Analysis of Steiner subtrees of Random Trees for Traceroute Algorithms

Fabrice Guillemin, Philippe Robert

ANALYSIS OF STEINER SUBTREES OF RANDOM TREES FOR TRACEROUTE ALGORITHMS

FABRICE GUILLEMIN AND PHILIPPE ROBERT

Abstract. We consider in this paper the problem of discovering, via a traceroute algorithm, the topology of a network, whose graph is spanned by an infinite branching process. A subset of nodes is selected according to some criterion. As a measure of efficiency of the algorithm, the Steiner distance of the selected nodes, i.e. the size of the spanning sub-tree of these nodes, is investigated. For the selection of nodes, two criteria are considered: A node is randomly selected with a probability, which is either independent of the depth of the node (uniform model) or else in the depth biased model, is exponentially decaying with respect to its depth. The limiting behavior the size of the discovered subtree is investigated for both models.

Contents

1. Introduction

In the past ten years, the Internet has known an extraordinary expansion and still experiences today a sustained growth. The counterpart of this success is that the different autonomous systems composing the global Internet have been independently developed by different operators. This raises some issue since the Internet is by construction a flat network, where the different components are interdependent in terms of connectivity availability, security, quality of service etc. It thus turns out that the knowledge of the physical layout of a network is of prime interest for network operators. The physical topology of a component of the Internet is in general very difficult to describe. To establish a representation of the whole or a part of the Internet, some topology exploration methods have to be devised. Various topology discovery experiments have been initiated by different organizations in order to infer the topology of the global Internet, notably the Skitter project by

Date: June 6, 2008.
Key words and phrases. Traceroute Algorithm. Steiner Distance. Branching Processes. Oscillating Behavior. Asymptotic Expansions.

CAIDA [3], the DIMES project [14] and many other initiatives. The method generally proposed for analyzing the topology of a network is based on the traceroute facility offered by routers. Roughly speaking, a traceroute procedure consists of sending traceroute messages between hosts as follows:

TRACEROUTE ALGORITHM
If H and G are hosts participating in the topology discovery experiment, H sends to G a traceroute message so that all the hosts/routers on the path (H, G) are identified.

The purpose of this paper is to investigate the efficiency of the traceroute algorithm. While a large number of experimental papers are available in the technical literature on the analysis of the topology of the Internet, a very few studies provide analytical insight into the efficiency of these topology discovery methods; see Vespignani *et al.* [5] for a discussion and Azzana *et al.* [2] for an analysis in the case of specific deterministic trees.

In this paper, a more realistic model is proposed to include some randomness in the degree of the nodes of the graph representing the topology of a network. One specifically considers a network with a random tree architecture spanned by a Galton-Watson branching process. We shall restrict the analysis to the case of offspring distributions, which have a finite second momenmt. This notably precludes the case of power law distributions with infinite second moments, typically distributions G such that $\mathbb{P}(G \geq n) \sim Cn^{-\alpha}$ with $\alpha \in (0, 2)$.

The Internet graph is definitely not a tree, since many studies (see the Skitter project) show that there is a core of highly connected routers. Nevertheless, some components of the Internet have a topology close to a tree structure. This is notably the case of access or collect networks, which play the role of capillarity networks in charge of collecting and distributing traffic between customers and the core of the Internet. This latter component is not critical for the problem we study in this paper since core routers are easy to discover by traceroute procedures. This is why we focus on collect networks, which can be represented by a tree architecture, spanned by a branching process. In addition, to get more insight into the topology discovery process in the case of a large network, it is assumed that the underlying branching process does not terminate with probability 1; in particular the depth of the tree is infinite.

The discovery process is as follows: a random number of nodes are selected among the nodes of the tree. After the selected nodes have performed the traceroute algorithm, the set of the nodes discovered is the spanning tree of the selected nodes. The performance criterion used in this paper is simply the size of this sub-tree. In graph theory it is known as the *Steiner distance* of the selected nodes (with the slight difference that the selected nodes are not counted). It has been the subject of a recent interest by Mahmoud and Neininger [9] and Christophi and Mahmoud [4] which considered the asymptotic behavior of the distance between two random nodes of the tree. Panholzer [12], Panholzer and Prodinger [13] proved central limit theorems when multiple points are considered. The asymptotics investigated in these papers concern the size of the random tree. In our paper, we will study two situations: when the size of the tree and the number of selected nodes go to

infinity and also when the infinite tree is fixed and the number of selected nodes grows. See Panholzer [12] for a thorough discussion of the literature in this domain.

