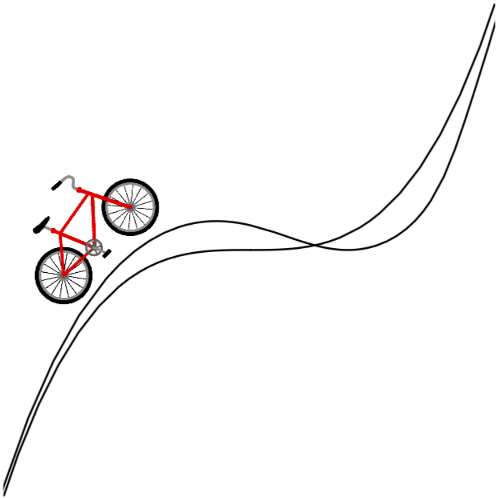
Hiding and finding the source

Jacob Richey

joint with: Miki Racz, Chris Hoffman, Gourab Ray

Cornell, May 2023

‘Geometry and Imagination’ (Conway, Doyle, Gilman, Thurston)



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 n, then guess the starting point.

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

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

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

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

Which was the most likely starting point?

A: They're all equally likely!

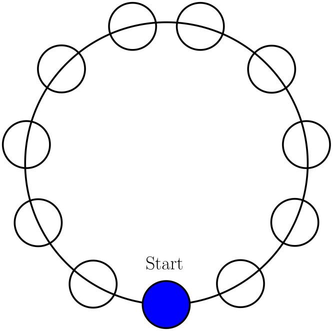
A: They're all equally likely!

Re-index SRW by record times, compute explicitly.

A: They're all equally likely!

Re-index SRW by record times, compute explicitly.

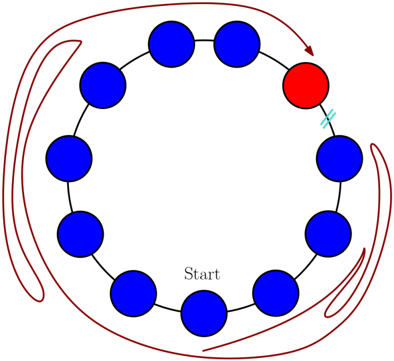
OR: last vertex visited by SRW on the ring is uniform.



A: They're all equally likely!

Re-index SRW by record times, compute explicitly.

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



Consider an infection spreading on the d-regular tree, $d \geq 3$

- The infection starts from site $v^* =$ 'patient zero'
- Infected sites can infect neighbors (with a speed limit)
- Observer sees all infected sites at a fixed (large) time

Classical example: SI (susceptible/infected)

Consider an infection spreading on the d-regular tree, $d \geq 3$

- The infection starts from site $v^* =$ 'patient zero'
- Infected sites can infect neighbors (with a speed limit)
- Observer sees all infected sites at a fixed (large) time

Classical example: SI (susceptible/infected)

The infection is spread by a random algorithm known to the observer

Observer makes their best guess for patient zero using a single snapshot

G_t = set of infected sites at time t

Maximum likelihood estimator

For any set $A \subset \mathbb{T}_d$,

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

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

$\mathbb{P}(G_t^v = A) = L(v, A)$ is called the (quenched) 'likelihood'

Detection probability

The observer correctly identifies the source with probability

$$\mathbb{P}(\hat{v}_{MLE}(G_t) = v^*)$$

Motivation: protecting user anonymity in a computer network

Goals for the rumor/infection spreading algorithm:

- *Spreading*: spread to many sites
- *Obfuscation*: minimize the detection probability for patient zero
- *Multiple observations*: obfuscate even if observer has > 1 independent observations

Motivation: protecting user anonymity in a computer network

Goals for the rumor/infection spreading algorithm:

- *Spreading*: spread to many sites
- *Obfuscation*: minimize the detection probability for patient zero
- *Multiple observations*: obfuscate even if observer has > 1 independent observations
- *Local spreading* (new): spread to all sites near patient zero

Previous results: SI model, rumor centrality

Theorem (Shah, Zaman, '10)

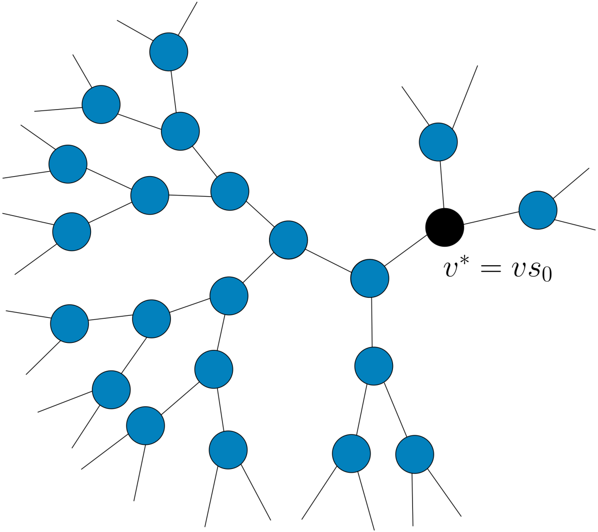
Consider the SI spreading model on the d-regular tree for $d \geq 3$. The detection probability is bounded away from 0 as $t \to \infty$.

