Finding the source

Jacob Richey (Renyi)

joint with Miki Racz (Northwestern)

Lednice, October 2023

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 $\mathbb{Z}$.

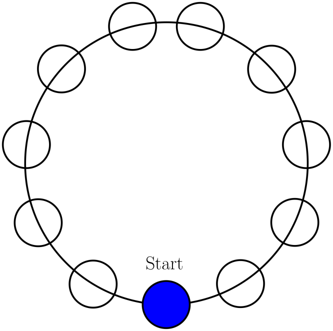
Problem: Run until the range has size n, then guess the starting point.

Which was the most likely starting point?

A: They're all equally likely!

Re-index SRW by record times, compute explicitly.

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



Random growth process on a connected graph $G = (V, E)$

- The source: start with $A_0 = \{v^*\}$
- Given A_t, randomly generate $A_{t+1} \supset A_t$
- Spread along edges of A, speed at most 1: $A_{t+1} \subset A_t \cup \partial A_t$

e.g. SI model: spread along each $e \in \partial_E A_t$ with probability p

Given a 'snapshot,' A_t at some (large) time t, try to guess v^*

A_t = set of infected sites at time t, started from $A_0 = \{v^*\}$

(Maximum) likelihood

For any set $A \subset G$, $v \in A$,

$$L(v|A) = \mathbb{P}(A_t = A | v^* = v).$$

Maximum likelihood estimator:

$$\widehat{v}_{ML} = \arg\max_{v \in A_t} L(v|A_t).$$

Often ML is hard to compute, can work with other estimators.

Detection probability

The observer correctly identifies the source with probability

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

Motivation: protecting privacy of metadata

Goals for the message 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

SI: edges pass information at rate 1 all independently

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 $t \to \infty$.

ML is described by 'rumor centrality':

$$R(v) = \prod_{w \in A_t} |T_w^v|.$$

T_w^v = subtree rooted at w 'away' from v

Fast spread and local spread, but no obfuscation.

Similar results for SI model on GW-trees; multiple observations on T_d

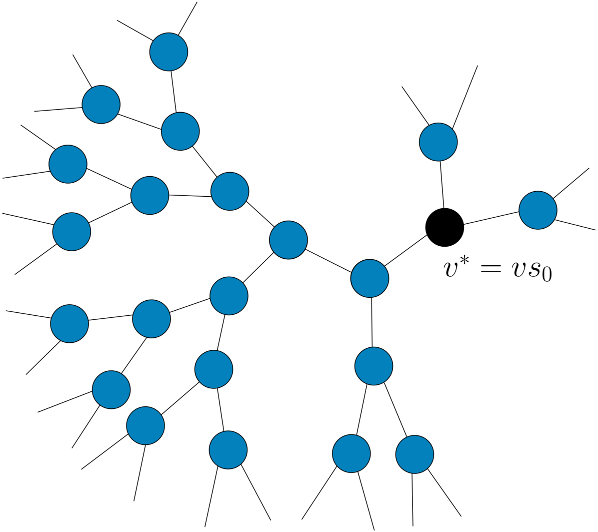
Adaptive diffusions: designed to hide the source

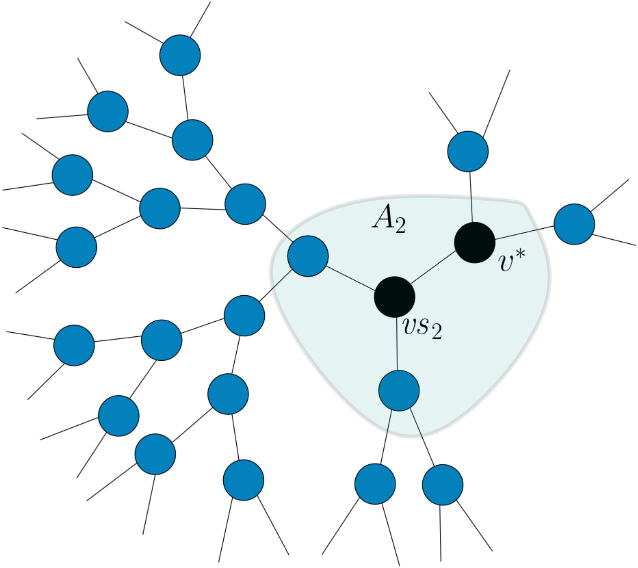
- Fix transition probabilities $h(t,x)$ for a random walk $H(t)$ on $\mathbb{Z}^{\geq 0}$:

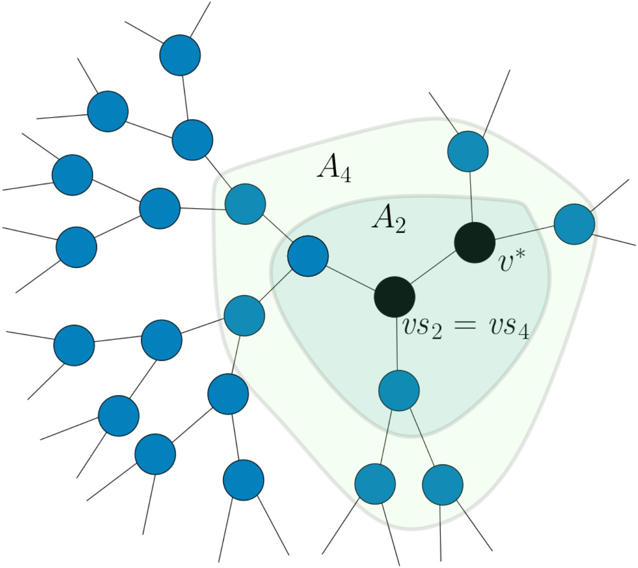
$$h(t,x) = \mathbb{P}(H(t+1) - H(t) = 1 | H(t) = x)$$

