Problems in network archaeology: root finding and broadcasting

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

network archaeology

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 (a function of) 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 simplest preferential attachment tree 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.

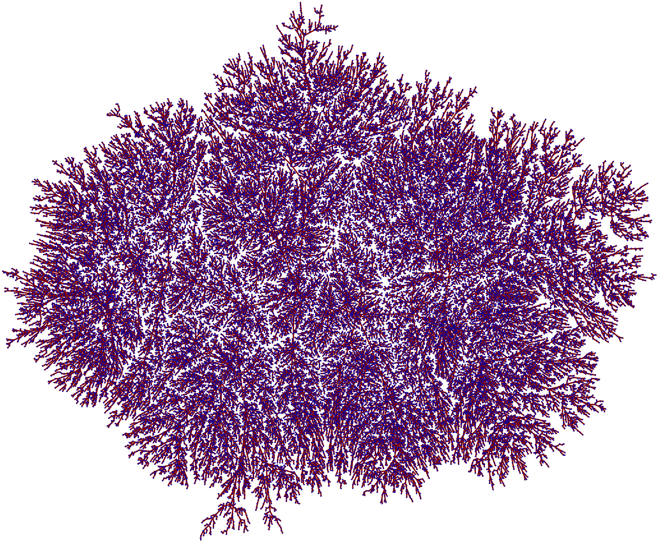
Formal setup. Given $\epsilon > 0$, select a set $S(\epsilon)$ of $K = K(n, \epsilon)$ vertices such that

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

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

uniform attachment tree UA(50000)

Picture by Igor Kortchemski.



results

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

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

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

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

sorting by degrees

A simple idea is to include in $\boldsymbol{S}(\epsilon)$ 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 $n \to \infty$.

Laura Eslava (2020) proves a central limit theorem for the distances of high-degree vertices from the root. For example, the vertices of the K highest degree all have depth $\sim c \log n$ for $c \approx 0.278652\ldots$

a simple method

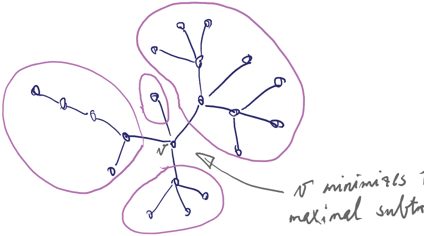
Select K vertices with smallest maximal subtree size.

Theorem. (Bubeck, Devroye, Lugosi (2017).) If $K \geq 11 \log(1/\epsilon)/\epsilon$, then $\mathbb{P}\{\text{root} \in S(\epsilon)\} \geq 1 - \epsilon$.

The winning vertex is a centroid of the tree.

"Jordan centrality."

The proof uses simple Pólya urn arguments.



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

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

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.

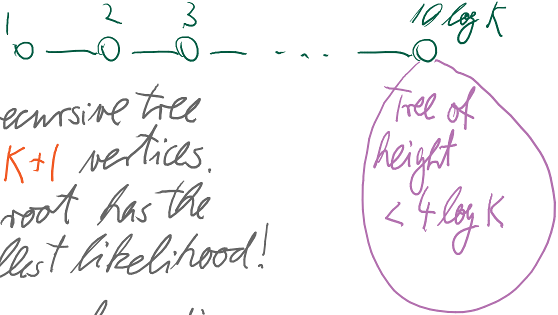
$\mathbf{Aut}(v,(T,u))$ involves automorphisms of subtrees.

general lower bound

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

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

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



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 $\boldsymbol{K}$ vertices with largest likelihood does not form a tree.

We relax the likelihood criterion: instead of minimizing

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

we minimize

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

Shah and Zaman (2011) call $\phi_T(u)$ rumor centrality.

