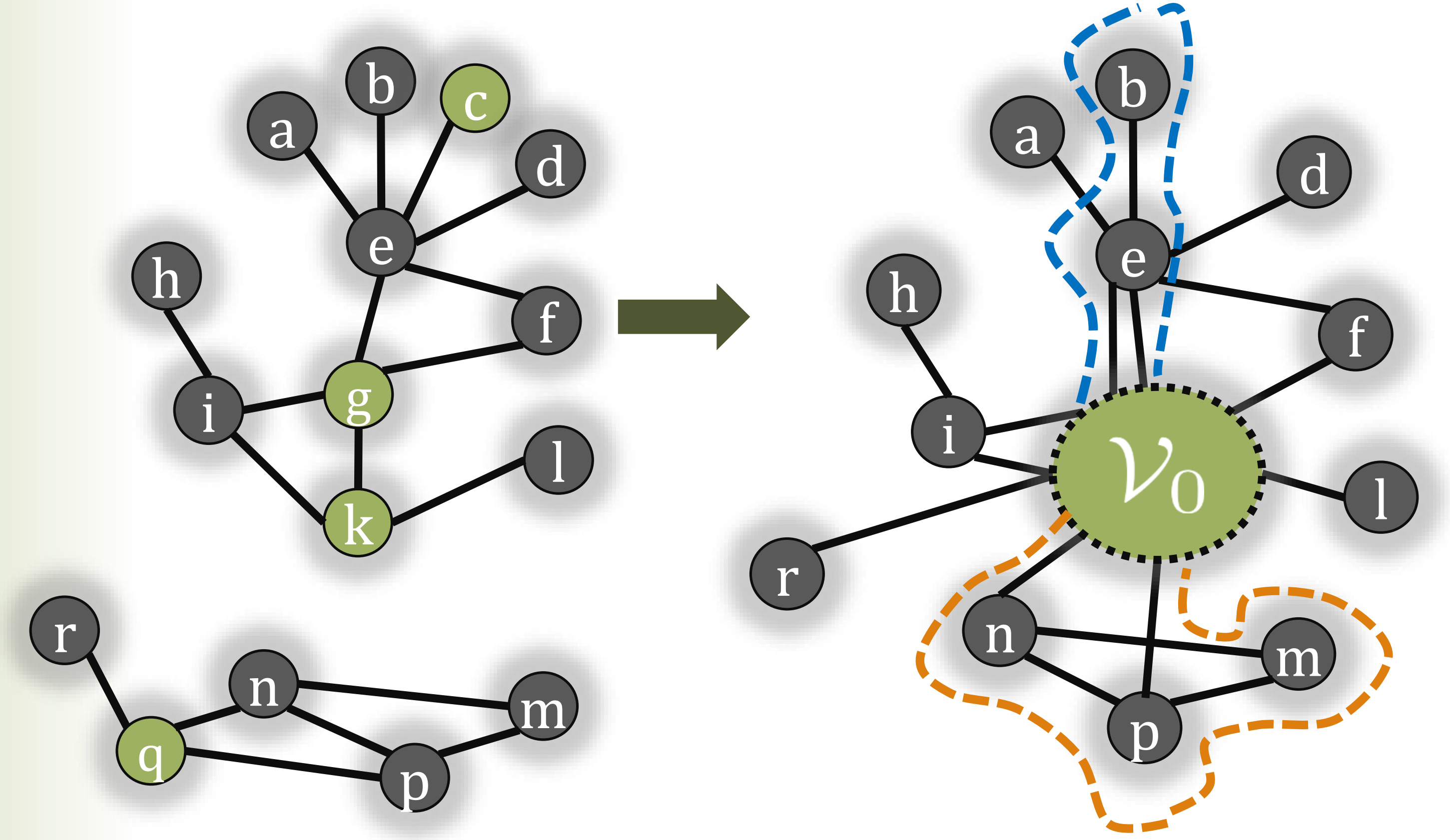


Sample

$z \sim$ Uniform$(\mathcal{V}_0)$

Idea of tours

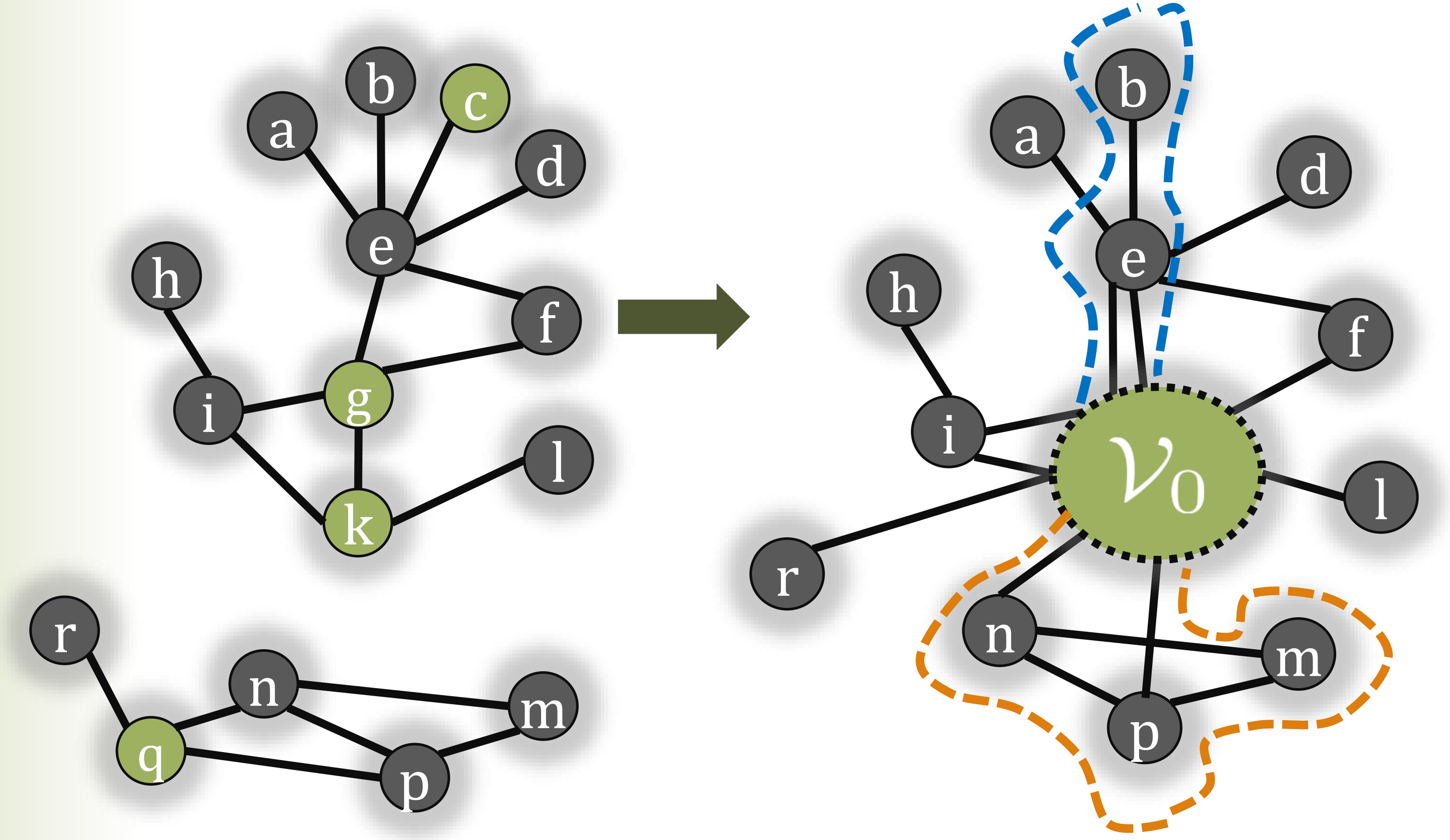
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



