Root finding and broadcasting in random recursive trees

Gábor Lugosi
based on joint work with
Louigi Addario-Berry
Sébastien Bubeck
Luc Devroye
Vasiliki Velona

network archeology

Dynamically growing networks appear in social networks, epidemiology, rumor spreading, computer networks, protein networks, etc.

Often one only observes a present-day snapshot of the network.

What can one say about the past?

Possible questions:

- Who is patient zero?
- Who started a rumor?
- Have influential/central nodes always been central?
- What was the original configuration?

uniform and preferential attachment models

Complex networks can often be accurately modeled by simple random growth dynamics.

Nodes arrive one by one.

The new vertex attaches to one (or more) already present node at random.

The simplest model is uniform attachment: the vertex to attach to is chosen uniformly at random.

Often large networks have very uneven degree distribution—"Power-law networks."

These may be accurately modeled by preferential attachment models: the vertex to attach to is chosen at random, with probability proportional to the degree of the vertex.

uniform and preferential attachment trees

We consider the simplest possible network models—trees.

The arriving vertex selects one vertex to attach to.

The uniform attachment tree is sometimes called Uniform Random Recursive Tree.

The preferential attachment three is also known as the Plane-Oriented Random Recursive Tree.

finding adam

Upon observing a large unlabelled tree of size n, we wish to identify the root of the tree.

Is this possible? In what sense? Vertex 1 and vertex 2 are indistinguishable.

We are allowed to select a set of vertices. The root should be among them with high probability.

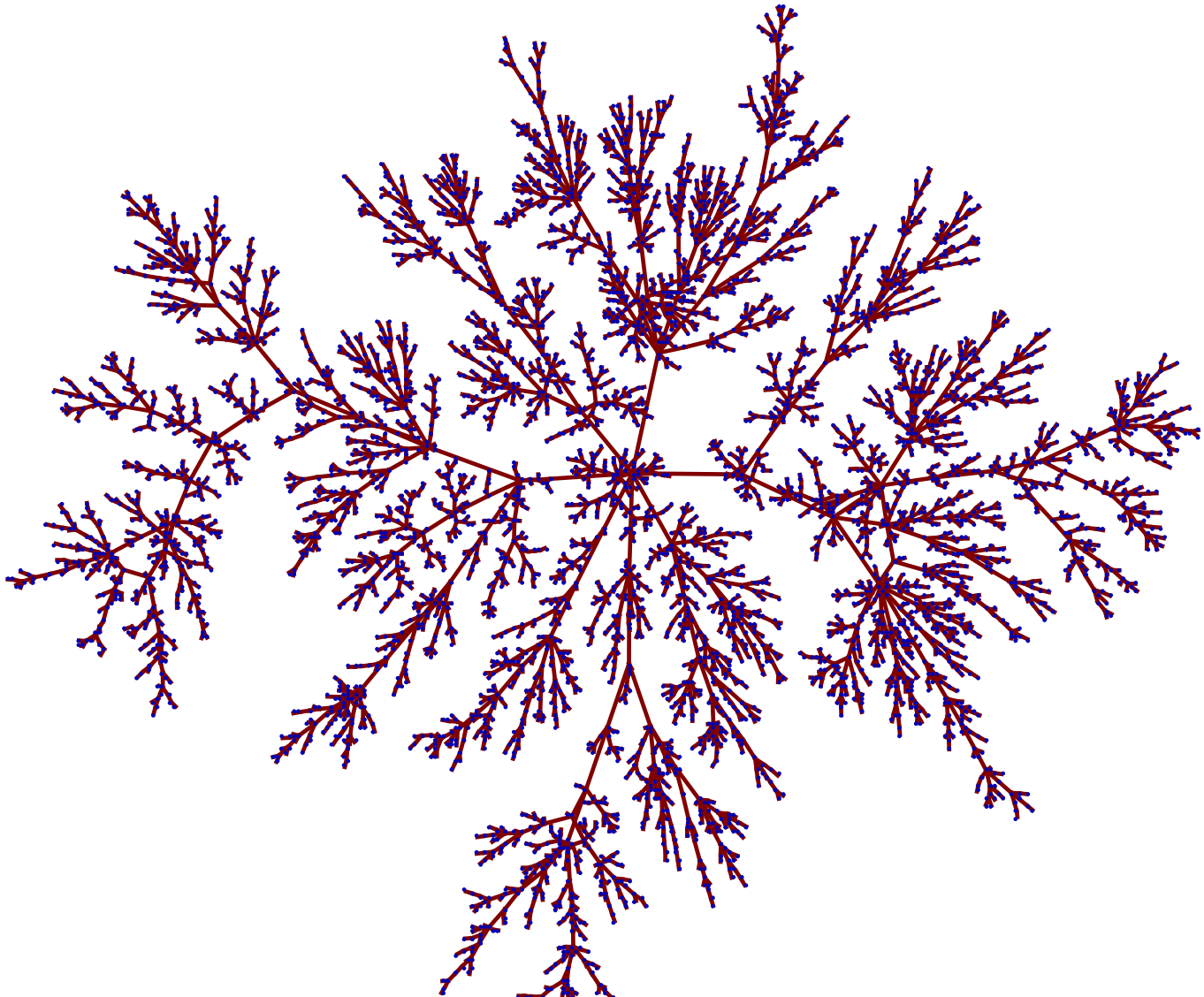
Formal setup. Given $\epsilon > 0$, select a set $S(\epsilon)$ of $K = K(n, \epsilon)$ vertices such that