Crane and Xu (2020) prove that $\phi_T(u)$ induces the same order of vertices as $\zeta_T(u)$.

sub-polynomial performance

For this method a sub-polynomial value of $\boldsymbol{K}$ is sufficient:

Theorem. (Bubeck, Devroye, Lugosi (2017).) If $\boldsymbol{K} \geq \boldsymbol{c} \exp\left(\frac{\log(1/\epsilon)}{\log\log(1/\epsilon)}\right)$, then

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

preferential attachment

Selecting K vertices with smallest maximal subtree size works again:

Theorem. Banerjee and Bhamidi (2020) improving Bubeck, Devroye, Lugosi (2017).) If $K \geq C\epsilon^{-(2+o(1))}$, 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 $K < 1/(4\epsilon^2)$.

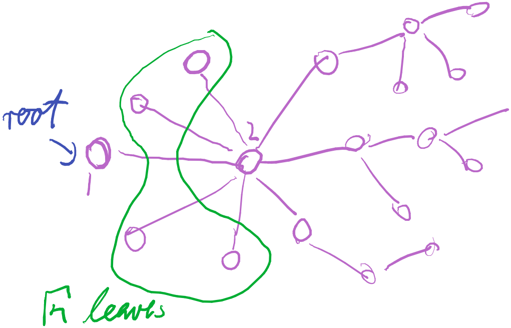
preferential attachment is harder

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

$$\frac{c}{\epsilon} \quad \text{and} \quad \frac{C}{\epsilon^{2+o(1)}}$$

for linear preferential attachment.

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

$$\frac{c}{\epsilon} \quad \text{and} \quad \frac{C}{\epsilon^{2+o(1)}}$$

for linear preferential attachment.

Recently, Alice Contat, Nicolas Curien, Perrine Lacroix, Etienne Lasalle, and Vincent Rivoirard (2023) closed the gap.

related work

Rumor spreading model of Shah and Zaman (2011). In a fixed graph–for example in an infinite $\boldsymbol{d}$-regular tree–a rumor spreads according to a diffusion model. See also Khim and Loh (2017).

Jog and Loh (2016) show that the central node eventually remains the same forever: "persistence of centrality".

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

Two papers of Banerjee and Bhamidi (2020) study persistence of centrality and persistence of vertices of highest degree in general nonlinear attachment trees.

seeded trees

Suppose that the tree is grown from a small initial "seed" tree. 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 problem

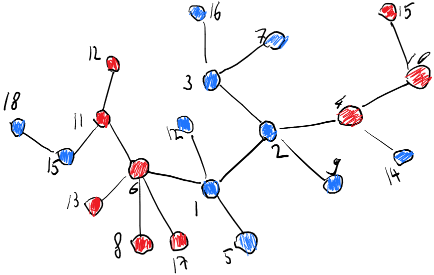
In a uniform or preferential attachment tree, vertices are colored red or blue.

The root has a random (unknown) color.

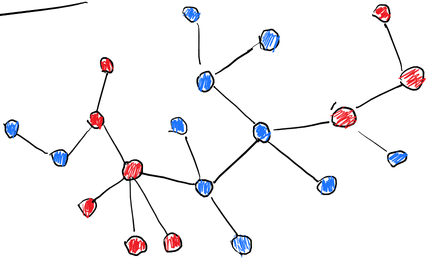
When a new vertex attaches to a parent, it takes the same color with probability $1 - \boldsymbol{q}$ and the opposite color with probability $\boldsymbol{q}$.

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

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



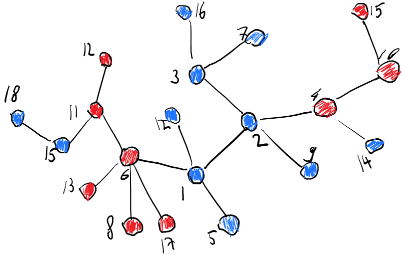
OBSERVED:



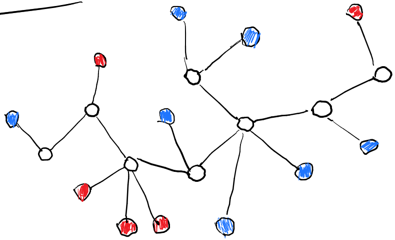
a more difficult variant

Sometimes only the colors of the leaves can be observed.