Sample
$z \sim \text{Uniform}(\mathcal{V}_0)$

Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



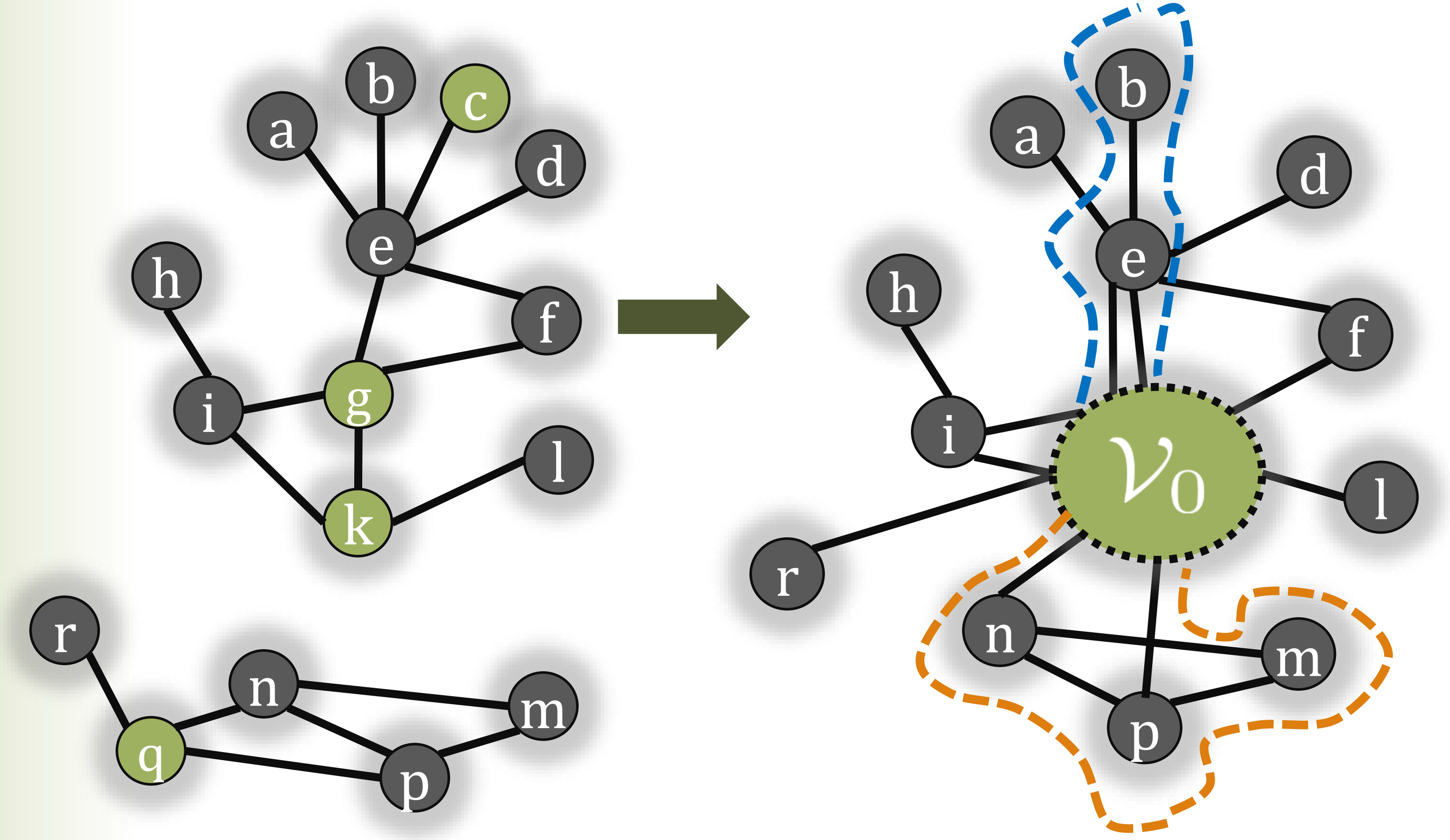
Properties of tours:

Sample

$z \sim$ Uniform$(\mathcal{V}_0)$

Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



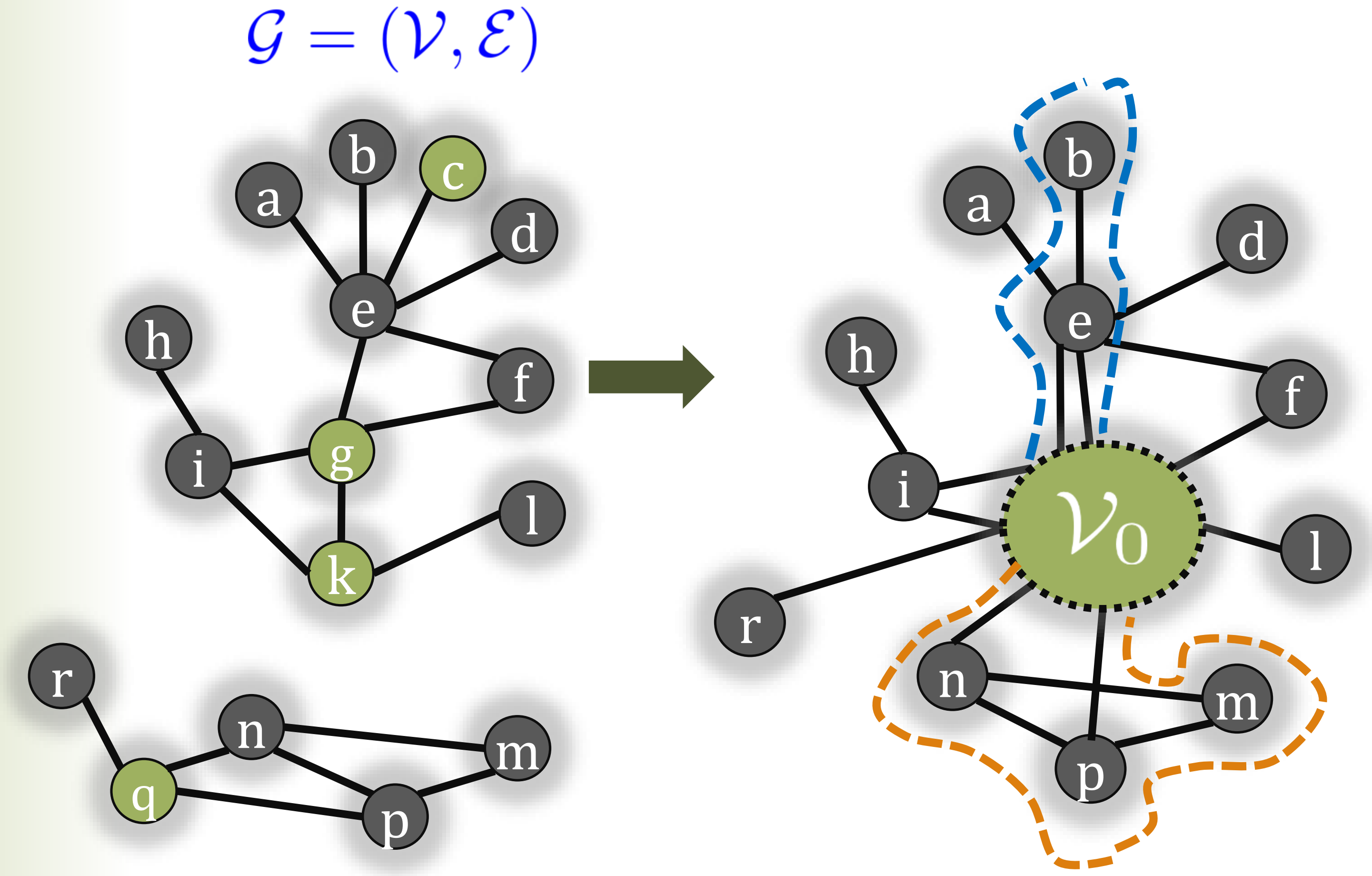
Properties of tours:

- Tours are independent

Sample

$z \sim \text{Uniform}(\mathcal{V}_0)$

Idea of tours



Properties of tours:

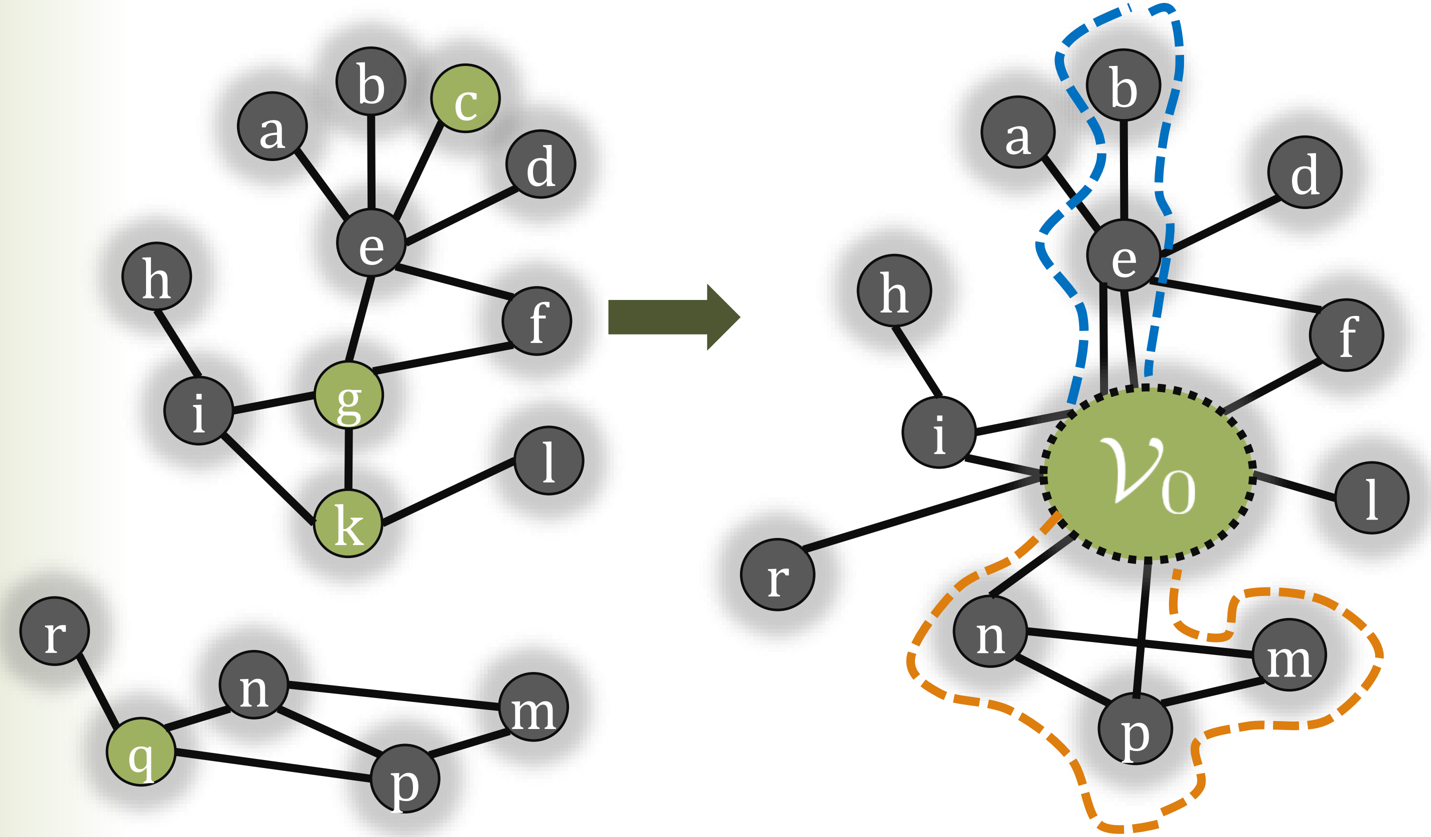
- Tours are independent
- Fully distributed crawler

