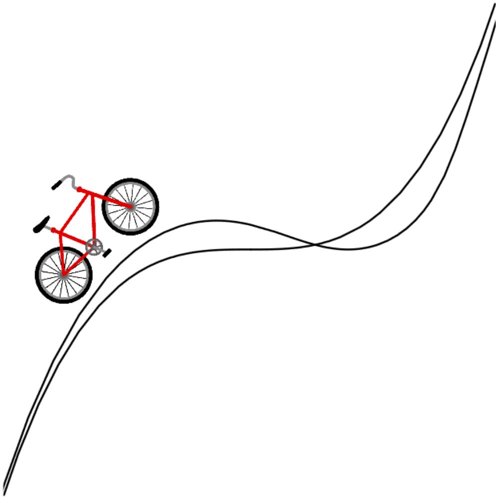
Finding the source

Jacob Richey

joint with: Miki Racz, Chris Hoffman, Gourab Ray

UBC, October 2022



Which way did the bicycle go?

Q: Given a snapshot of a (random) process, what can be determined?

Q: Given a snapshot of a (random) process, what can be determined?

- Starting/ending point?
- Most/least visited points?
- Step distribution/generator?
- Properties of the underlying graph?

Q: Given a snapshot of a (random) process, what can be determined?

- Starting/ending point?
- Most/least visited points?
- Step distribution/generator?
- Properties of the underlying graph?

Warmup: simple random walk on $\mathbb{Z}$.

Problem: Run until the range has size 11, then guess the starting point.

Warmup: simple random walk on $\mathbb{Z}$.

Problem: Run until the range has size 11, then guess the starting point.

Warmup: simple random walk on $\mathbb{Z}$.

Problem: Run until the range has size 11, then guess the starting point.

Given that the range is an interval of length N, what's the most likely starting point? Purple, red, or green?

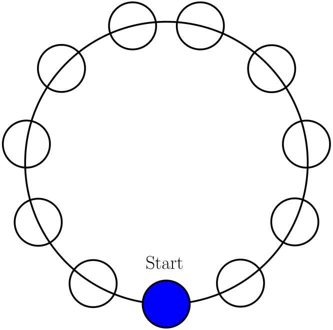
A: They're all equally likely!

Proof sketch: think of the range as a 'coin switching' markov chain, compute transition probabilities recursively.

A: They're all equally likely!

Proof sketch: think of the range as a 'coin switching' markov chain, compute transition probabilities recursively.

Alternatively: last vertex visited by SRW on the ring is uniform.



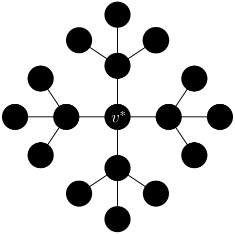
Consider a rumor spreading through a network.

- The rumor starts from a ‘source’ vertex
- At integer times, nodes can pass the rumor to neighbors
- Observer sees which nodes have the rumor at some time

Consider a rumor spreading through a network.

- The rumor starts from a ‘source’ vertex
- At integer times, nodes can pass the rumor to neighbors
- Observer sees which nodes have the rumor at some time

For today, focus on the d-regular tree.



The rumor is spread by a random algorithm known to the observer.

Goals for the rumor spreader:

- *Spreading*: spread the rumor to many nodes
- *Obfuscation*: minimize the probability that the observer guesses the source correctly
- *Multiple observations*: obfuscate the source even when the observer has access to multiple independent rumors

The rumor is spread by a random algorithm known to the observer.

Goals for the rumor spreader:

- *Spreading*: spread the rumor to many nodes
- *Obfuscation*: minimize the probability that the observer guesses the source correctly
- *Multiple observations*: obfuscate the source even when the observer has access to multiple independent rumors
- *Local spreading* (new): ensure that all vertices close to the source learn the rumor quickly

Social media metadata

Obfuscating the source $\leftrightarrow$ protecting user data

Social media metadata

Obfuscating the source $\leftrightarrow$ protecting user data

Contact tracing / finding patient zero

Previous work: SI/SIR. MLE well understood. Rumor centrality

New algorithm: adaptive diffusion

- $G = d$-regular tree
- $G_t =$ set of nodes that know the rumor at time t
- $vs_t =$ virtual source at time t
- **G_t is a ball of radius $t/2$ centered at vs_t at even times t**
- Defined by transition probabilities $\alpha(t, h)$ for the virtual source

Virtual source evolves according to:

- Start with $vs_0 = v^*$

Virtual source evolves according to:

- Start with $vs_0 = v^*$
- vs_2 is a uniform neighbor of v^*.