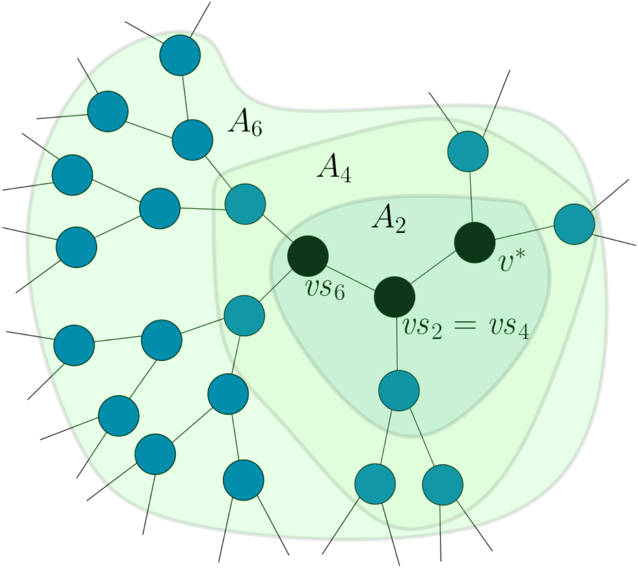
$$H(t) - H(t) \in \{0,1\} \text{ for all } t$$

- Evolve a single particle VS_t on G with $VS_0 = v^*$ and VS_t at depth $H(t)$ for all t (choose uniform child when stepping)
- For $t = 2, 4, 6, 8, \ldots$, $A_t =$ ball of radius $t/2$ in G centered at VS_t


$v^* = vs_0$


A_2
v^*
vs_2


A_4
A_2
v^*
$vs_2 = vs_4$


A_6
A_4
A_2
v^*
vs_6
$vs_2 = vs_4$

Spreading

For adaptive diffusion,

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

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

Detection

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

SI: good spread and local spread, perfect detection. [Shah, Zaman '10]

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

Let $G = T_d = d$-regular tree. There exists an adaptive diffusion algorithm that achieves no detection:

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

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

Local spreading?

Definition

The *local spread* $L(t)$ is the radius of the largest ball centered at v^* and contained in A_t.

The adaptive diffusion algorithm that cannot be detected has constant order local spread, $L(t) = \Theta(1)$ – no local spread!

Spreading/detection trade-off [Racz, R. '18]

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

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

Then the average local spreading is bounded from above:

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

Obfuscation (non-detection) and local spreading are **inversely linked** in this case.

The trade-off is essentially tight:

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

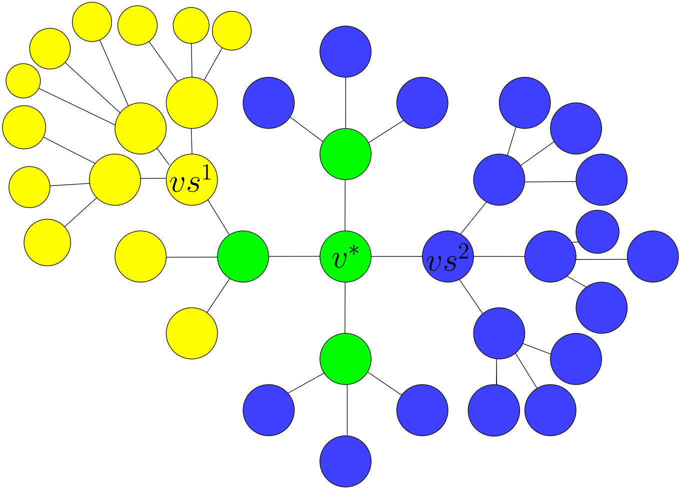
For every $\gamma \in (0,1)$, there exists an adaptive diffusion with both polynomial detection of order γ,

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

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 $\{A_t^i\}_{i=1}^k$ of the diffusion started from the same source v^*.



Two independent observations (Racz, R. '18)

Suppose the observer has two iid adaptive diffusion snapshots A_t^1 and A_t^2 started from the same source v^*. For any t,

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

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

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

Nearly perfect detection now!

It gets worse:

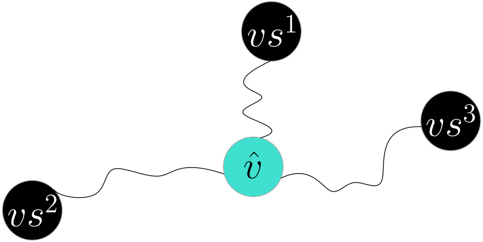
Three or more independent observations (Racz, R. '18)

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

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

Perfect detection!

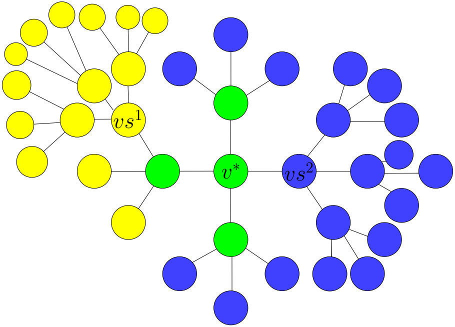
Proof: Pick any three virtual sources and draw the paths between them.



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 to hide the source under multiple observations

Simple estimator: guess a green vertex



Question

Does there exist a spreading algorithm that achieves order-optimal spreading and at most polynomial detection 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.

GW-trees? Real-world networks?