$$\mathbb{P}\{\text{root} \in S(\epsilon)\} \geq 1 - \epsilon \, .$$

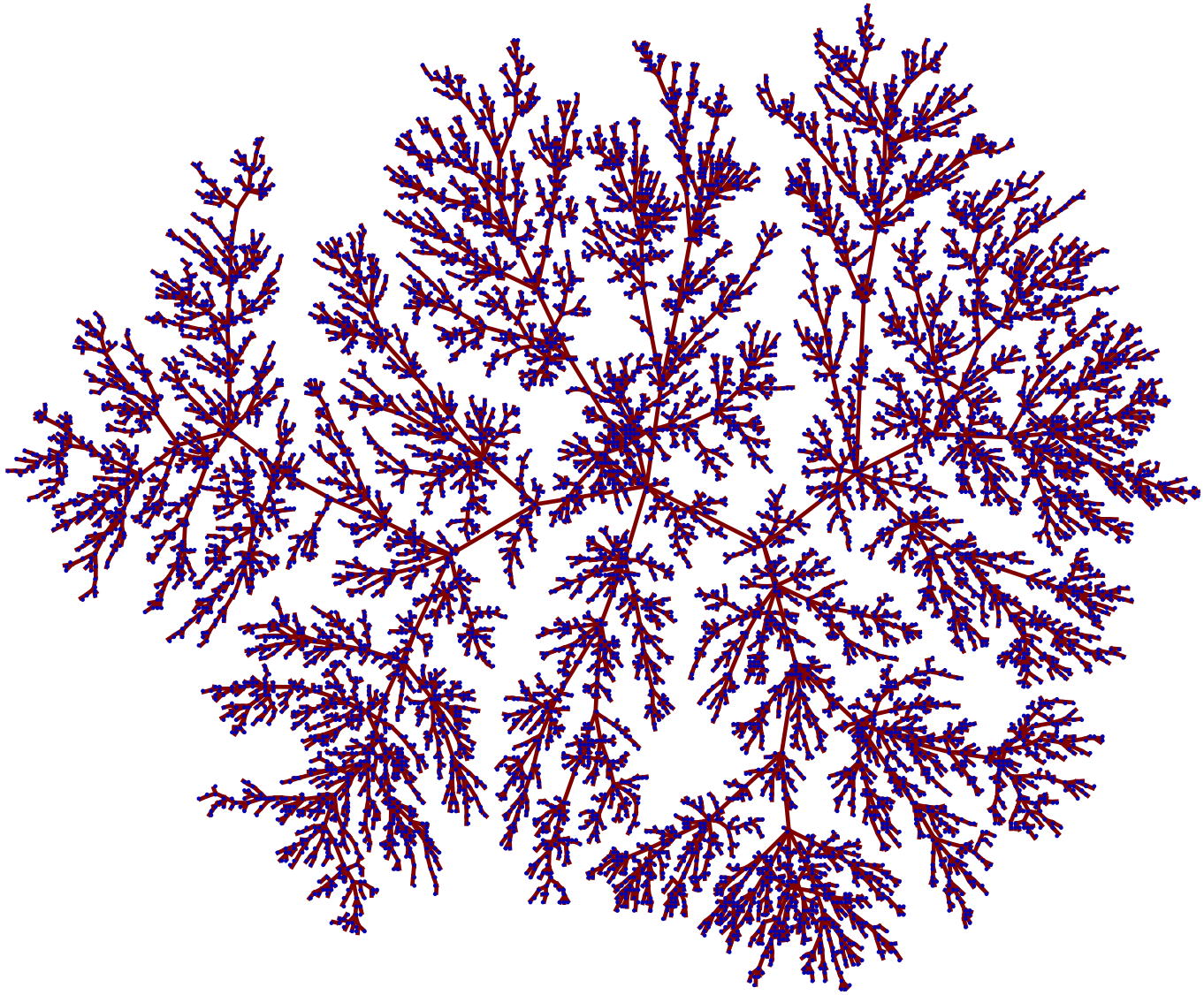
How large does $K(n, \epsilon)$ have to be?

uniform attachment tree UA(5000)

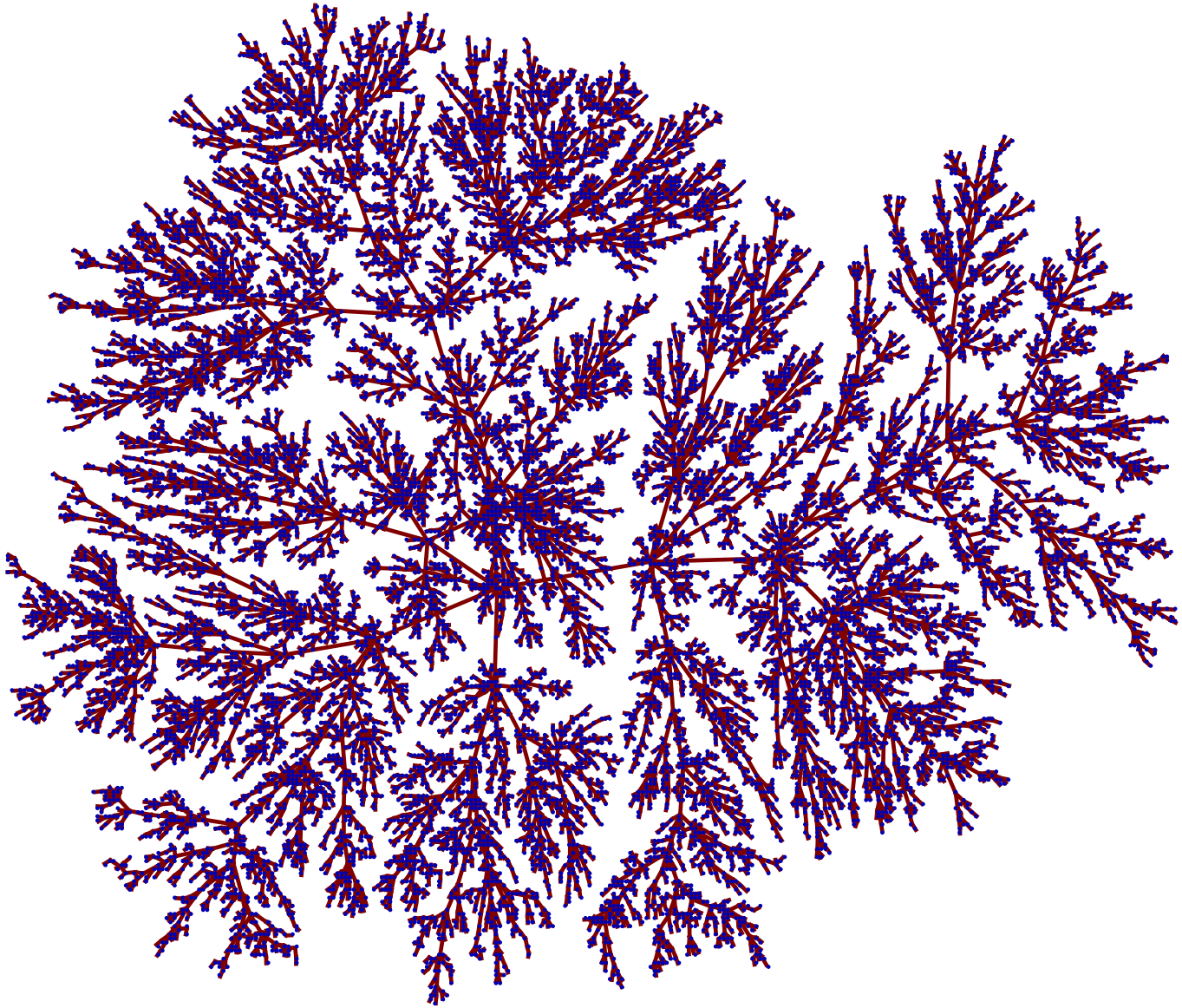
Thanks to Igor Kortchemski for the pictures.