Sample

$z \sim \text{Uniform}(\mathcal{V}_0)$

Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



Properties of tours:

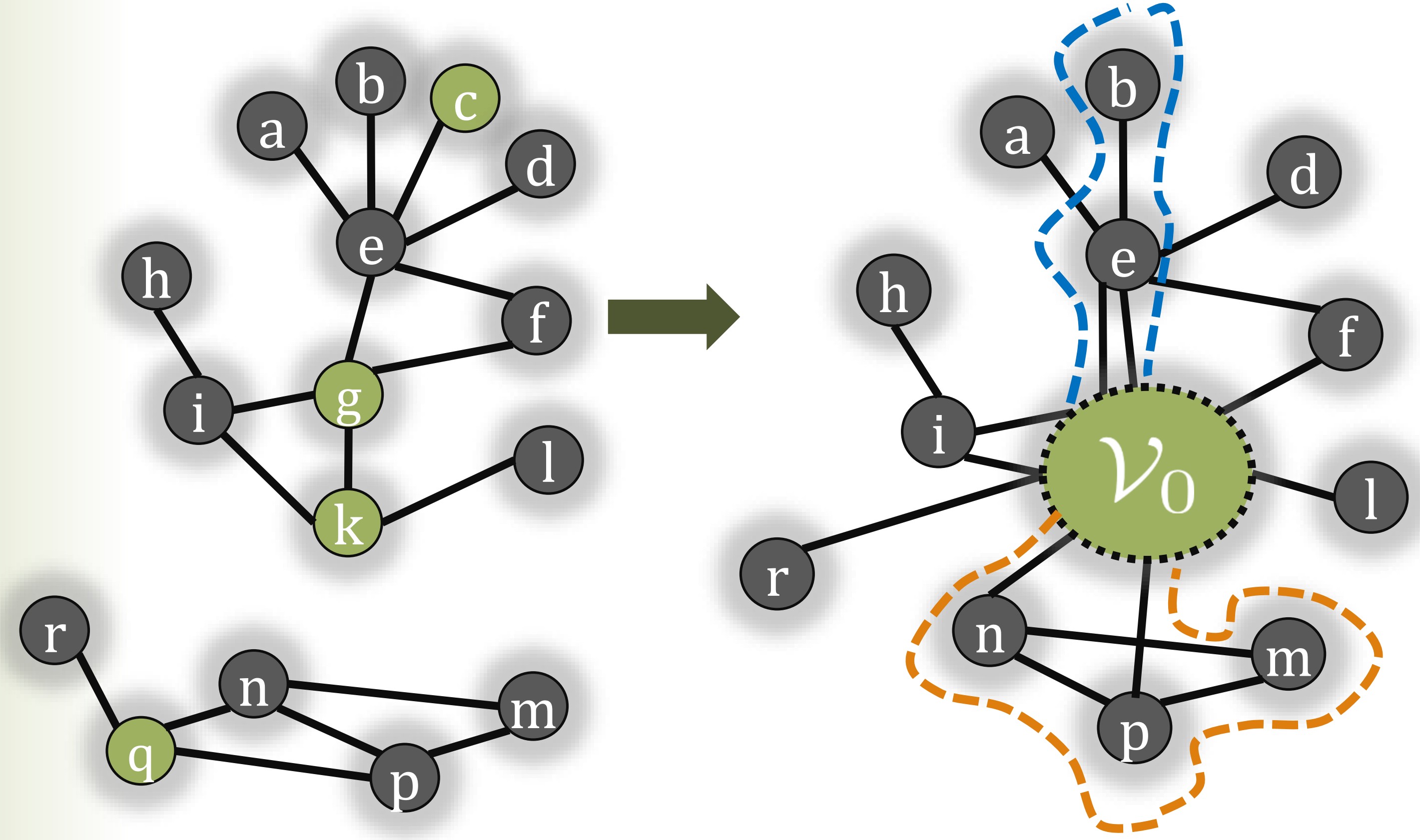
- Tours are independent
- Fully distributed crawler implementation

Sample

$z \sim \text{Uniform}(\mathcal{V}_0)$

Idea of tours

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$



Properties of tours:

- Tours are independent
- Fully distributed crawler implementation
- Larger super node size, shorter the tours

Sample

$z \sim \text{Uniform}(\mathcal{V}_0)$

Details in the paper

Reinforcement Learning technique (contd.)

Stochastic Approximation Algorithm

Details in the paper

Reinforcement Learning technique (contd.)

For each node i in $\mathcal{V}_0$

Seed set

Stochastic Approximation Algorithm

Details in the paper

Reinforcement Learning technique (contd.)

Stochastic Approximation Algorithm

For each node i in $\mathcal{V}_0$ ← Seed set

$$V_{n+1}(i) = V_n(i) + a(n)\mathbb{I}\{z = i\} \left[\underbrace{\left(\sum_{m=1}^{\xi(n)} f(X_m^n) \right)}_{\text{Function sum inside a tour}} - \frac{\xi(n)}{|\mathcal{V}_0|} \sum_{j \in \mathcal{V}_0} V_n(j) + V_n(X_{\xi(n)}^n) - V_n(i) \right]$$

Cost function

Details in the paper

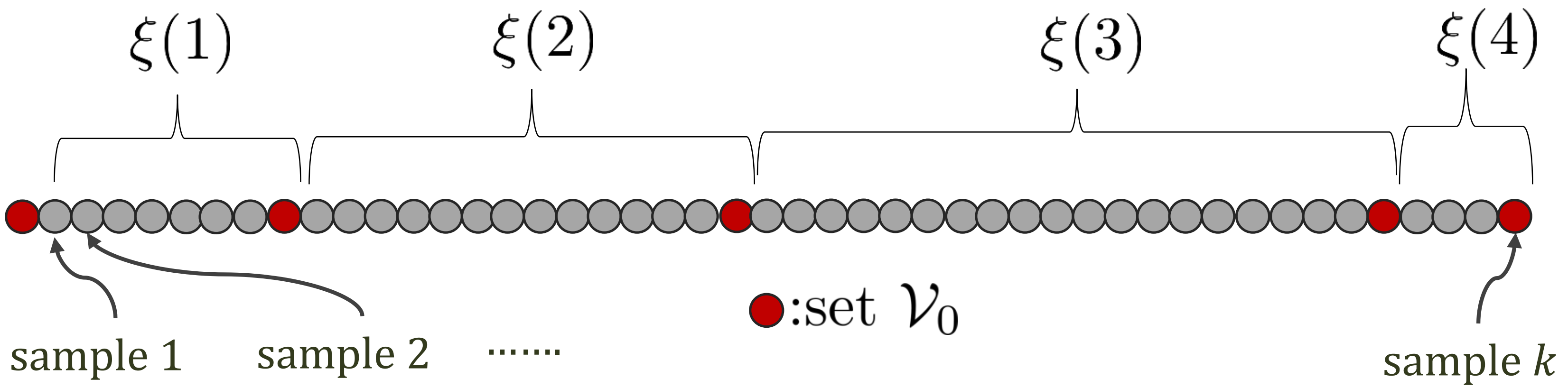
Reinforcement Learning technique (contd.)

Stochastic Approximation Algorithm

For each node i in $\mathcal{V}_0$ Seed set

$$V_{n+1}(i) = V_n(i) + a(n)\mathbb{I}\{z = i\}\left[\overbrace{\left(\sum_{m=1}^{\xi(n)} f(X_m^n) \right)}^{\text{Function sum inside a tour}} - \frac{\xi(n)}{|\mathcal{V}_0|} \sum_{j \in \mathcal{V}_0} V_n(j) + V_n(X_{\xi(n)}^n) - V_n(i) \right]$$

Cost function



Reinforcement Learning technique (contd.)

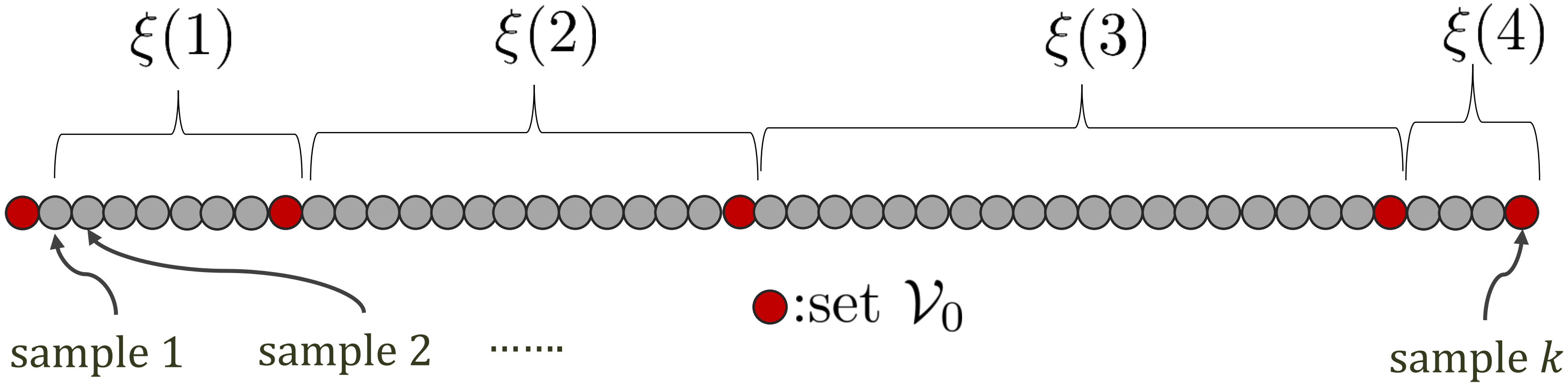
Stochastic Approximation Algorithm

For each node i in $\mathcal{V}_0$ Seed set

$$V_{n+1}(i) = V_n(i) + a(n)\mathbb{I}\{z = i\}\left[\overbrace{\left(\sum_{m=1}^{\xi(n)} f(X_m^n)\right)}^{\text{Function sum inside a tour}} - \frac{\xi(n)}{|\mathcal{V}_0|}\sum_{j\in\mathcal{V}_0} V_n(j) + V_n(X^n_{\xi(n)}) - V_n(i)\right]$$

Cost function

$a(n) > 0$ are stepsizes satisfying $\sum_n a(n) = \infty,\ \sum_n a(n)^2 < \infty.$



Reinforcement Learning technique (contd.)