Virtual source evolves according to:

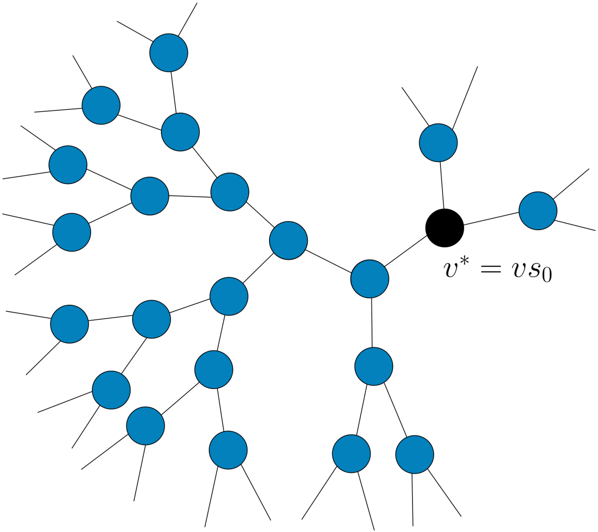
- Start with $vs_0 = v^*$
- vs_2 is a uniform neighbor of v^*.
- Let $h = \text{dist}(vs_t, v^*)$
- Probability $\alpha(t, h)$: $vs_{t+2} =$ uniform neighbor of vs_t excluding previous virtual sources
- Probability $1 - \alpha(t, h)$: $vs_{t+2} = vs_t$

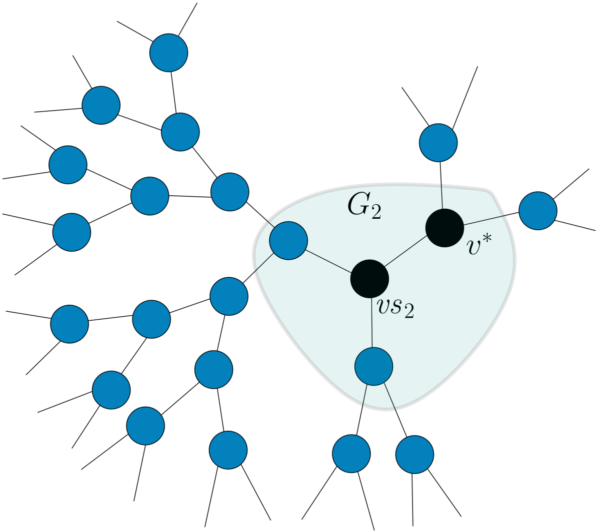
Virtual source evolves according to:

- Start with $vs_0 = v^*$
- vs_2 is a uniform neighbor of v^*.
- Let $h = \text{dist}(vs_t, v^*)$
- Probability $\alpha(t,h)$: $vs_{t+2} =$ uniform neighbor of vs_t excluding previous virtual sources
- Probability $1 - \alpha(t,h)$: $vs_{t+2} = vs_t$

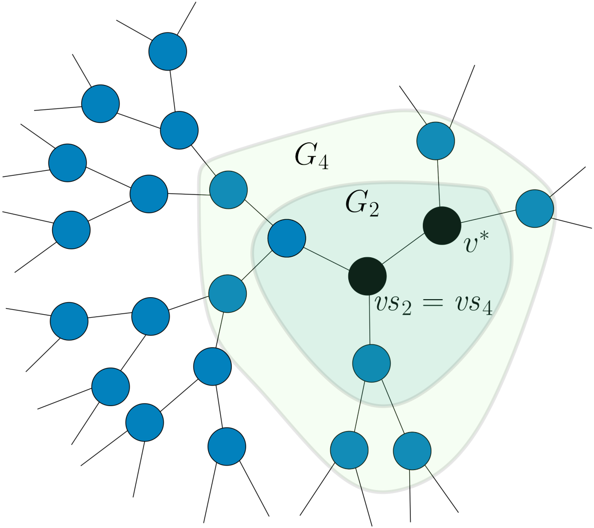
When the virtual source moves, it always moves in a uniform direction away from v^*.