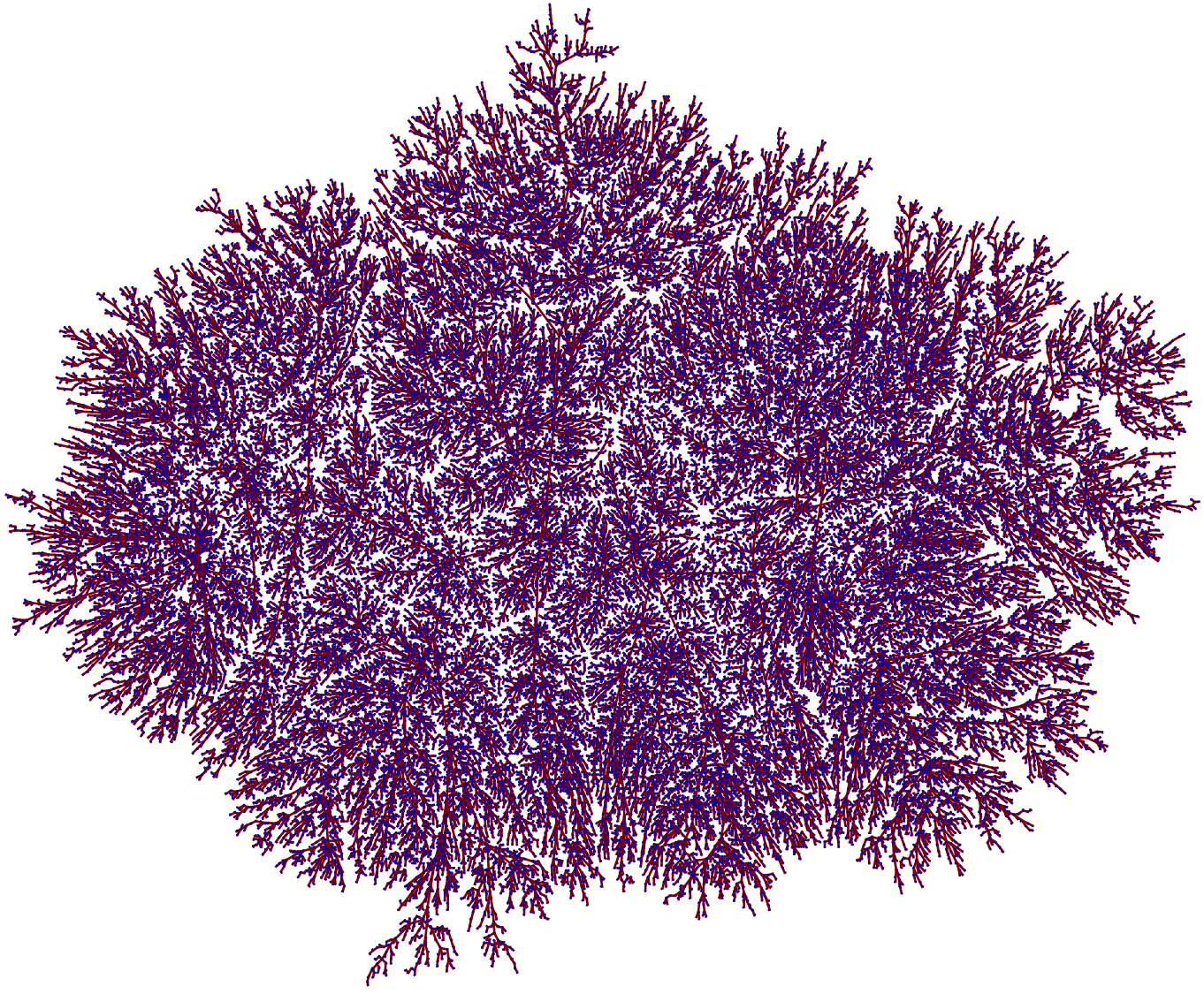
uniform attachment tree UA(10000)