Stochastic Approximation Algorithm

For each node i in $\mathcal{V}_0$ ← Seed set

$$V_{n+1}(i) = V_n(i) + a(n)\mathbb{I}\{z = i\}\left[\left(\overbrace{\sum_{m=1}^{\xi(n)} f(X_m^n)}^{\text{Function sum inside a tour}}\right) - \frac{\xi(n)}{|\mathcal{V}_0|}\sum_{j \in \mathcal{V}_0} V_n(j) + V_n(X_{\xi(n)}^n) - V_n(i)\right]$$

Cost function ↑ ($V_{n+1}(i)$)

$a(n) > 0$ are stepsizes satisfying $\sum_n a(n) = \infty,\ \sum_n a(n)^2 < \infty.$

$$\frac{1}{|\mathcal{V}_0|}\sum_{j \in \mathcal{V}_0} V_n(j) \to \sum_{u \in \mathcal{V}} \pi_u f(u)$$

Which Random Walk method to select ?

Which Random Walk method to select ?

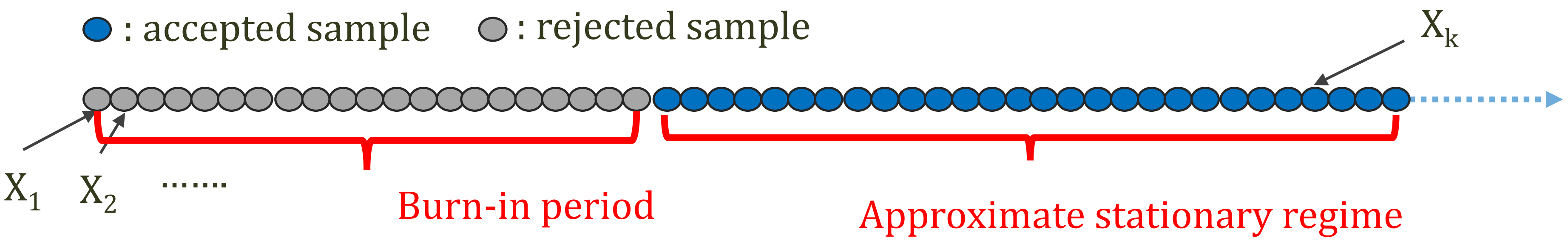
- Mixing time

Not a good criterion here due to burn-in period.

Which Random Walk method to select ?

- Mixing time

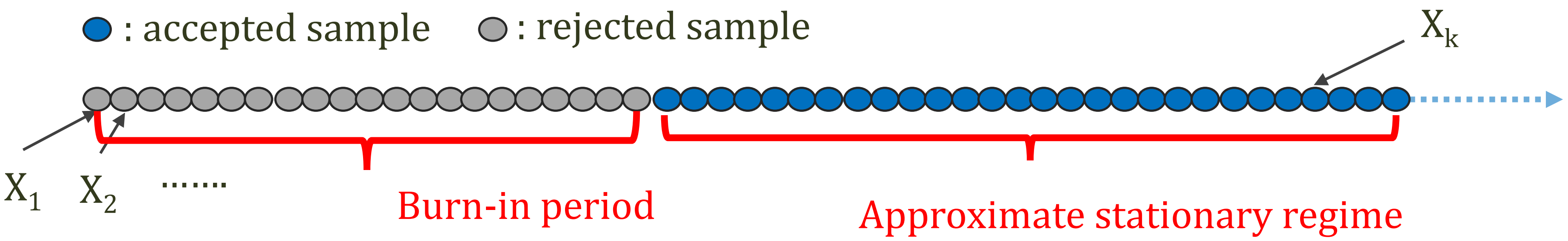
Not a good criterion here due to burn-in period.



Which Random Walk method to select ?

- Mixing time

Not a good criterion here due to burn-in period.



Reinforcement Learning technique does not require burn-in period

Which Random Walk method to select ?

- Mixing time

Not a good criterion here due to burn-in period.

- Efficiency of the estimator:

How many samples are needed to achieve certain accuracy

Asymptotic Variance

Asymptotic Variance

Asymptotic variance of the estimator

$$\sigma^2 \triangleq \lim_{n\to\infty} n \,\mathrm{Var}\left(\mu^{(n)}(\mathcal{G})\right)$$

Asymptotic Variance

Asymptotic variance of the estimator

$$\sigma^2 \triangleq \lim_{n\to\infty} n \, \mathrm{Var}\left(\mu^{(n)}(\mathcal{G})\right)$$

Also from Central Limit Theorem equivalent

$$\sqrt{n}\left(\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

Details in the paper

Asymptotic Variance (contd.)

Details in the paper

Asymptotic Variance (contd.)

- For Metropolis-Hastings Sampling,

$$\sqrt{n}\left(\hat{\mu}^{(n)}_{\mathrm{MH}}(\mathcal{G}) - \mu(\mathcal{G})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\mathrm{MH}})$$

$$\text{where} \quad \sigma^2_{\mathrm{MH}} = \frac{2}{n}\mathbf{f}^T\mathbf{Z}\mathbf{f} - \frac{1}{n}\mathbf{f}^T\mathbf{f} - \left(\frac{1}{n}\mathbf{f}^T\mathbf{1}\right)^2$$

Fundamental matrix of Markov chain

Details in the paper

Asymptotic Variance (contd.)

- For Metropolis-Hastings Sampling,

$$\sqrt{n}\left(\hat{\mu}^{(n)}_{\mathrm{MH}}(\mathcal{G}) - \mu(\mathcal{G})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\mathrm{MH}})$$

where $\sigma^2_{\mathrm{MH}} = \frac{2}{n}\mathbf{f}^T\mathbf{Z}\mathbf{f} - \frac{1}{n}\mathbf{f}^T\mathbf{f} - \left(\frac{1}{n}\mathbf{f}^T\mathbf{1}\right)^2$

Fundamental matrix of Markov chain

- For Respondent Driven Sampling,

$$\sqrt{n}(\hat{\mu}^{(n)}_{\mathrm{RDS}}(\mathcal{G}) - \mu(\mathcal{G})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\mathrm{RDS}}) \quad , \sigma^2_{\mathrm{RDS}} = \text{function}(\deg, \mathbf{Z}, \ \mathbf{f})$$

Details in the paper

Asymptotic Variance (contd.)

- For Metropolis-Hastings Sampling,

$$\sqrt{n}\left(\hat{\mu}^{(n)}_{\mathrm{MH}}(\mathcal{G}) - \mu(\mathcal{G})\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\mathrm{MH}})$$

where $\sigma^2_{\mathrm{MH}} = \frac{2}{n}\mathbf{f}^T\mathbf{Z}\mathbf{f} - \frac{1}{n}\mathbf{f}^T\mathbf{f} - \left(\frac{1}{n}\mathbf{f}^T\mathbf{1}\right)^2$

Fundamental matrix of Markov chain

- For Respondent Driven Sampling,

$$\sqrt{n}(\hat{\mu}^{(n)}_{\mathrm{RDS}}(\mathcal{G}) - \mu(\mathcal{G})) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2_{\mathrm{RDS}}) \quad , \sigma^2_{\mathrm{RDS}} = \text{function}(\deg, \mathbf{Z}, \ \mathbf{f})$$

- For Reinforcement Learning based sampling,

$$\mathbb{E}\left[|\hat{\mu}^{(n)}_{\mathrm{RL}}(\mathcal{G}) - \mu(\mathcal{G})|^2\right] = \mathcal{O}\left(\frac{1}{n}\right)$$

Numerical Studies

Numerical Studies

Normalized Root Mean Square Error (NRMSE) vs Budget B

Numerical Studies

Normalized Root Mean Square Error (NRMSE) vs Budget B

$$\mathrm{NRMSE} := \frac{\sqrt{\mathrm{MSE}}}{\mu(\mathcal{G})} \quad \text{where } \mathrm{MSE} = \mathbb{E}[(\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G}))^2]$$

Numerical Studies

Normalized Root Mean Square Error (NRMSE) vs Budget B

$$\mathrm{NRMSE} := \frac{\sqrt{\mathrm{MSE}}}{\mu(\mathcal{G})} \qquad \text{where } \mathrm{MSE} = \mathbb{E}[(\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G}))^2]$$

Why MSE ?