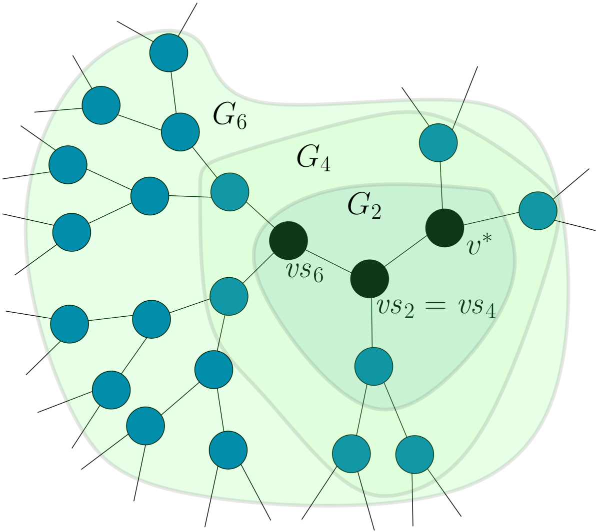
Equivalently, work with $p(t,h) = \mathbb{P}(\text{dist}(vs_t, v^*) = h)$.


$v^* = vs_0$



G_2
v^*
vs_2


G_4
G_2
v^*
$vs_2 = vs_4$



G_6
G_4
G_2
v^*
vs_6
$vs_2 = vs_4$

MLE for the source vertex: for any trajectory ω,

$$\hat{v}_{MLE}(\omega) = \arg\max_{v \in \omega} \mathbb{P}(G_t^v = \omega),$$

where G_t^v is an independent copy of G_t started from v.

Think of $\hat{v}_{MLE} = \hat{v}_{MLE}(G_t)$ as a random variable.

MLE for the source vertex: for any trajectory ω,

$$\hat{v}_{MLE}(\omega) = \arg\max_{v \in \omega} \mathbb{P}(G_t^v = \omega),$$

where G_t^v is an independent copy of G_t started from v.

Think of $\hat{v}_{MLE} = \hat{v}_{MLE}(G_t)$ as a random variable.

Fact

$\mathbb{P}(G_t^v = G_t)$ is maximized at any vertex at distance h^* from vs_t, where

$$h^* = \arg\max_{h \in \{1,2,\ldots,t/2\}} \frac{p(t,h)}{(d-1)^h}.$$

So $\hat{v}_{MLE}$ picks any vertex at distance h^* from vs_t.

Spreading

For adaptive diffusion,

$$|G_t| = N_t = \frac{1}{d-2}(d-1)^{t/2}.$$

deterministically at even times t. (Order-optimal spreading)

Obfuscation

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = \begin{cases} \Theta(N_t^{-1}) & \text{(perfect obfuscation)} \\ \Theta(N_t^{-\gamma}) & \text{(polynomial obfuscation)} \\ o(1) & \text{(weak obfuscation)} \end{cases}$$

SI/SIR: good spreading, weak obfuscation; not even weak obfuscation under multiple observations. [Shah, Zaman, Dong, Tan, Wang, Zhang]

Adaptive diffusion (Fanti, Kairouz, Oh, Viswanath '15)

Let $G = d$-regular tree. There exists an adaptive diffusion algorithm that achieves perfect obfuscation:

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = \Theta(N_t^{-1})$$

SI/SIR: good spreading, weak obfuscation; not even weak obfuscation under multiple observations. [Shah, Zaman, Dong, Tan, Wang, Zhang]

Adaptive diffusion (Fanti, Kairouz, Oh, Viswanath '15)

Let $G = d$-regular tree. There exists an adaptive diffusion algorithm that achieves perfect obfuscation:

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = \Theta(N_t^{-1})$$

Proof sketch: choose $p(t, h) \sim (d-1)^h$, so the MLE picks a uniform random vertex in G_t. Show this is realizable for some values $\alpha(t, h)$.

Q: Does it have good local spreading?

Q: Does it have good local spreading?

Definition

The *local spread* R_t is the radius of the largest ball centered at v^* and contained in G_t.

For adaptive diffusion, $R_t = \#$ of times the virtual source has *not* moved:

$$R_t = t/2 - \text{dist}(v^*, vs_t).$$

Q: Does it have good local spreading?

Definition

The *local spread* R_t is the radius of the largest ball centered at v^* and contained in G_t.