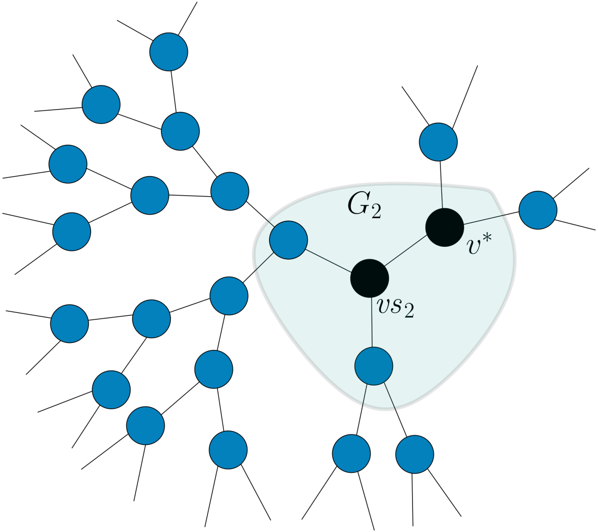
Fast spread and local spread, but no obfuscation.

Similar results for SI model on random trees.

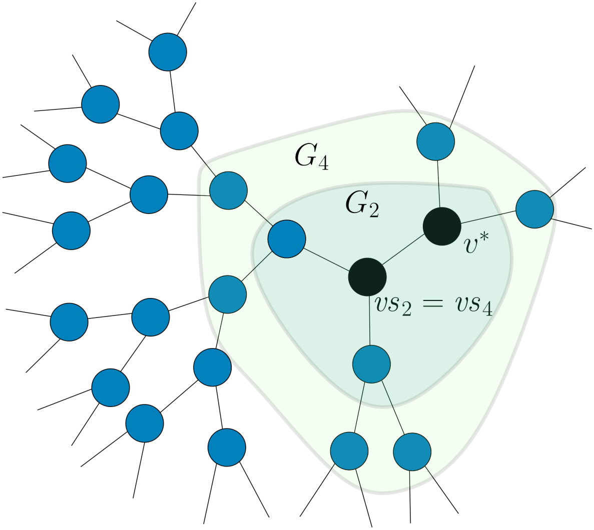
New class of random spreading algorithms: adaptive diffusions

- vs_t = virtual source at time t
- Every two time units, the virtual source either stays put or moves to a neighboring site
- When the virtual source moves, it chooses uniformly among the $d - 1$ options away from v^*
- **G_t is a ball of radius $t/2$ centered at vs_t at even times t**
- Characterized by transition probabilities for the virtual source

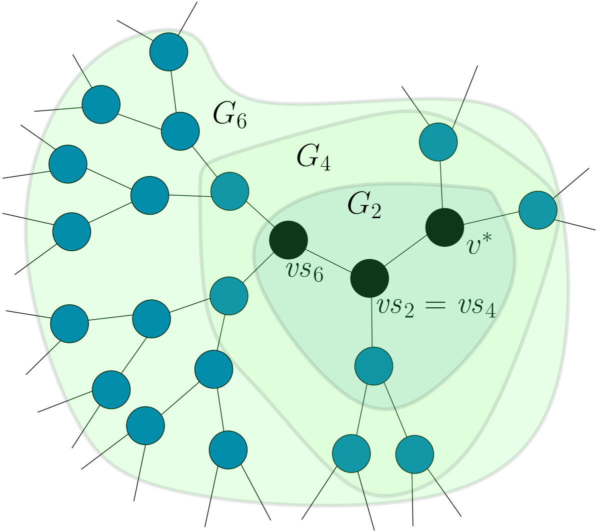

$v^* = vs_0$



G_2
v^*
vs_2



G_4
G_2
v^*
$vs_2 = vs_4$


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

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

$$\mathbb{P}(\hat{v}_{MLE} = v^*) = \begin{cases} \Theta(N_t^{-1}) & \text{perfect obfuscation} \\ \Theta(N_t^{-\gamma}) & \text{polynomial obfuscation} \\ o(1) & \text{weak obfuscation} \\ \Theta(1) & \text{no obfuscation} \end{cases}$$

SI: good spread and local spread, no obfuscation. [Shah, Zaman '10]

SI: good spread and local spread, no obfuscation. [Shah, Zaman '10]

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: good spread and local spread, no obfuscation. [Shah, Zaman '10]

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

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

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

Pf sketch: Choose transition probabilities for the virtual source so that it is uniformly distributed over a ball

Local spreading?