Normalized Root Mean Square Error (NRMSE) vs Budget B

$$\text{NRMSE} := \frac{\sqrt{\text{MSE}}}{\mu(\mathcal{G})} \quad \text{where MSE} = \mathbb{E}[(\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G}))^2]$$

Why MSE ? $\mathbb{P}\left[|\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G})| \geq \varepsilon\right] \leq \frac{\text{MSE}}{\varepsilon^2}$

Numerical Studies

Normalized Root Mean Square Error (NRMSE) vs Budget B

$$\text{NRMSE} := \frac{\sqrt{\text{MSE}}}{\mu(\mathcal{G})} \quad \text{where MSE} = \mathbb{E}[(\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G}))^2]$$

Why MSE ? $\mathbb{P}\left[|\hat{\mu}^{(n)}(\mathcal{G}) - \mu(\mathcal{G})| \geq \varepsilon\right] \leq \frac{\text{MSE}}{\varepsilon^2}$

Budget B: number of allowed samples

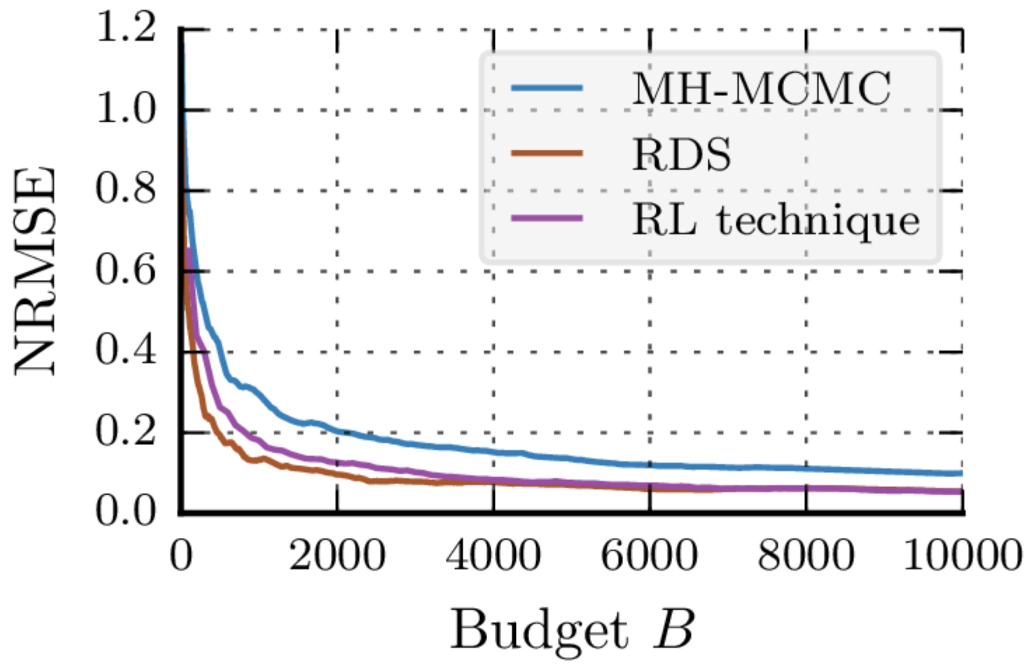
Les Misérables network

Les Misérables network

Number of nodes: 77, number of edges: 254.

Les Misérables network

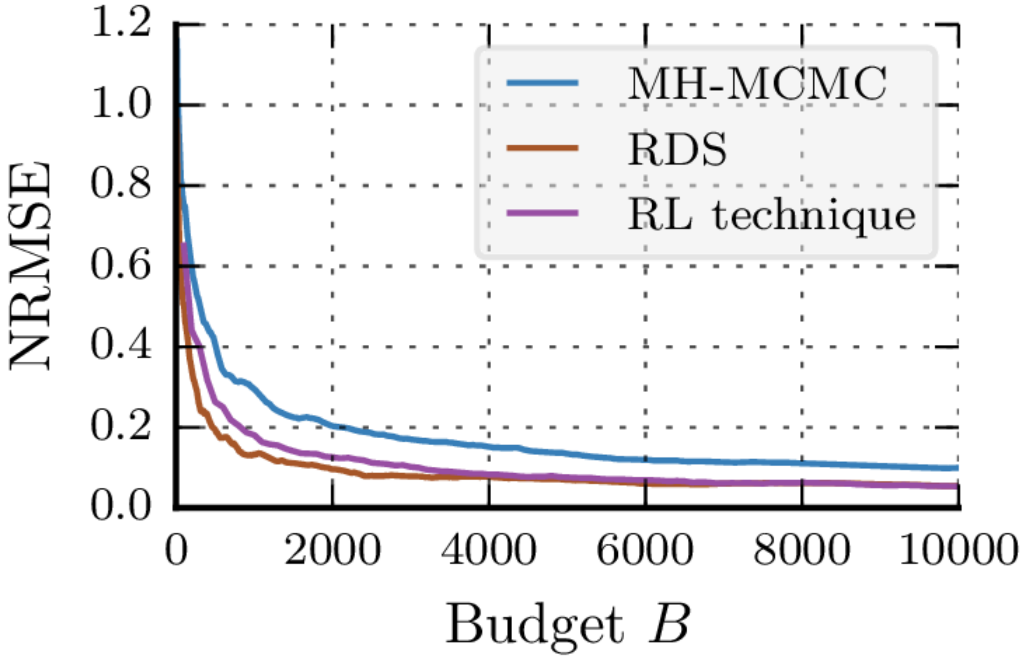
Number of nodes: 77, number of edges: 254.



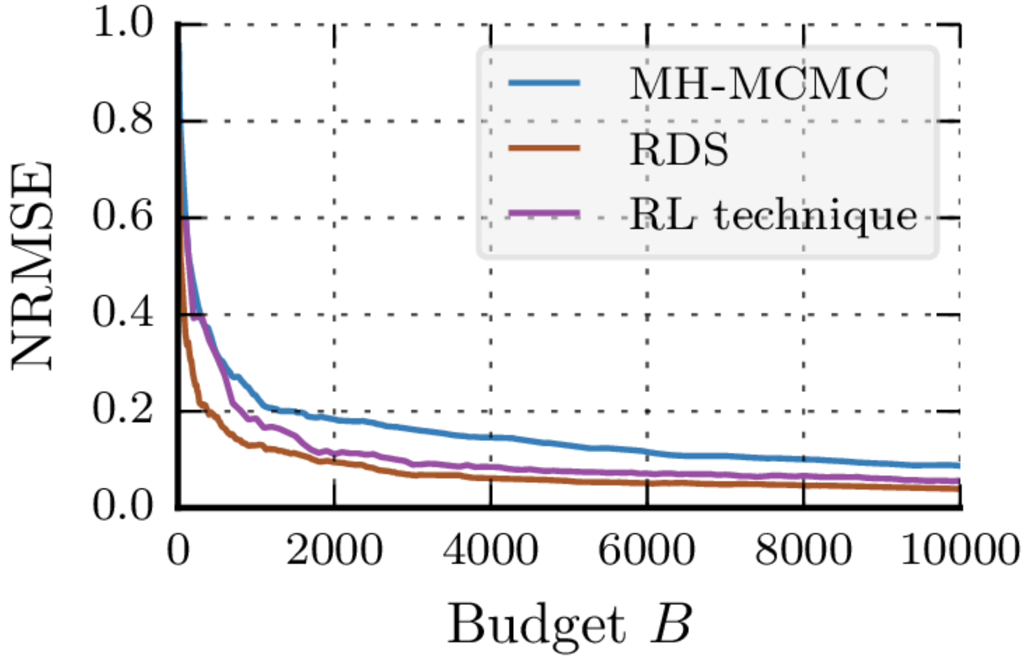
$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbb{I}\{\deg(\mathrm{u}) > 10\}$$

Les Misérables network

Number of nodes: 77, number of edges: 254.



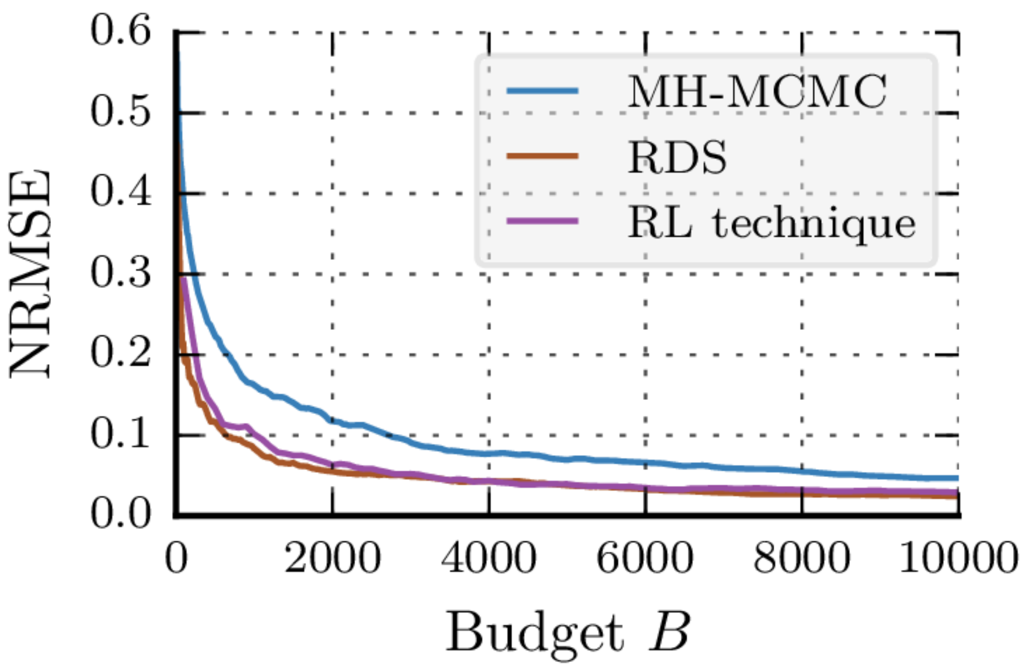
$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbb{I}\{\mathrm{deg(u)} > 10\}$$



$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbb{I}\{\mathrm{deg(u)} < 4\}$$