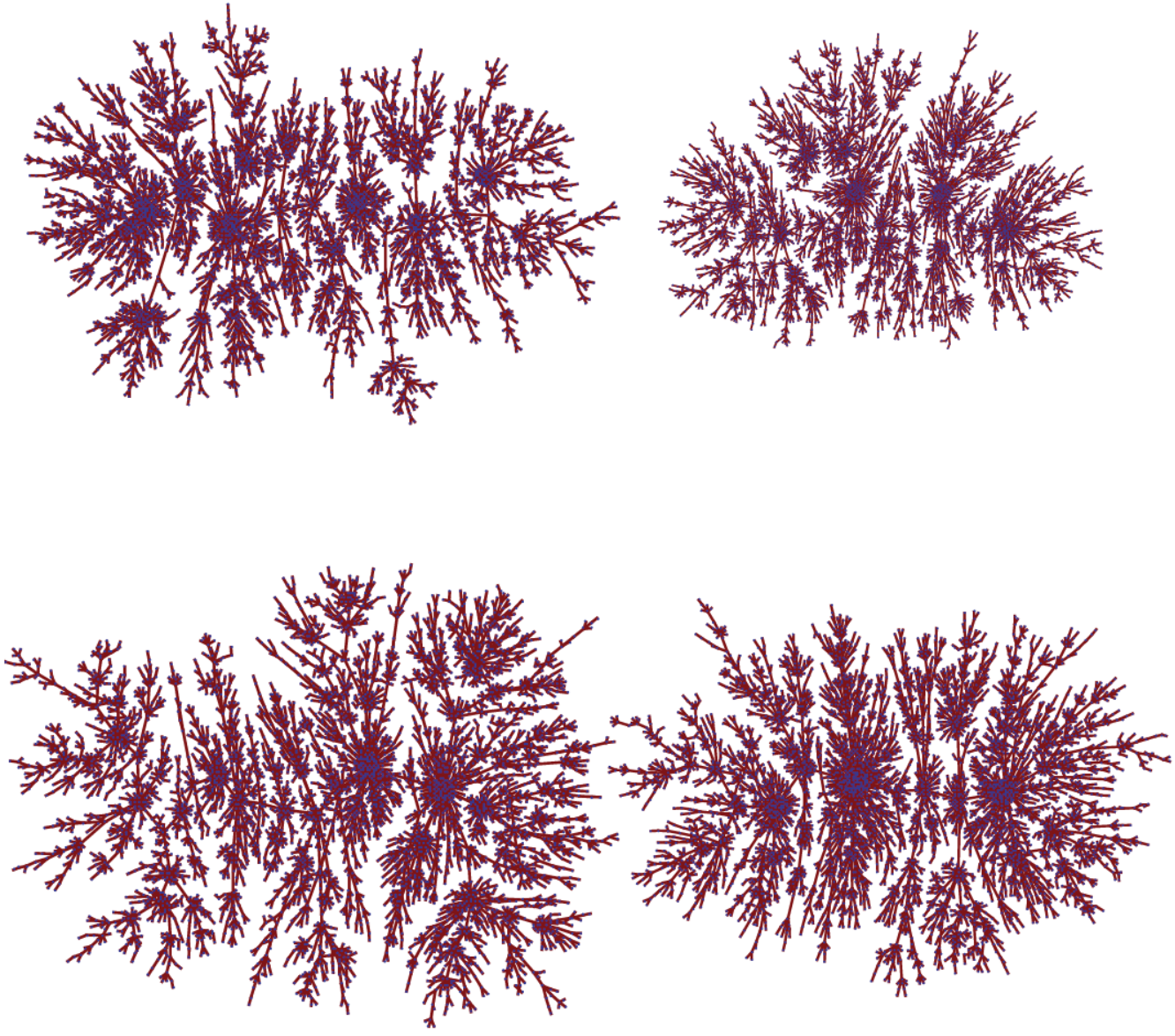
uniform attachment tree UA(15000)



uniform attachment tree UA(50000)



preferential attachment trees



results

In both models $K(n, \epsilon)$ is independent of n.

We obtain lower and upper bounds for $K = K(\epsilon)$ in both models.

Finding the root of a preferential attachment tree is harder than of a uniform attachment tree.

sorting by degrees

A simple idea is to include in $S(\epsilon)$ vertices with highest degree.

In a uniform attachment tree, the expected degree of the root is

$$1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n-1} \approx \ln n \ .$$

However, the maximum degree is much larger: $\approx \log_2 n$.

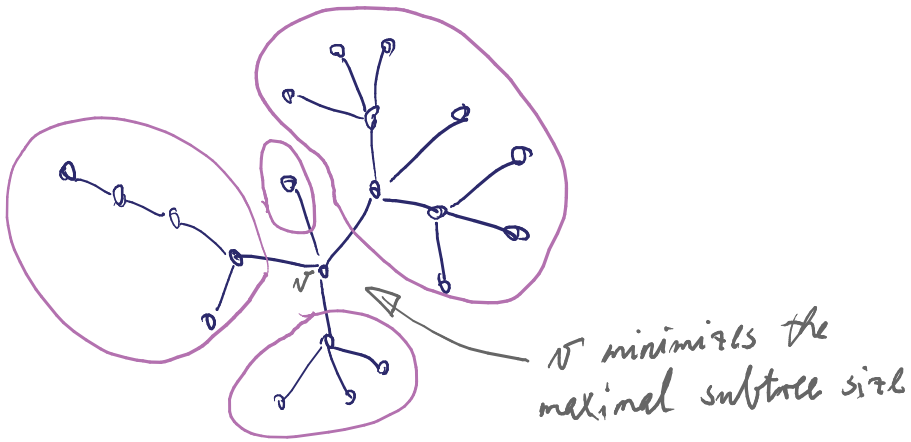
The index of the root in the ordering by degrees goes to infinity as $n \to \infty$.

In the preferential attachment model this idea works better but still not optimal. (Maximum degree is of the order of $\sqrt{n}$).

a simple method

Select K vertices with
smallest maximal subtree size.

Theorem. If
$K \geq 11 \log(1/\epsilon)/\epsilon$, then
$\mathbb{P}\{\text{root} \in S(\epsilon)\} \geq 1 - \epsilon$.



"Jordan centrality."

Proof sketch: Label the vertices chronologically. Let $\psi(i)$ denote the max subtree size of vertex i.