Local spreading?

Definition

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

The adaptive diffusion algorithm that achieves perfect obfuscation has constant order local spread, $R_t = \Theta(1)$ – no local spread!

Spreading/obfuscation trade-off [Racz, R. '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 average 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**.

The trade-off is essentially tight:

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

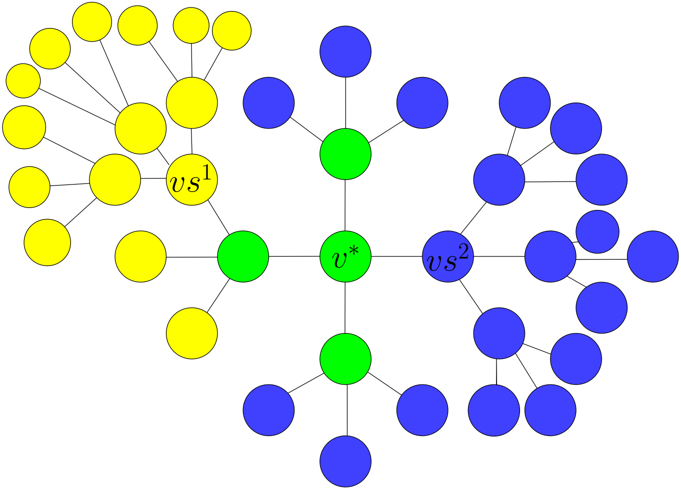
For every $\gamma \in (0,1)$, there exists an adaptive diffusion with both polynomial obfuscation of order γ,

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

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^*. For any t,

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

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:

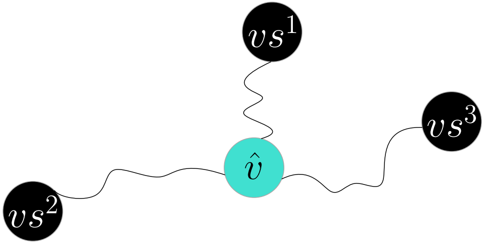
Three or more independent observations (Racz, Richey '18)

Suppose the observer has $k \geq 3$ iid snapshots G_t^i, $i \in [k]$ started from the same source v^*. For any t,

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

No obfuscation!

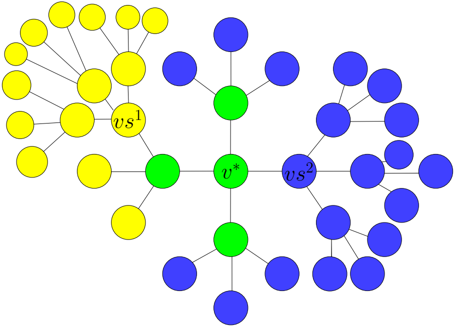
Proof: Pick any three virtual sources and draw the paths between them.



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

Necessary condition for obfuscation under multiple observations

Simple estimator: guess a green vertex



Question

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

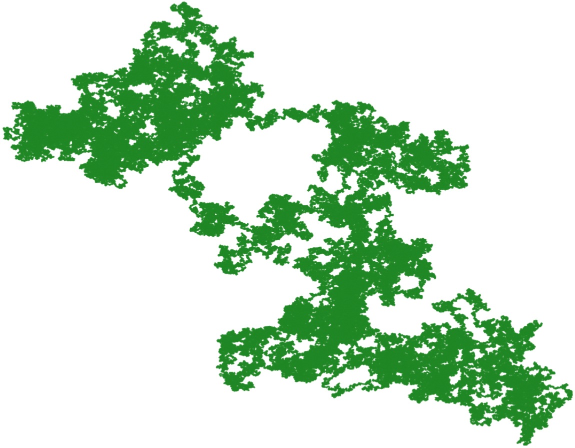
Should look at algorithms that have order-optimal local spreading:

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

RHS is large if local spread is typically small

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

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



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

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

Where did the Brownian particle go: given local time of BM on a sphere (Pemantle, Peres, Pitman, Yor '00)

Theorem

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

Where did the Brownian particle go: given local time of BM on a sphere (Pemantle, Peres, Pitman, Yor '00)