For adaptive diffusion, $R_t = \#$ of times the virtual source has *not* moved:

$$R_t = t/2 - \text{dist}(v^*, vs_t).$$

The algorithm from the theorem doesn't even achieve weak local spread!

$$p(t, h) \sim (d-1)^h \implies \text{dist}(vs_t, v^*) \approx t/2 - O(1).$$

Spreading/obfuscation trade-off [Racz, Richey '18]

Consider any adaptive diffusion with polynomial obfuscation of order $\gamma \in (0, 1)$, i.e.

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = O(N_t^{-\gamma}).$$

Then the local spreading is bounded from above:

$$\mathbb{E}[R_t] \leq (1 - \gamma)\frac{t}{2} + O(\log t).$$

Spreading/obfuscation trade-off [Racz, Richey '18]

Consider any adaptive diffusion with polynomial obfuscation of order $\gamma \in (0,1)$, i.e.

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = O(N_t^{-\gamma}).$$

Then the local spreading is bounded from above:

$$\mathbb{E}[R_t] \leq (1-\gamma)\frac{t}{2} + O(\log t).$$

Obfuscation and local spreading are **inversely linked** in this case.

The trade-off is essentially tight:

Spreading/obfuscation trade-off [Racz, Richey '18]

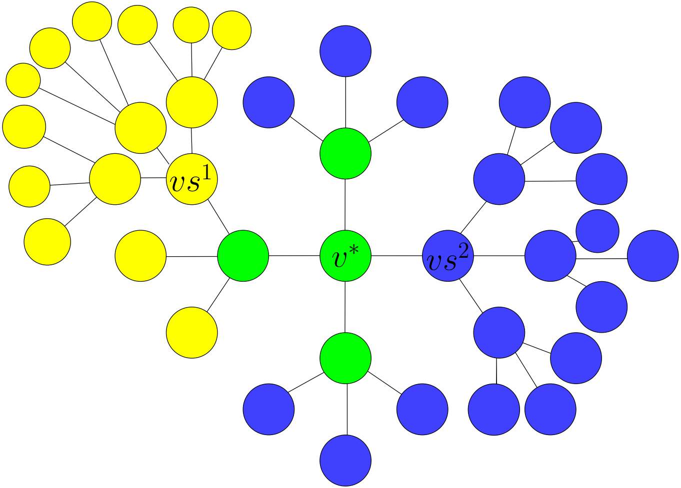
For every $\gamma \in (0,1)$, there exists an adaptive diffusion with both polynomial obfuscation of order γ,

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = O(N_t^{-\gamma}),$$

and order optimal local spreading

$$\mathbb{E}[R_t] \geq (1-\gamma)\frac{t}{2}.$$

Suppose the observer has access to $k > 1$ independent snapshots $\{G_t^i\}_{i=1}^k$ of the diffusion started from the same source v^*.



Two independent observations (Racz, Richey '18)

Suppose the observer has two iid adaptive diffusion snapshots G_t^1 and G_t^2 started from the same source v^*. There exists a nice estimator $\hat{v}$, not depending on the spreading algorithm, such that for any t,

$$\mathbb{P}(\hat{v} = v^*) \geq \frac{d-1}{d} \cdot \frac{2}{t}.$$

Moreover, there exists a protocol such that for any t,

$$\mathbb{P}(\hat{v}_{MLE} = v^*) \leq \frac{d-1}{d} \cdot \frac{7}{t}.$$

Only weak obfuscation now!

It gets worse:

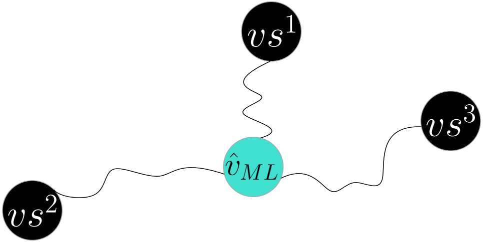
Three or more independent observations (Racz, Richey '18)

Suppose the observer has $k \geq 3$ iid snapshots G_t^i, $i \in [k]$ started from the same source v^*. There exists a nice estimator $\hat{w}$, not depending on the spreading algorithm, such that for any t,

$$\mathbb{P}(\hat{w} = v^*) \geq 1 - d \exp\left(-\frac{(d-2)^2}{2d^2} k\right).$$

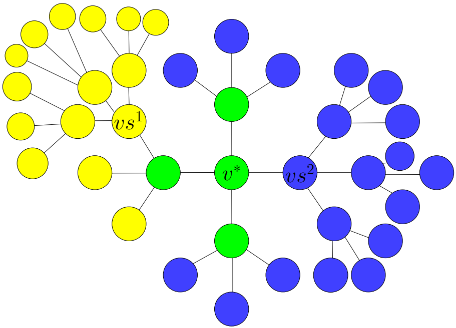
Not even weak obfuscation!