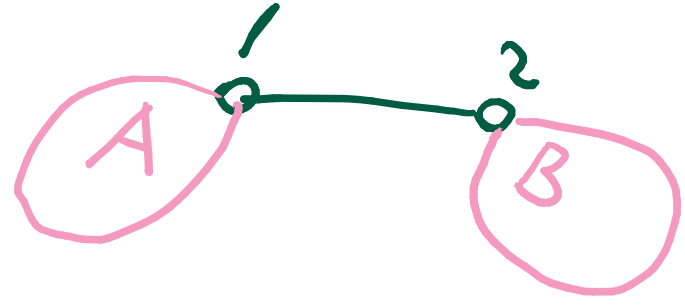
$$\mathbb{P}(1 \notin S(\varepsilon)) \le \mathbb{P}(\exists i > K : \psi(i) \le \psi(1))$$

$$\le \mathbb{P}(\psi(1) \ge (1-\varepsilon)n) + \mathbb{P}(\exists i > K : \psi(i) \le (1-\varepsilon)n)$$

First term:

$$\psi(1) \le \max(A, B) \approx n \cdot \max(U, V)$$

uniform $[0,1]$



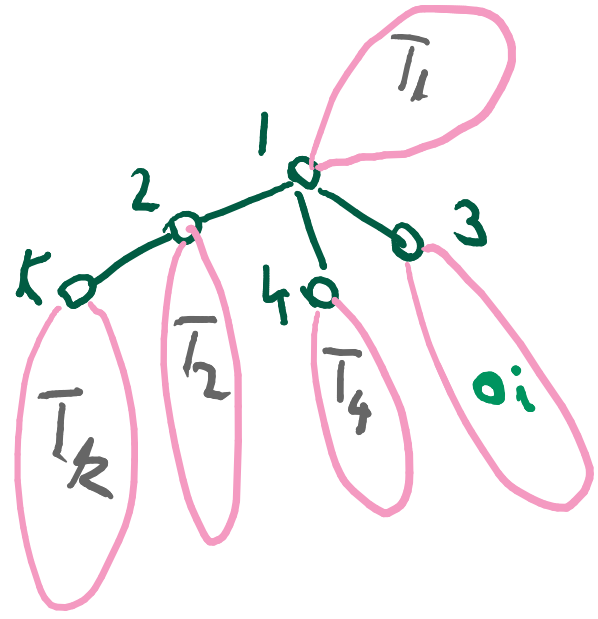
$$\mathbb{P}(\psi(1) \ge (1-\varepsilon)n) \approx 2\,\mathbb{P}(U \ge 1-\varepsilon) = 2\varepsilon$$

Second term:

$$\mathbb{P}(\exists i > K : \psi(i) \le (1-\varepsilon)n)$$

$$\le \mathbb{P}\Big(\exists k < K : \underbrace{\sum_{j \ne k} T_j}_{\approx n \cdot \mathrm{Beta}(K-1,1)} < (1-\varepsilon)n\Big)$$

$$\lesssim K(1-\varepsilon)^{K-1}$$



We have

$$\mathbb{P}(1 \notin S(\varepsilon)) \le 2\varepsilon + K(1-\varepsilon)^{K-1} \le 3\varepsilon$$

if $K \asymp O\left(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon}\right)$

minimum largest subtree size

The function ψ is monotone along paths towards the minimum.

Computation is easy (linear time). The algorithm outputs a tree.

The analysis above is essentially tight.

In a uniform attachment tree of size $3K$, the root is a leaf with probability

$$\frac{1}{2} \cdot \frac{2}{3} \cdot \cdots \cdot \frac{3K-2}{3K-1} = \frac{1}{3K-1}$$

Also, whp, there are about $3K/2$ leaves, so the simple method errs with probability $> \epsilon$ if $K < 1/(3\epsilon)$.

Can we do better?

maximum likelihood

A recursive tree is a rooted labeled tree. The root has label 1. Labels increase along paths away from the root.

By counting the number of recursive labelings of an unlabelled rooted tree, we may determine the maximum likelihood estimator of the root.

It minimizes the function

$$\zeta_T(u) = \prod_{v \in V(T)} |(T,u)_{v\downarrow}| \cdot \mathbf{Aut}(v,(T,u)).$$

Where $(T,u)_{v\downarrow}$ is the subtree of the tree T rooted at u, starting at v.

$\mathbf{Aut}(v,(T,u))$ involves automorphisms of subtrees.

general lower bound

We may use this to derive lower bounds for any method.

Theorem. If $K < \exp(\sqrt{(1/30)\log(1/2\epsilon)})$, then for any method outputting a set S of size K,

$$\mathbb{P}\{\text{root} \in S\} < 1 - \epsilon \, .$$

A recursive tree of $K+1$ vertices. The root has the smallest likelihood!

This configuration happens with probability $\geq \frac{1}{2} \exp(-30 \log^2 K) \geq \varepsilon$

a relaxation of maximum likelihood

Maximum likelihood is difficult to analyze.

The set of K vertices with largest likelihood does not form a tree.

We relax the likelihood criterion: instead of minimizing

$$\zeta_T(u) = \prod_{v \in V(T)} |(T,u)_{v\downarrow}| \cdot \mathrm{Aut}(v,(T,u)),$$

we minimize

$$\phi_T(u) = \prod_{v \in V(T)} |(T,u)_{v\downarrow}|$$

a relaxation of maximum likelihood

For this method a sub-polynomial value of K is sufficient:

Theorem. If $K \geq c \exp\left(\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}\right)$, then

$$\mathbb{P}\{\text{root} \in S(\epsilon)\} \geq 1 - \epsilon \ .$$

preferential attachment

Selecting K vertices with smallest maximal subtree size works again:

Theorem. If $K \geq C \log^2(1/\epsilon)/\epsilon^4$, then

$$\mathbb{P}\{\text{root} \in S(\epsilon)\} \geq 1 - \epsilon \, .$$

The bound is worse than for uniform attachment.

It cannot be improved by much: the probability that the root is a leaf in a tree of size $4K$ is

$$\frac{1}{2} \cdot \frac{3}{4} \cdot \ldots \cdot \frac{2(4K-2)-1}{2(4K-2)} \approx \frac{1}{\sqrt{4K}}$$

The number of leaves is about $8K/3$ so the method errs with probability $> \epsilon$ if $K < 1/(4\epsilon^2)$.