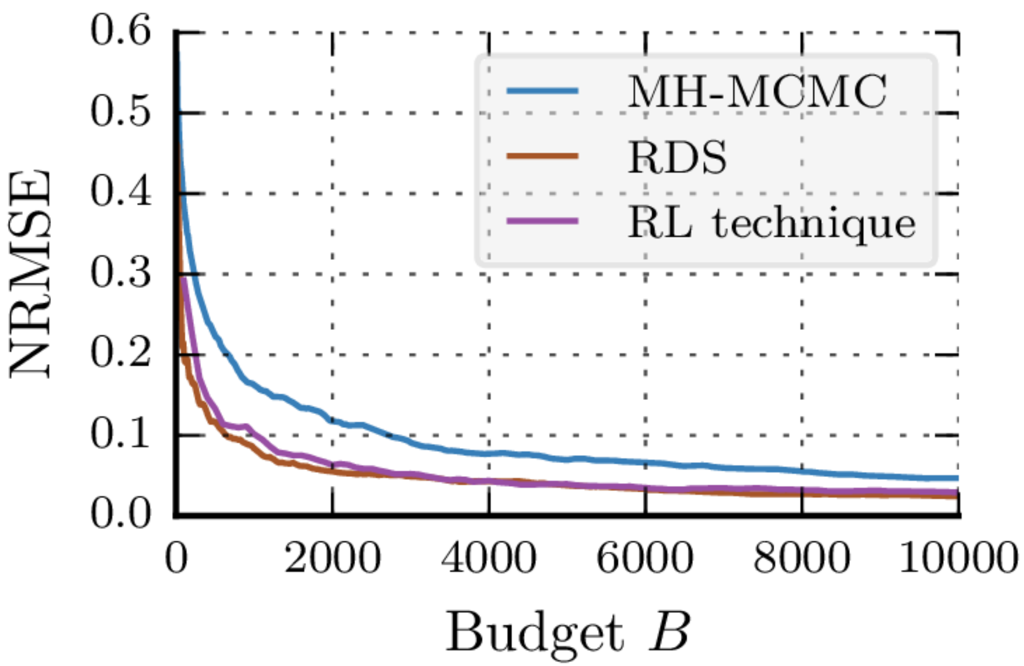
Les Misérables network contd.

Les Misérables network contd.

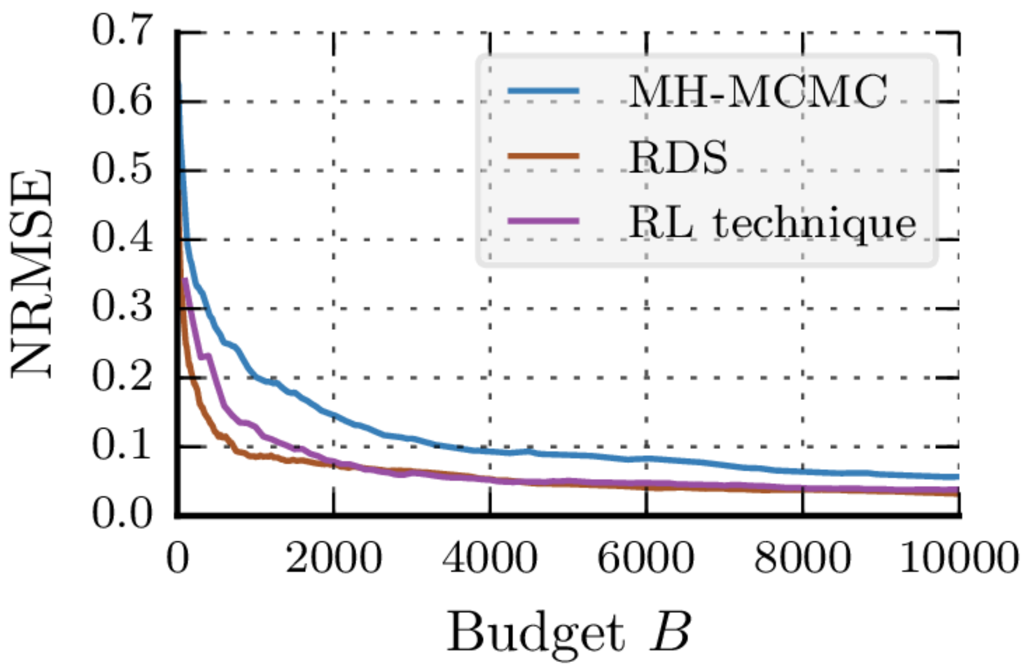


$$\mu(\mathcal{G}) = \text{Average degree}$$

Les Misérables network contd.



$\mu(\mathcal{G})$ = Average degree



$\mu(\mathcal{G})$ = Average clustering coefficient

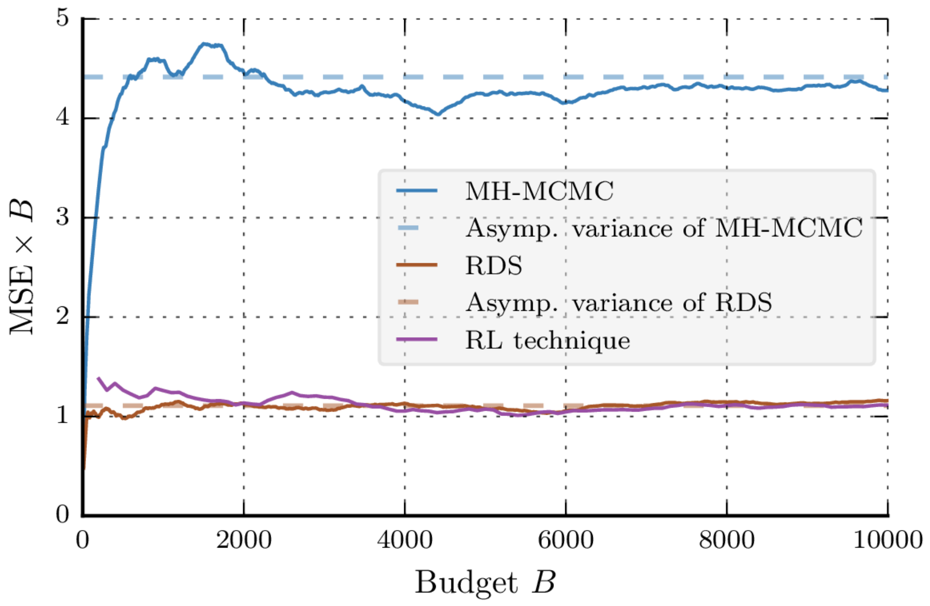
Les Misérables network contd.

Les Misérables network contd.

Study of asymptotic variance

Les Misérables network contd.

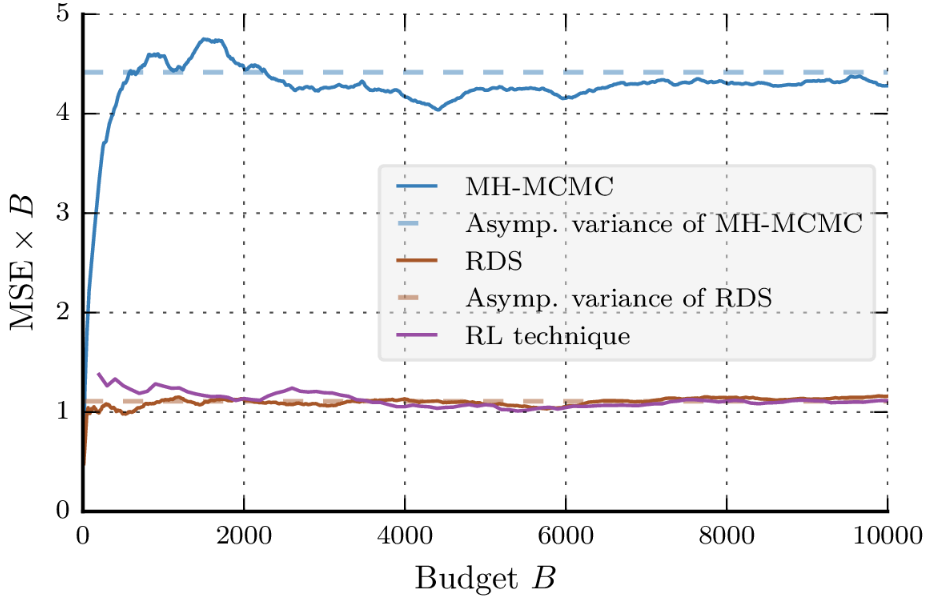
Study of asymptotic variance

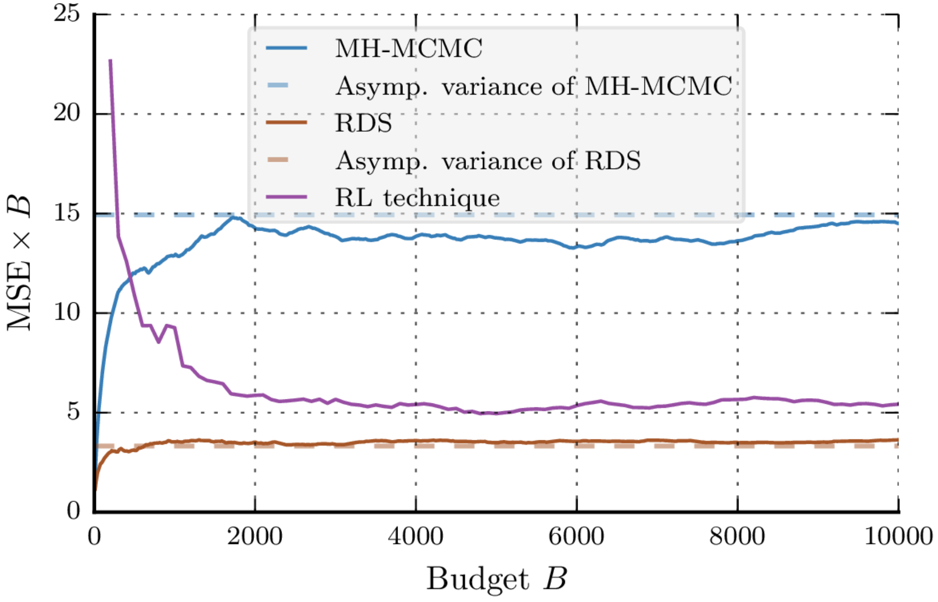


$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbb{I}\{\mathrm{deg(u)} > 10\}$$

Les Misérables network contd.