This is the problem of broadcasting on trees.



OBSERVED:



colored trees

L. Addario-Berry, L. Devroye, G. Lugosi, V. Velona.
Broadcasting on random recursive trees.
Annals of Applied Probability, 2022.

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

colored trees

Optimal classifier: We say blue if

$$\sum_{\text{blue vertices } \boldsymbol{u}} \zeta_{\boldsymbol{T}}(\boldsymbol{u}) > \sum_{\text{red vertices } \boldsymbol{u}} \zeta_{\boldsymbol{T}}(\boldsymbol{u})$$

where

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

is the number of recursive labelings of the tree with $\boldsymbol{u}$ as root.
Difficult to analyze. We need simple classifiers.

root-finding

One may use root-finding algorithms.

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

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

The probability of error is at most

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

Optimizing in ϵ, we obtain that

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

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

small distance to the root helps

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

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

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

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

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

root-finding with proximity

Theorem. (Moon (2002).)
Let $\boldsymbol{v}^*$ be the centroid of a uniform attachment tree. Then

$$\mathbb{E}\,\text{distance}(\boldsymbol{v}^*, \text{root}) = 1 + \boldsymbol{o_n}(1)\ .$$

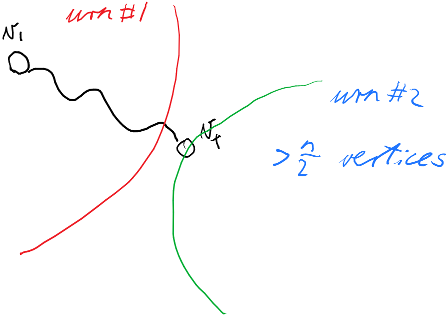
For preferential attachment trees Wagner and Durant (2019) prove analogous results.

centroid rule

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

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

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



root-finding with small diameter

It follows that in the uniform attachment model

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

Moreover,

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

whenever $\boldsymbol{q} \leq 1/2$.

Similar bounds hold for preferential attachment.

when only leaf-colors are observed

In the more difficult model when only the colors of the leaves are observed,

$$\boldsymbol{R}^*(\boldsymbol{q},\boldsymbol{n}) \le 13\boldsymbol{q}$$

and

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

whenever $\boldsymbol{q} < 1/2$.

We only need the fact that there is at least one leaf close to $\boldsymbol{v}^*$. 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,

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

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,

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

beyond trees: random uniform k-dags

A uniform random k-dag is the union of k independent uniform random recursive trees on the same vertex set $[n]$.

A uniform random k-dag may be generated recursively; each vertex $i \in \{2, 3, \ldots, n\}$ is attached by an edge to a k vertices chosen uniformly at random (with replacement) among the vertices $\{1, \ldots, i-1\}$.

Multiple edges are collapsed so that the resulting graph is simple.

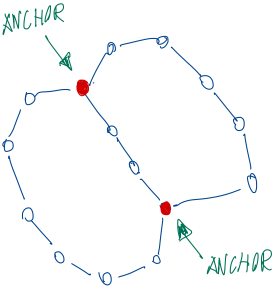
Can one find the root vertex?

Challenge: centrality-based methods the worked for trees do not have an obvious extension.