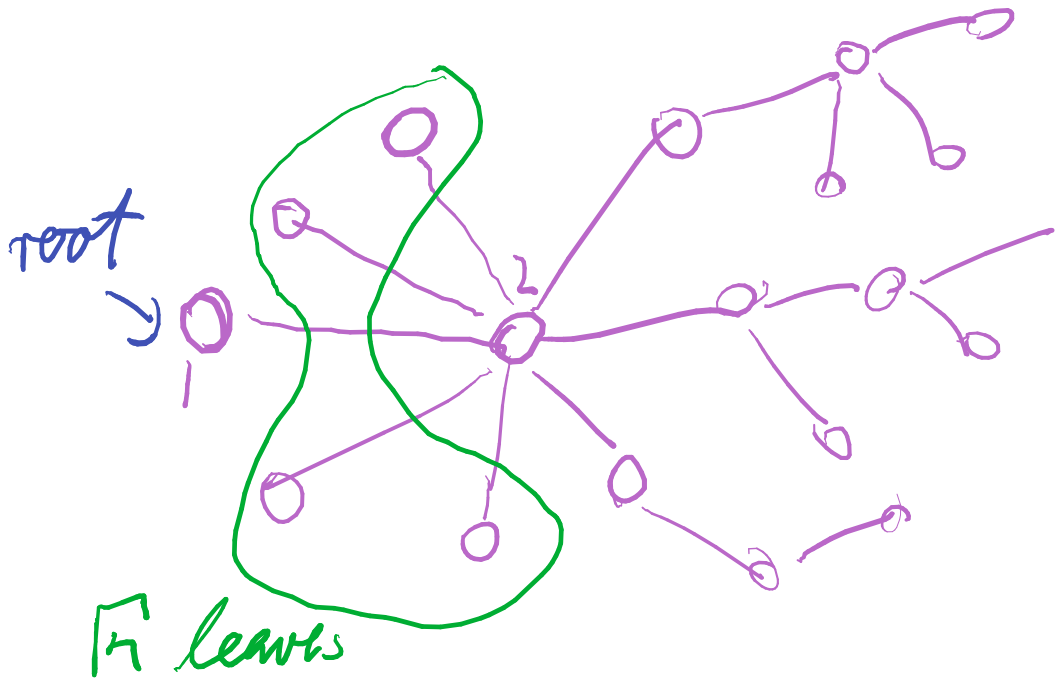
preferential attachment is harder

No sub-polynomial bounds are possible for preferential attachment.

Theorem. If $K < c/\epsilon$, then for any method outputting a set S of size K, $\mathbb{P}\{\text{root} \in S(\epsilon)\} < 1 - \epsilon$.

Proof: In a PA tree of size $n \sim 1/\epsilon^2$ the root is a leaf with probability ϵ and vertex 2 has $\sqrt{n}$ neighbors that are leaves.

The root is indistinguishable from $2K$ other vertices.



summary

The optimal value of $K(\epsilon)$ is between

$$\exp\left(\sqrt{\frac{1}{30}\log\frac{1}{2\epsilon}}\right) \quad \text{and} \quad c\exp\left(\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}\right)$$

for uniform attachment and between

$$\frac{c}{\epsilon} \quad \text{and} \quad \frac{C\log^2(1/\epsilon)}{\epsilon^4}$$

for preferential attachment.

broadcasting problem

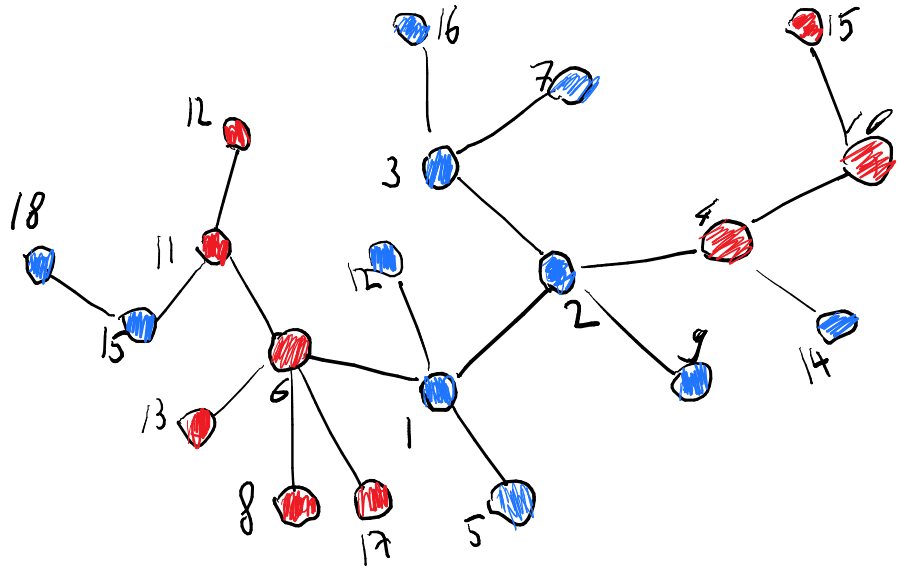
In a uniform or preferential attachment tree, vertices are colored red or blue.

The root has a random (unknown) color.

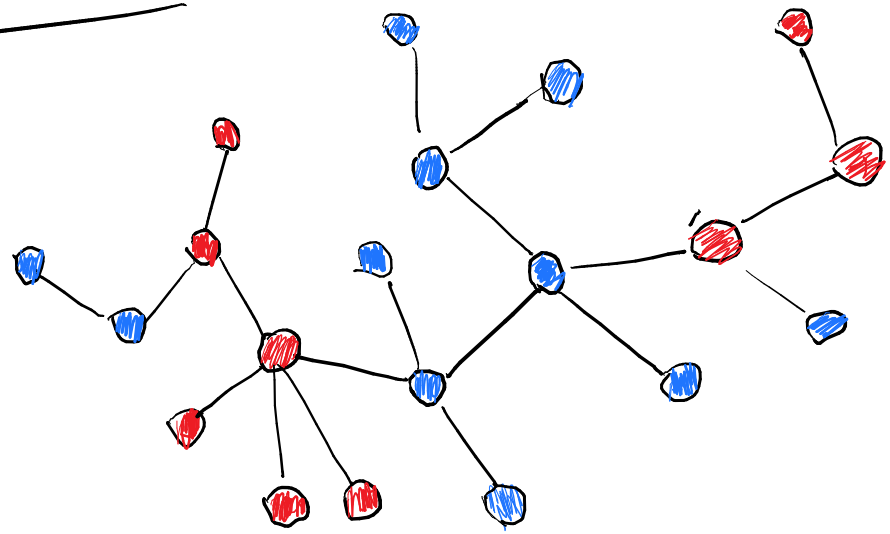
When a new vertex attaches to a parent, it takes the same color with probability $1 - q$ and the opposite color with probability q.