Study of asymptotic variance



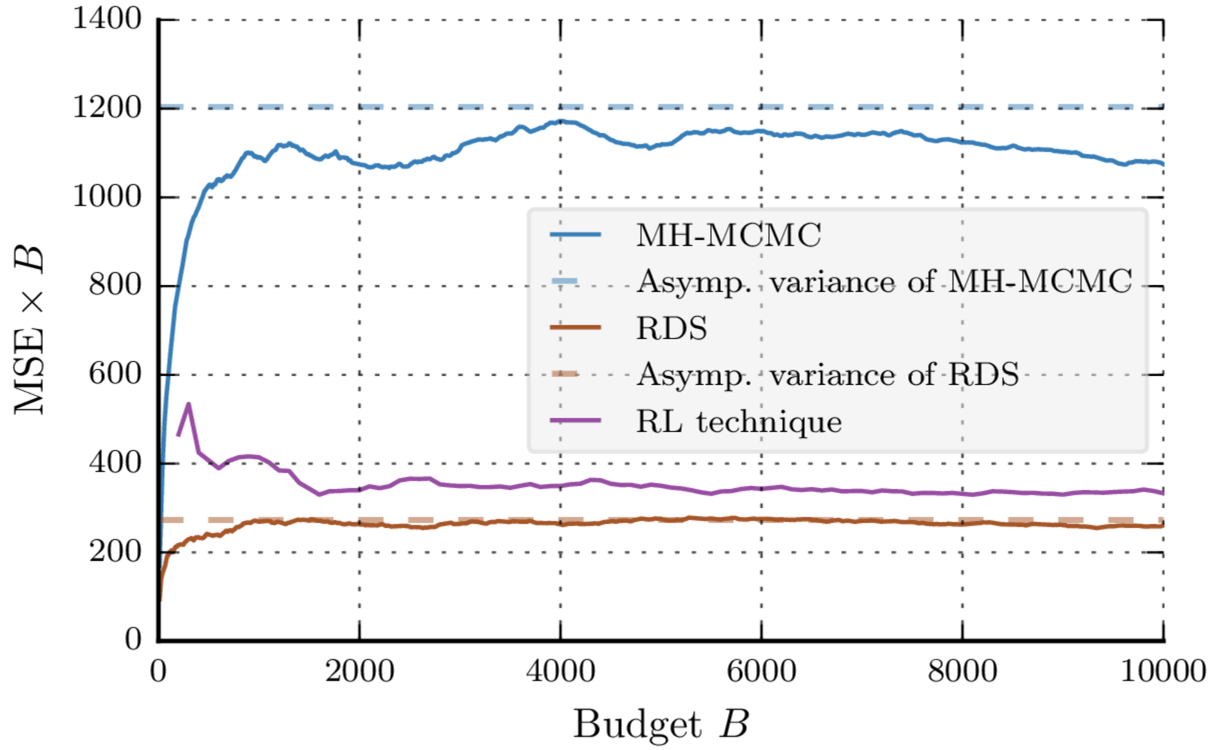


$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} \mathbb{I}\{\deg(\mathrm{u}) > 10\}$$

$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} \mathbb{I}\{\deg(\mathrm{u}) < 4\}$$

Les Misérables network contd.

Study of asymptotic variance



$$\mu(\mathcal{G}) = \text{Average degree}$$

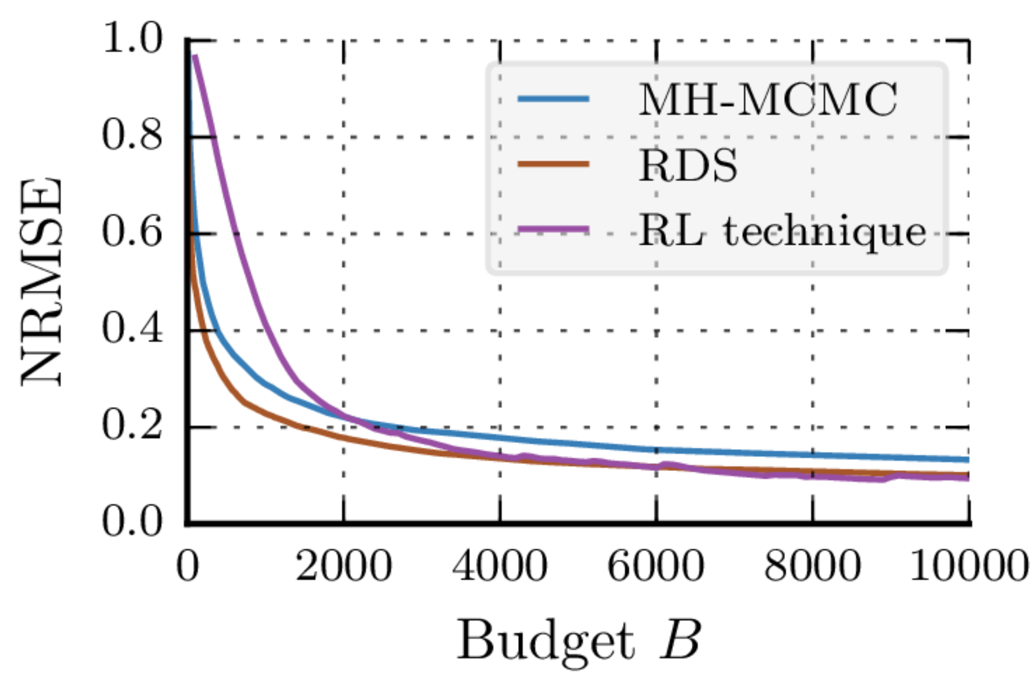
Friendster network

Friendster network

Number of nodes ~ 65K   number of edges ~ 1.25M

Friendster network

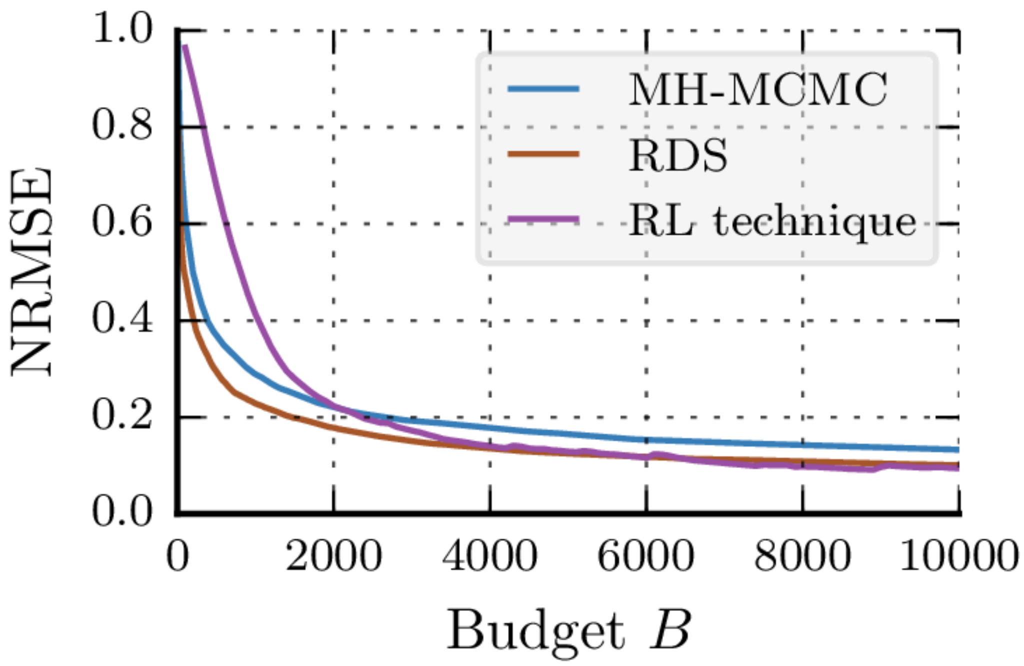
Number of nodes ~ 65K   number of edges ~ 1.25M

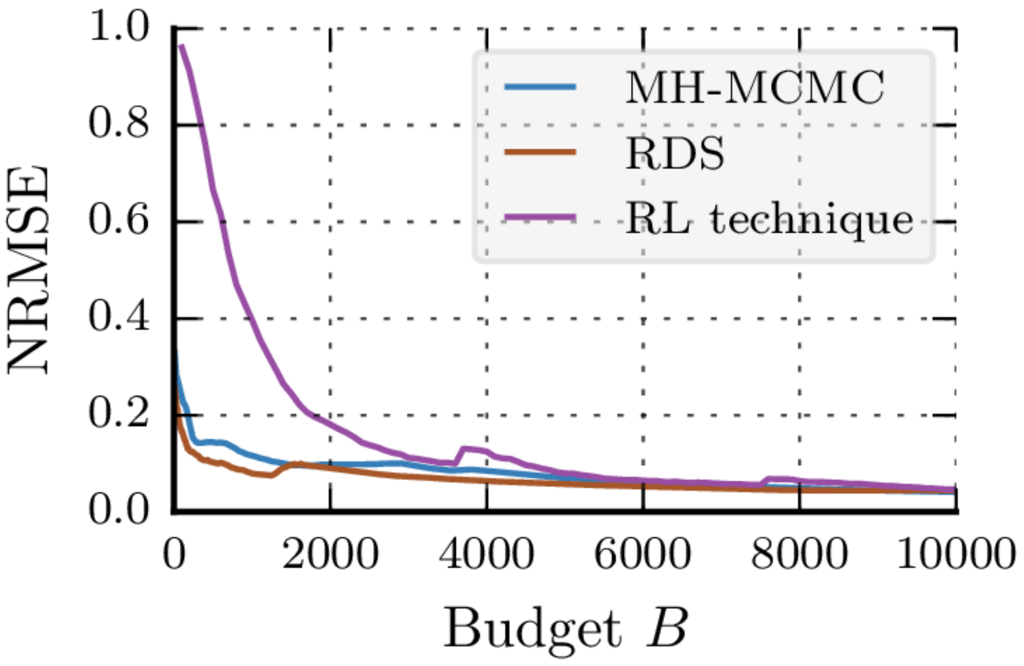


$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbb{I}\{\mathrm{deg(u)} > 50\}$$

Friendster network

Number of nodes ~ 65K   number of edges ~ 1.25M





$$\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} \mathbb{I}\{\deg(\mathrm{u}) > 50\}$$

$$\mu(\mathcal{G}) = \text{Average clustering coefficient}$$

Friendster network contd.