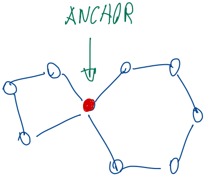
beyond trees: random uniform $\boldsymbol{k}$-dags

S. Briend, F. Calvillo, G. Lugosi.
Archaeology of random recursive dags and Cooper-Frieze random networks.
Combinatorics, Probability, and Computing, 2023.

double cycles and their anchors



"DOUBLE CYCLES"



adam is an anchor of a short double cycle

For a positive integer M, let $S_M \subset [n]$ be the set of vertices i such that i is an anchor of a double cycle of size (K, L) for some $K \leq M$ and $L \leq M$.
For any ϵ, one may take $M = M_\epsilon = \frac{42}{k} \log(1/\epsilon)$ such that

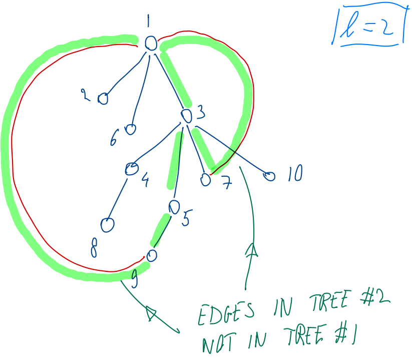
$$\mathbb{P}\left\{1 \in S_M \text{ and } |S_M| \leq \frac{4}{\epsilon} k^{2M}(2M)!\right\} \geq 1 - \epsilon \; .$$

When $\epsilon \leq e^{-ck}$ we may take

$$K(\epsilon) = \frac{1}{\epsilon}\left(c \log \frac{1}{\epsilon}\right)^{c \log(1/\epsilon)/k} \; .$$

This is probably far from the optimal dependence.

adam is an anchor of a short double cycle



cooper-frieze networks

The same method works in "uniform Cooper-Frieze random graphs."

At each step of the growth process, a coin is flipped.

Either a new vertex is added and attached to a random vertex or a random edge is added.

One may take

$$K(\epsilon) = \left(c \log \frac{1}{\epsilon}\right)^{c \log(1/\epsilon)} .$$

broadcasting on k-dags

We may extend the broadcasting problem from trees to k-dags.

- Initially, there are k "roots", each colored red or blue (where k is odd).
- For $n > k$, when vertex n arrives, it chooses k parents uniformly at random (with replacement) and attaches to them.
- The color of each one of the k parents is flipped with probability p.
- The color of the new vertex is assigned according to the majority.

> Upon observing a large random uniform k-dag with its vertex colors, we would like to infer the majority color of the k roots.

majority rule

We studied the simple majority rule: count red and blue vertices and decide according to the majority vote.

This rule disregards the graph structure.

S. Briend, L. Devroye, G. Lugosi.
Broadcasting in random recursive dags, 2023.

a randomized nonlinear urn model

The proportion of red and blue vertices $(R_n, 1 - R_n)$ evolves according to a generalized Pólya urn:

- Initially, there are kR_k red and $k(1 - R_k)$ blue balls.

- For $n > k$, k balls are drawn at random (with replacement).

- The color of each ball is independently flipped with probability p.

- A new ball is added to the urn, whose color is the majority of the flipped colors.

probability of error

Denote by b_n^{maj} the majority color at time n. (In case of ties, randomize.)

The probability of error is

$$R^{maj}(n,p) = \mathbb{P}\left\{ b_n^{maj} \neq b_k^{maj} \right\} \ .$$

For what values of p do we have

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \ ?$$

proportion of red vertices

The first step us to understand the asymptotic behavior of $\boldsymbol{R_n}$.
Note that

$$R_{n+1} = R_n + \frac{\mathbb{1}_{\{c_{n+1}=\text{red}\}} - R_n}{n+1}$$

where

$$\mathbb{P}\left\{c_{n+1} = \text{red} | R_n\right\} \;=\; \mathbb{P}\left\{\text{Bin}\left(k, f(R_n)\right) \geq k/2 | R_n\right\}$$