Classification problem: Upon observing an (unlabeled) tree of size n together with the vertex colors, guess the color of the root.

Question: What is the optimal probability of error $R^*(q, n)$? How does it depend on n and q?

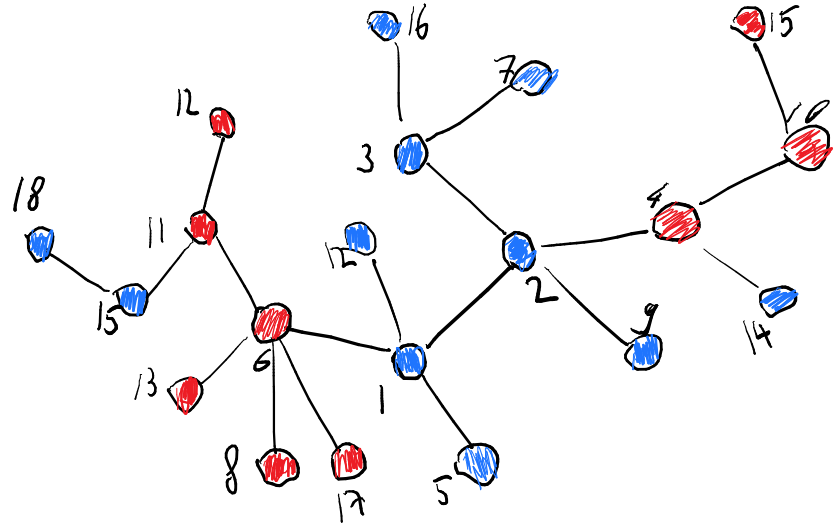


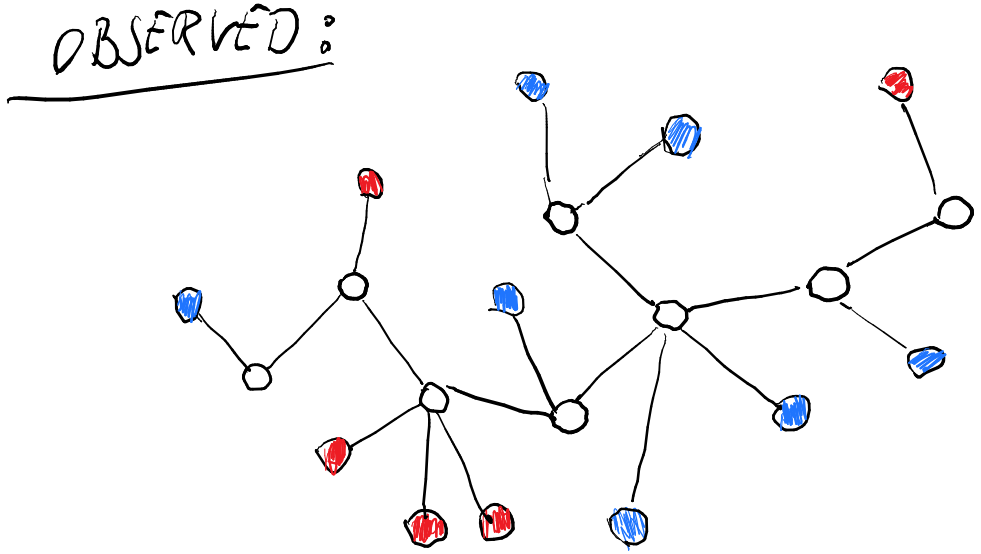
OBSERVED:



a more difficult variant

Sometimes only the colors of the leaves can be observed.
This is the problem of broadcasting on trees.





colored trees

Obviously, $R^*(q, n) \geq q/2$ for all n.

colored trees

Optimal classifier: We say blue if

$$\sum_{\text{blue vertices } u} \zeta_T(u) > \sum_{\text{red vertices } u} \zeta_T(u)$$

where

$$\zeta_T(u) = \prod_{v \in V(T)} |(T,u)_{v\downarrow}| \cdot \mathbf{Aut}(v,(T,u)),$$

is the number of recursive labelings of the tree with u as root. Difficult to analyze. We need simple classifiers.

root-finding

One may use root-finding algorithms.

Let $S(\epsilon)$ be a subtree of $K(\epsilon)$ vertices that contains the root with probability $1-\epsilon$,

Take a majority vote over the vertex colors of $S(\epsilon)$.

The probability of error is at most

$$\begin{aligned}\epsilon + \mathbb{P}\{S(\epsilon) \text{ is not monochromatic}\} &\leq \epsilon + \left(1-(1-q)^{K(\epsilon)-1}\right) \\ &\leq \epsilon + qK(\epsilon)\end{aligned}$$

Optimizing in ϵ, we obtain that

$R^*(q,n) \leq Cq^{1/\log\log q}$ for uniform attachment and

$R^*(q,n) \leq Cq^{1/4}\log(1/q)$ for preferential attachment.

small distance to the root helps

The analysis above cannot be improved if $S(\epsilon)$ is a path.

Suppose that by observing the uncolored tree, we choose vertex v.
Denote $D = \text{distance}(v, \text{root})$.
Choose the color of v.