Proof: Pick any three virtual sources and draw the paths between them.

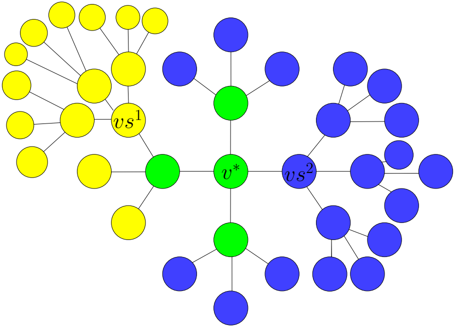


When the three virtual sources lie in different sub-trees away from the root, there will be a unique intersection point $\hat{w}$.

Simple estimator: guess a green vertex!



Simple estimator: guess a green vertex!



Obfuscation and local spreading are **positively linked** in this case:

$$\mathbb{P}(\hat{v}_{MLE} = v^*) \geq \mathbb{E}\left[\left|\bigcap_{i=1}^{k} G_t^i\right|^{-1}\right]$$

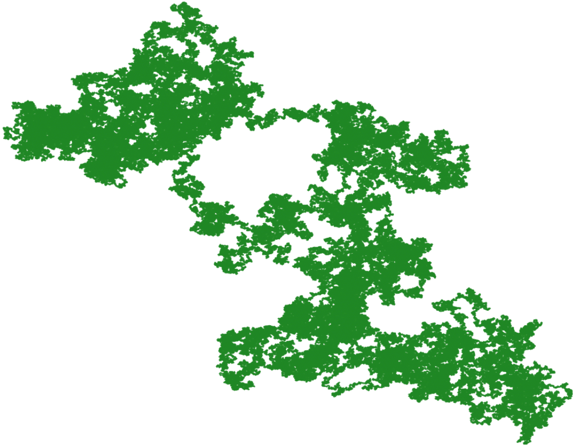
Question

Does there exist a spreading algorithm that achieves order-optimal spreading and polynomial obfuscation given ≥ 2 observations?

Should look at algorithms that have order-optimal local spreading.

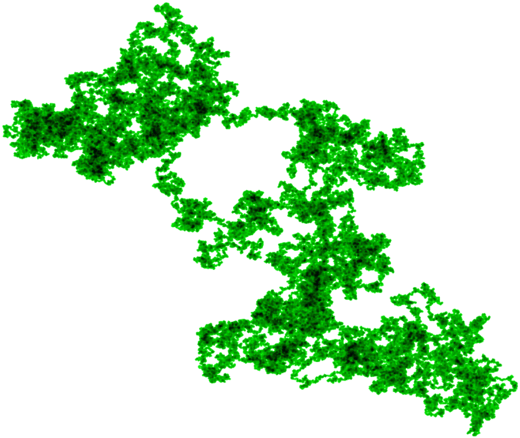
Also, need more randomness: adaptive diffusion is given by the path of a single particle (the virtual source). Too simple!

Simple random walk on $\mathbb{Z}^2$, run for $5 \cdot 10^6$ steps.



Previous work

Same SRW as before, with occupation times



Previous work

Assume *partial* information about the occupation measure.

Previous work

Assume *partial* information about the occupation measure.

(Warren, Yor '98) Brownian burgler: BM conditioned on local times

Previous work

Assume *partial* information about the occupation measure.

(Warren, Yor ‘98) Brownian burgler: BM conditioned on local times

Theorem (Pemantle, Peres, Pitman, Yor ’00)

Let $d \geq 3$, and consider Brownian motion in $\mathbb{R}^d$ run for time 1.

Given the occupation measure of the path projected onto the sphere, you can recover the range and the endpoint with probability 1.

Previous work

Assume *partial* information about the occupation measure.