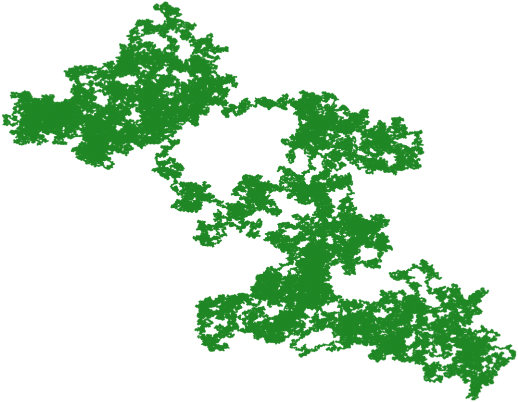
Theorem

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

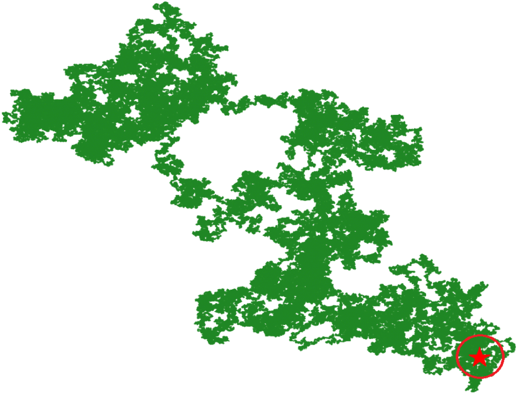
In dimension $d = 2$, the range cannot be recovered.

SRW in $\mathbb{Z}^d$



Q: where is the starting point?

SRW in $\mathbb{Z}^d$



A: there!

R_t = range of SRW up to time t, started from $0 \in \mathbb{Z}^d$

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 up to time t and $\hat{v}(\omega) \in \omega$ for every ω, and Ξ is uniform$(0, 1)$ independent of everything.

R_t = range of SRW up to time t, started from $0 \in \mathbb{Z}^d$

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 up to time t and $\hat{v}(\omega) \in \omega$ for every ω, and Ξ is uniform$(0, 1)$ independent of everything.

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

R_t = range of SRW up to time t, started from $0 \in \mathbb{Z}^d$

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 up to time t and $\hat{v}(\omega) \in \omega$ for every ω, and Ξ is uniform$(0,1)$ independent of everything.

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 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 detection probability of an estimator $\hat{v}$ is

$$\text{Detect}(\hat{v}) = \mathbb{P}(\hat{v} = 0).$$

Definition

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

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

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

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

Theorem (Hoffman, R. '19)

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

i. *For $d = 1$,*

$$Detect(\hat{v}_{MLE}) = \Theta(t^{-1/2}).$$

ii. *For $d = 2$,*

$$Ratio(\hat{v}_{MLE}, R) \to_p 0.$$

iii. *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 $c_d \in (0, 1)$, and $c_d = \frac{2}{d+2}$ for $d = 5, 6$.

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

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

Conjecture

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

Quenched detection result on a d-regular tree:

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

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

Quenched detection result on a d-regular tree:

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

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

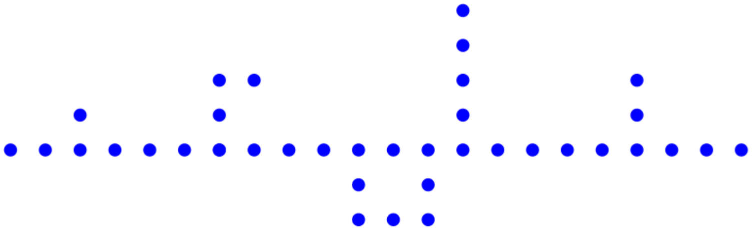
A similar result holds for SRW on a random d-regular graph on $[n]$, run up to time $t = n^{1-\gamma}$ for any $\gamma > 0$.

Todos:

- Biased RW on $\mathbb{Z}^d$
- 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, by bounding long returns.
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!

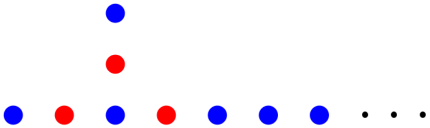


Figure: The three red ‘divider’ points can’t all be cut points.

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!

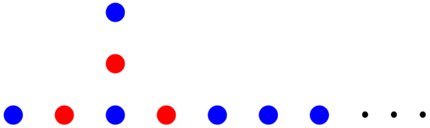


Figure: The three red 'divider' points can't all be cut points.

Need more information about how cutpoints are distributed.

Thanks!