Then the probability that the chosen color is different from the root is

$$\begin{aligned} \mathbb{E}\mathbb{1}_{\{\text{Bin}(D,q) \text{ is odd}\}} &= \frac{1 - \mathbb{E}(-1)^{\text{Bin}(D,q)}}{2} \\ &= \frac{1 - \mathbb{E}(1-2q)^{D}}{2} \leq q\mathbb{E}D \end{aligned}$$

Can we find a vertex close to the root? Only expected distance matters!

root-finding with proximity

Lemma.
Let v^* be vertex that minimizes the maximal subtree size. Then in both uniform and preferential attachment trees, the root is within distance $L = C \log(1/\epsilon)$ of v^*, with probability at least $1 - \epsilon$.
In particular,

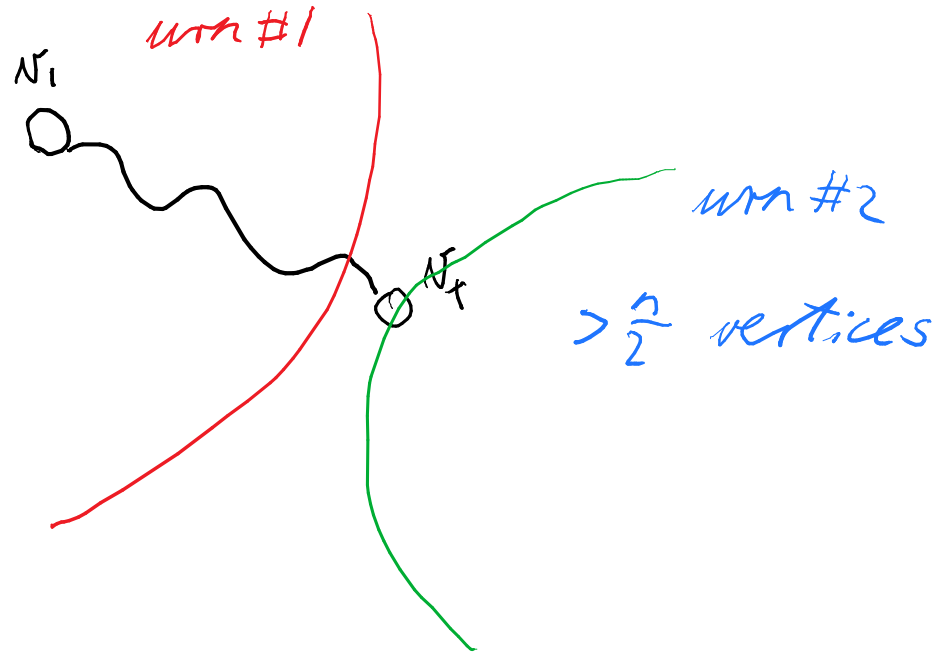
$$\mathbb{E}\ \text{distance}(v^*, \text{root}) \leq \frac{9}{2} + o_n(1)\ .$$

proof

For vertex v_t to be more central than the root, in a Pólya urn initialized with 1 white ball and $t-1$ black balls, after time n, there must be more than $n/2$ white balls.

The probability of this is at most $t2^{-(t-2)}$.

So the probability that the central vertex has index $> C\log(1/\epsilon)$ is at most ϵ.



root-finding with small diameter

Then we may define $S(\epsilon)$ as the subtree of all vertices within distance $L = C \log(1/\epsilon)$ of v^*. This set has diameter at most $2L$.

It follows that in both uniform and preferential attachment models,

$$\frac{q}{2} \leq R^*(q, n) \leq Cq \ .$$

Moreover,

$$R^*(q, n) < \frac{1}{2}$$

whenever $q < 1/2$.

when only leaf-colors are observed

The same bounds hold in the more difficult model when only the colors of the leaves are observed.

We only need the fact that there is at least one leaf close to v^*. But this holds since there are leaves close to the root.

majority

The simplest classifier takes the majority of the observed colors.
One may prove that the probability of error satisfies

$$\overline{R}(q, n) \leq cq \ .$$

Moreover,

$$\lim_{n \to \infty} \overline{R}(q, n) \left\{ \begin{array}{ll} < 1/2 & \text{if } q < 1/4 \\ = 1/2 & \text{if } q > 1/4 \end{array} \right\}$$

when can we do better than random guessing?

The optimal probability of error $R^*(q, n)$ satisfies

$$\lim_{n\to\infty} R^*(q,n) \quad \begin{array}{ll} < 1/2 & \text{if } q < 1 \\ = 1/2 & \text{if } q = 1 \end{array}$$

when the colors of all vertices are observed.
When only leaf colors are available,

$$\lim_{n\to\infty} R^*(q,n) \quad \begin{array}{ll} < 1/2 & \text{if } q \in [0, 1/2) \cup (1/2, 1) \\ = 1/2 & \text{if } q \in \{1/2, 1\} \ . \end{array}$$

Refereces

S. Bubeck, L. Devroye, and G. Lugosi.
Finding Adam in random growing trees.
Random Structures and Algorithms, 2017.

L. Addario-Berry, L. Devroye, G. Lugosi, V. Velona.
Broadcasting on random recursive trees.
preprint, 2020.

related work

Rumor spreading model of Shah and Zaman (2011).

In a fixed graph–for example in an infinite d-regular tree–a rumor spreads according to a diffusion model.

Jog and Loh (2016) show that the central node eventually remains the same forever.

Extensions of root finding to nonlinear attachment: Banerjee and Bhamidi (2020)

Bubeck, Eldan, Mossel, and Rácz (2015) and Curien, Duquesne, Kortchemski, and Manolescu (2015) show that the process "never forgets" the initial (seed) tree.

Finding the seed tree: Devroye and Reddad (2019); Lugosi and Pereira (2019).

Broadcasting on trees: large body of literature. Evans, Kenyon, Peres, and Schulman (2000), Janson and Mossel (2004), Mossel (1998, 2001), Daskalakis, Mossel, and Roch (2006, 2011), Mézard and Montanari (2006)

questions

Other tree models: split trees, Galton-Watson trees, phylogenetic trees.

Models beyond trees: uniform and preferential attachment graphs, diffusion networks, etc.