with $f(t) = (1-p)t + p(1-t)$.
The expected increase is

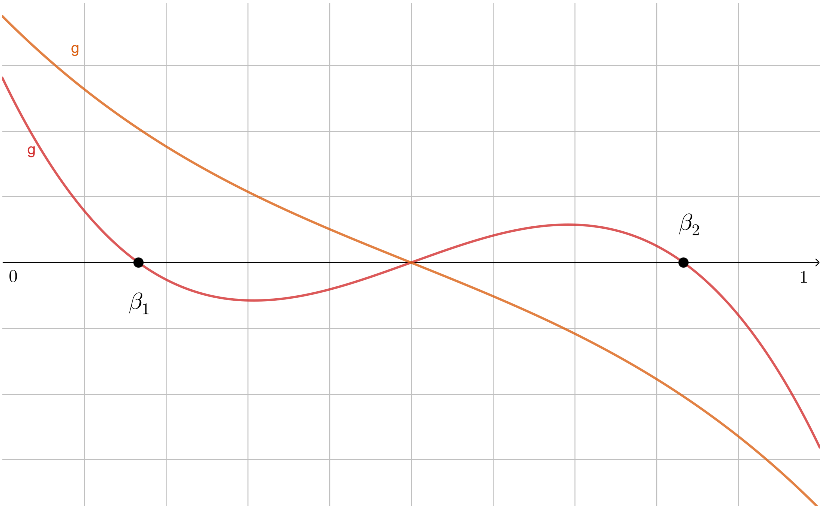
$$g(t) = \mathbb{P}\left\{\text{Bin}(k, f(t)) > k/2\right\} - t\ .$$

The behavior of R_n depends on the nature of zeroes of $g(t)$.

Such reinforced random processes have been extensively studied, see Pemantle (2007).

the function g

The function g for low-, and high-mutation probabilities.



the function g

Define

$$\alpha_k = \frac{1}{2^{k-2}} \sum_{i>k/2}^{k} \binom{k}{i} (i - k/2) \, .$$

Note that $\alpha_1 = 1$, $\alpha_3 = 3/2$, and $\alpha_k \sim \sqrt{k/(2\pi)}$ for large k.

Three regimes:

$$\begin{aligned} 0 \le p < \frac{1}{2} - \frac{1}{2\alpha_k} &: \qquad \text{low rate of mutation;} \\ \frac{1}{2} - \frac{1}{2\alpha_k} \le p < \frac{1}{2} - \frac{1}{2\alpha_k} &: \qquad \text{medium rate of mutation;} \\ \frac{1}{2} - \frac{1}{4\alpha_k} \le p \le \frac{1}{2} &: \qquad \text{low rate of mutation;} \end{aligned}$$

For $k = 3$, the thresholds are at $1/6$ and $1/3$.

low rate of mutation

There exist constants $0 < \beta_1 < 1/2 < \beta_2 < 1$ such that, almost surely, either $R_n \to \beta_1$ or $R_n \to \beta_2$.

Moreover, if $R_k < 1/2$, then

$$\mathbb{P}\left\{R_n \to \beta_1 | R_k\right\} < \mathbb{P}\left\{R_n \to \beta_2 | R_k\right\}$$

and therefore

$$\limsup_{n\to\infty} R^{maj}(n,p) < \frac{1}{2} \ .$$

(This regime does not exist when $k = 1$.)

Note: $\beta_1 \leq e^{-2k(1/2-p)^2}$ and $\beta_2 = 1 - \beta_1$.

medium rate of mutation

R_n converges to $1/2$ almost surely.

However, if $R_k < 1/2$, then with positive probability, $R_n < 1/2$ for all n and therefore

$$\limsup_{n \to \infty} R^{maj}(n, p) < \frac{1}{2} \; .$$

high rate of mutation

R_n converges to $1/2$ almost surely.

Also, with probability one R_n crosses $1/2$ infinitely often and

$$\limsup_{n\to\infty} R^{maj}(n,p) = \frac{1}{2} \; .$$

further questions

- estimates of the asymptotic probability of error;
- methods that take the graph structure into account; what is the best possible probability of error?
- extensions to other models such as preferential attachment or geometric variants.