(Warren, Yor '98) Brownian burgler: BM conditioned on local times

Theorem (Pemantle, Peres, Pitman, Yor '00)

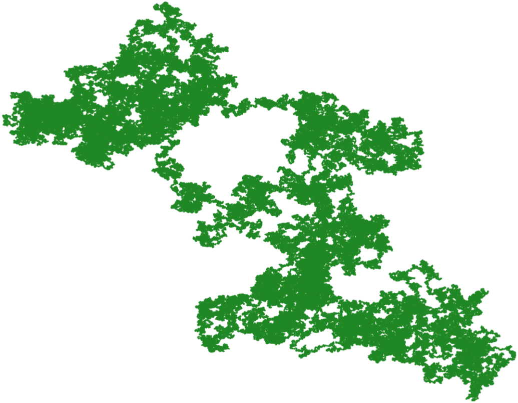
Let $d \geq 3$, and consider Brownian motion in $\mathbb{R}^d$ run for time 1.

Given the occupation measure of the path projected onto the sphere, you can recover the range and the endpoint with probability 1.

Conjecture (PPPY '00)

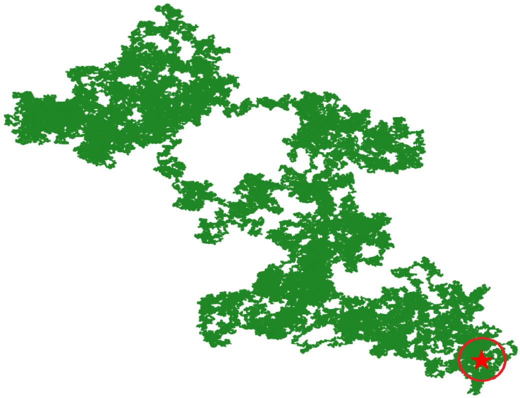
In dimension $d = 2$, the range cannot be recovered with probability 1.

SRW in $\mathbb{Z}^d$



Q: where is the starting point?

SRW in $\mathbb{Z}^d$



R_t = range of SRW up to time t.

Definition

An estimator $\hat{v}$ is a function

$$\hat{v} : (\Omega, \Xi) \to \mathbb{Z}^d,$$

where Ω is the space of simple random walk trajectories and $\hat{v}(\omega) \in \omega$ for every ω, and Ξ is uniform$(0, 1)$ independent of everything else.

R_t = range of SRW up to time t.

Definition

An estimator $\hat{v}$ is a function

$$\hat{v} : (\Omega, \Xi) \to \mathbb{Z}^d,$$

where Ω is the space of simple random walk trajectories and $\hat{v}(\omega) \in \omega$ for every ω, and Ξ is uniform$(0,1)$ independent of everything else.

Example: $\hat{v}(\omega)$ = uniform random closest point to the center of mass of ω.

R_t = range of SRW up to time t.

Definition

An estimator $\hat{v}$ is a function

$$\hat{v} : (\Omega, \Xi) \to \mathbb{Z}^d,$$

where Ω is the space of simple random walk trajectories and $\hat{v}(\omega) \in \omega$ for every ω, and Ξ is uniform$(0, 1)$ independent of everything else.

Example: $\hat{v}(\omega) =$ uniform random closest point to the center of mass of ω.

Definition

For $v \in \mathbb{Z}^d$ and $\omega \in \Omega$, the quenched likelihood of (v, ω) is

$$L(v, \omega) = \mathbb{P}(R^v = \omega),$$

where R^v is an independent copy of R started from v.

How to measure the strength of an estimator?

Definition

The (annealed) detection probability of an estimator $\hat{v}$ is

$$\mathrm{Detect}(\hat{v}) = \mathbb{P}(R^{\hat{v}(R)} = R) = \int_\Omega L(\hat{v}(\omega), \omega) d\omega$$

Definition

For $v \in \mathbb{Z}^d$ and $\omega \in \Omega$, the quenched likelihood ratio of (v, ω) is

$$\mathrm{Ratio}(v, \omega) = \frac{L(v, \omega)}{\sum_{u \in \omega} L(u, \omega)},$$

and the annealed likelihood ratio of an estimator $\hat{v}$ is

$$\mathrm{Ratio}(\hat{v}) = \int_\Omega \mathrm{Ratio}(\hat{v}(\omega), \omega) d\omega.$$

Theorem (Hoffman, R. '19)

The following hold for SRW in $\mathbb{Z}^d$ *as* $t \to \infty$.