Friendster network contd.

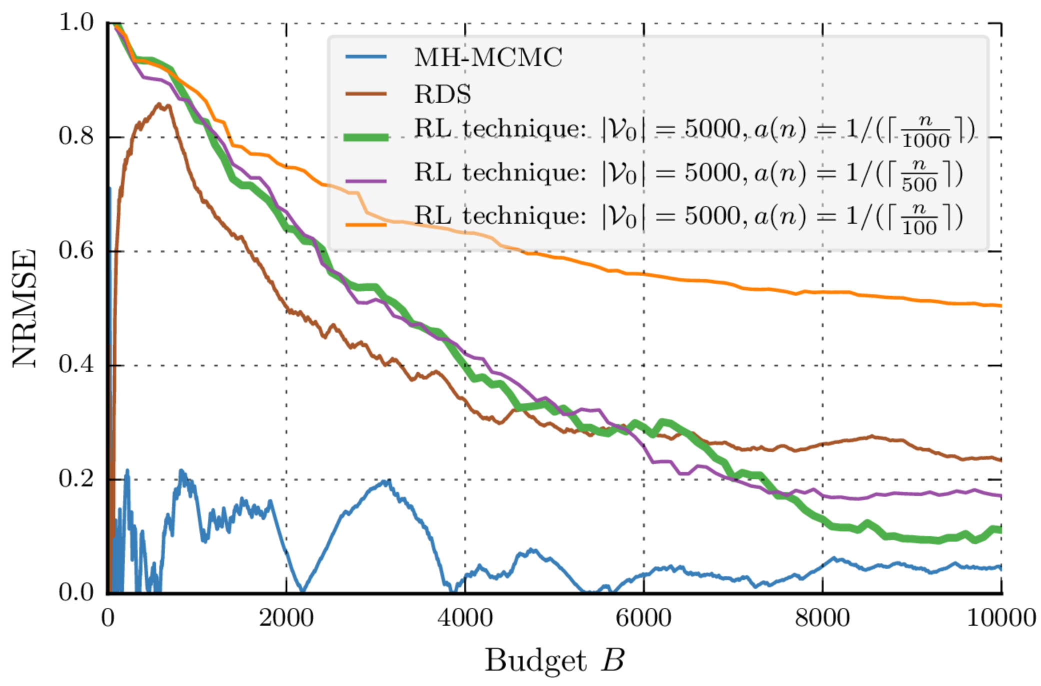
Stability of sample paths:

Friendster network contd.

Stability of sample paths: single path example

Friendster network contd.

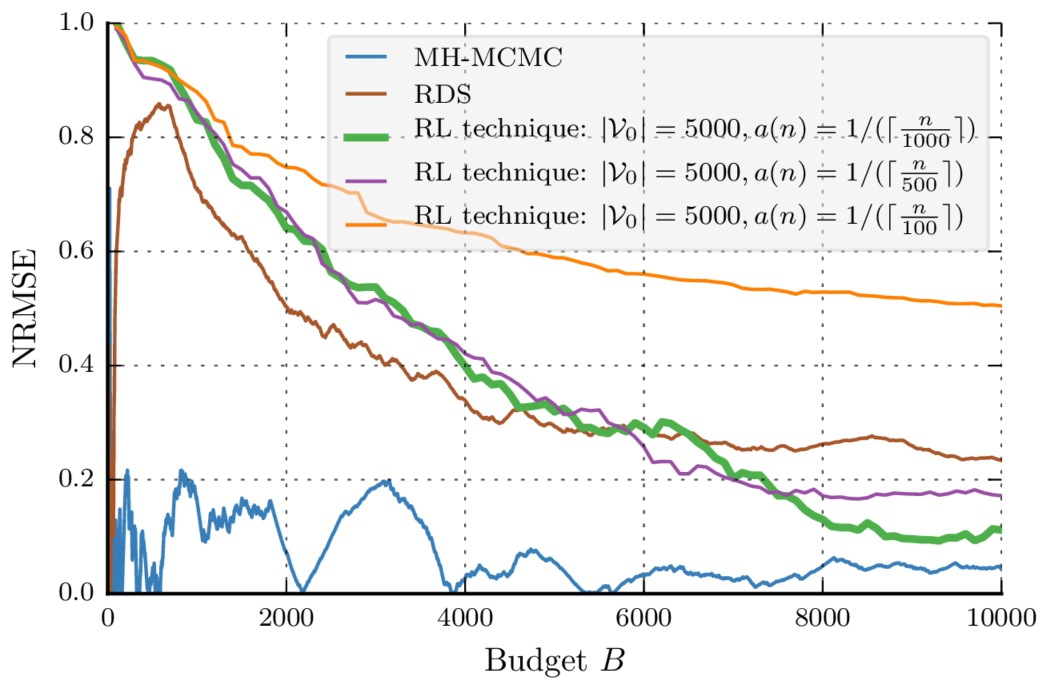
Stability of sample paths: single path example

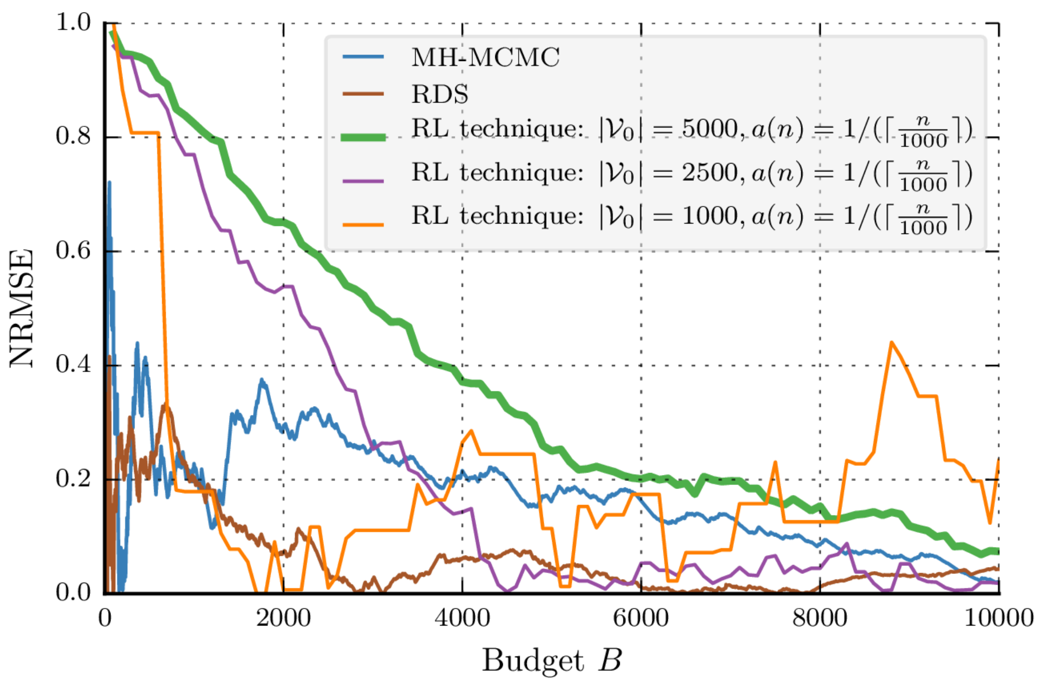


Varying super-node size

Friendster network contd.

Stability of sample paths: single path example





Varying super-node size

Varying step size

Conclusions

Conclusions

- Rand Walk based estimators of $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$

Conclusions

- Rand Walk based estimators of $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}} f(u)$
- Numerical and theoretical study of Mean Square Error & Asymptotic Variance of
  - ✓ Metropolis-Hastings sampling
  - ✓ Respondent Driven sampling (RDS)
  - ✓ New Reinforcement Learning based sampling (RL)

Conclusions

- Rand Walk based estimators of $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$
- Numerical and theoretical study of Mean Square Error & Asymptotic Variance of
  - ✓ Metropolis-Hastings sampling
  - ✓ Respondent Driven sampling (RDS)
  - ✓ New Reinforcement Learning based sampling (RL)
- Reinforcement Learning technique:
  - ✓ Tackles disconnected graph
  - ✓ A cross between deterministic iteration and MCMC
  - ✓ Can control the stability of the algorithm with step sizes

Conclusions

- Rand Walk based estimators of $\mu(\mathcal{G}) = \frac{1}{|\mathcal{V}|} \sum_{u \in \mathcal{V}} f(u)$
- Numerical and theoretical study of Mean Square Error & Asymptotic Variance of
  - ✓ Metropolis-Hastings sampling
  - ✓ Respondent Driven sampling (RDS)
  - ✓ New Reinforcement Learning based sampling (RL)
- Reinforcement Learning technique:
  - ✓ Tackles disconnected graph
  - ✓ A cross between deterministic iteration and MCMC
  - ✓ Can control the stability of the algorithm with step sizes
- RDS works better. RL technique comparable, yet more stable and no burn-in !

Thank you!

http://bit.do/Jithin