i. For $d = 2$,

$$\sup_{v \in R} Ratio(v, R) \to_p 0.$$

ii. For $d \in \{3, 4, 5, 6\}$, *there exists an estimator* $\hat{v}$ *such that*

$$Detect(\hat{v}) \geq \Theta(t^{-c_d})$$

for some constant $c_d \in (0, 1)$.

iii. For $d \geq 7$, *there exists an estimator* $\hat{u}$ *such that*

$$Detect(\hat{v}) = \Theta(1).$$

Conjecture

$$Detect(\hat{v}_{MLE}) = \begin{cases} o(1), & d = 2 \\ \Theta(1), & d \geq 5 \end{cases}$$

Theorem (Ray, R., 22+)

The following holds for SRW on the d-regular tree. There exists an estimator $\hat{v}$ *such that: for all* $\epsilon > 0$ *there exists* $\delta > 0$ *and a sequence of sets* $A_t \subset \Omega_t$ *such that* $\liminf_t \mathbb{P}(R_t \in A_t) \geq 1 - \epsilon$*, and*

$$\liminf_t \inf_{\omega_t \in A_t} Ratio(\hat{v}(\omega_t), \omega_t) > \delta.$$

Theorem (Ray, R., 22+)

The following holds for SRW on the d-regular tree. There exists an estimator $\hat{v}$ such that: for all $\epsilon > 0$ there exists $\delta > 0$ and a sequence of sets $A_t \subset \Omega_t$ such that $\liminf_t \mathbb{P}(R_t \in A_t) \geq 1 - \epsilon$, and

$$\liminf_t \inf_{\omega_t \in A_t} Ratio(\hat{v}(\omega_t), \omega_t) > \delta.$$

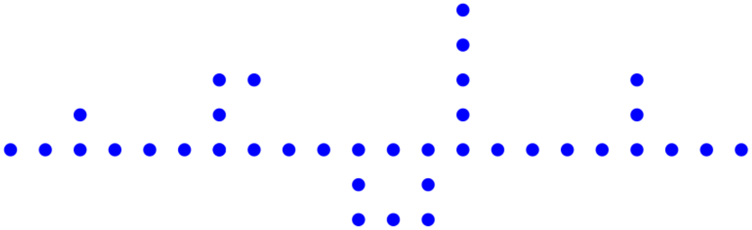
A similar result should hold for SRW on a random d-regular graph on $[n]$, run up to time $t = n^{1-\gamma}$ for any $\gamma > 0$.

Further Q's:

- Explicit formulas, biased RW on $\mathbb{Z}$
- Performance of 'longest path' estimator for transient RW's
- Good estimator for $\mathbb{Z}^3$

Proof ideas:

1. Get rid of the 'middle' of the range, using transience.
2. Infer chronological info using 'cut points.'



Proof sketch:

1. Get rid of the 'middle' of the range, using transience.
2. Infer chronological info using 'cut points.'

Ingredients:

1. Long cycles: return probabilities / self-intersection exponents (Lawler)

Ingredients:

1. Long cycles: return probabilities / self-intersection exponents (Lawler)
2. A cut time for X is a time $s \in [0, t]$ such that

$$X_{[0,s)} \cap X_{(s,t]} = \emptyset$$

If s is a cut time, X_s is called a cut point.

Theorem (James, Peres, '96)

In dimension $d \geq 3$, there are infinitely many cut times. In dimension $d \geq 5$, cut times have positive density.

Cutpoints are totally ordered (by their cut times).

Given all the cut points, find the ‘first’ and ‘last’ ones, pick uniformly from their small components.

Cutpoints are totally ordered (by their cut times).

Given all the cut points, find the ‘first’ and ‘last’ ones, pick uniformly from their small components.

Problem: not all ‘divider’ points are cut points!

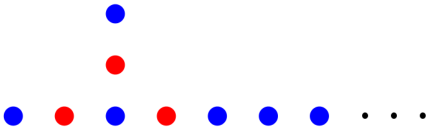


Figure: The three red ‘divider’ points can’t all be cut points.

Cutpoints are totally ordered (by their cut times).

Given all the cut points, find the 'first' and 'last' ones, pick uniformly from their small components.

Problem: not all 'divider' points are cut points!

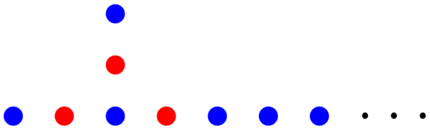


Figure: The three red 'divider' points can't all be cut points.

Need more information about how cutpoints are distributed.