Two stochastic models for selecting the nodes in the network are considered. In the first model, the uniform model, we adopt the point of view of an external observer to the network; a set of nodes is chosen at random and a traceroute algorithm is performed. In the second model, the depth biased model, the observer is located at one node (the root node) and it chooses more likely nodes not too far away. As it will be seen, in the uniform model , the selected nodes are basically in the "bottom" of the tree where most of the nodes are, while in the second model they are more concentrated at the "top" of the tree.

In the first model, referred to as the uniform model, nodes whose depth is less than $N > 0$ are randomly chosen with probability $1 - \exp(-\lambda)$ for some $\lambda > 0$ independently of the position of the node in the tree. The quantity analyzed here is the ratio $\rho_N(\lambda)$ of the mean size $\mathbb{E}(R_N)$ of the sub-tree discovered and the mean number $\mathbb{E}(T_N)$ of nodes of the tree whose depth is less than N. The quantity $\rho_N(\lambda)$ denotes the fraction of the tree discovered. The asymptotic results of this paper first determine the limit $\rho(\lambda)$ of $\rho_N(\lambda)$ as N tends to infinity. In a second step, the asymptotic behavior of $\rho(\lambda)$ for $\lambda \to 0$ is investigated. This last point gives an indication of the efficiency of the algorithm when only a few nodes are selected in the topology discovery experiment.

For the uniform model, it is shown in Theorem 1 that, for small λ, the exploration rate $\rho(\lambda)/\lambda$ is equivalent to $\log_m \lambda$ where m is the mean value of the offspring distribution of a node, so that at the first order the algorithm is very efficient. A second order analysis, Proposition 2, reveals that the standard deviation of the size of the discovered tree scales with the mean size of the tree, except when the offspring distribution is deterministic. This latter case is degenerate in the sense that the standard deviation is negligible when compared to the mean value.

In the second model, referred to as the depth biased model, the probability of selecting a node depends on its depth in the tree so that the mean number of selected nodes at depth n is α^n for some $\alpha > 0$. It is shown in Theorem 2 that the ratio of the average of the size $R(\alpha)$ of the sub-tree discovered and the average number of selected nodes is equivalent to $1/(1-\alpha)$.

The paper is organized as follows: In Section 2, the models for the selection of the nodes of the tree are introduced. The uniform model is investigated in Section 4 and the depth biased model in Section 5. The main ingredients for the analysis of these models are Kesten-Stigum Theorem and some results on the rates of convergence for Galton-Watson branching processes and a general limit theorem proved in Section 3.

Acknowledgments. The authors wish to thank two anonymous referees for their work, their detailed comments have helped us a lot to improve and correct mistakes in the first version of the paper.

2. Problem Formulation

Throughout this paper, we consider a Galton-Watson branching process, whose graph is a tree denoted by $\mathcal{T}$. Each element of the nth generation (or nth level) gives birth to G nodes at the $(n+1)$th generation independently of the other elements of the nth level, where the offspring G is some *integrable* random variable. (See Athreya and Ney [1] and Lyons and Peres [7] for an introduction to random trees.)

It is assumed that $\mathbb{P}(G{=}0){=}0$ and $P(G \geq 2) > 0$, in particular the tree is supercritical, i.e. $m = \mathbb{E}(G) > 1$. For $n \geq 0$, the variable Z_n denotes the number of nodes at level n, in particular $Z_0 = 1$. For $1 \leq \ell \leq Z_n$, a node of the tree can be represented as a pair (n, ℓ), where n is its generation and ℓ its rank within the generation. (For notational conventions, see Neveu [11] for example.) Let $\mathcal{T}_k^{n,\ell}$ denote the sub-tree of $\mathcal{T}$ with depth less than or equal to k and with root at node (n, ℓ). The size of $\mathcal{T}_k^{n,\ell}$ is denoted by $T_k^{n,\ell}$. When (n, ℓ) is the root node, i.e. $(n, \ell) = (0, 1)$, the upper index $(0, 1)$ is omitted. With the above notation, one gets easily that for all $N > 1$ and $n = 1, \ldots, N$

$$T_N = \sum_{i=0}^{n-1} Z_i + \sum_{\ell=1}^{Z_n} T_{N-n}^{n,\ell}. \tag{1}$$

Let us consider a counting measure $\mathcal{N}$ on the tree representing the distribution of the points selected in the tree: For a subset A of the nodes of the tree, $\mathcal{N}(A)$ denotes the total number of points in A. By selecting nodes, a sub-tree from $\mathcal{T}$ is obtained through the traceroute algorithm; this sub-tree is referred to as sampled tree. See Figure 1.

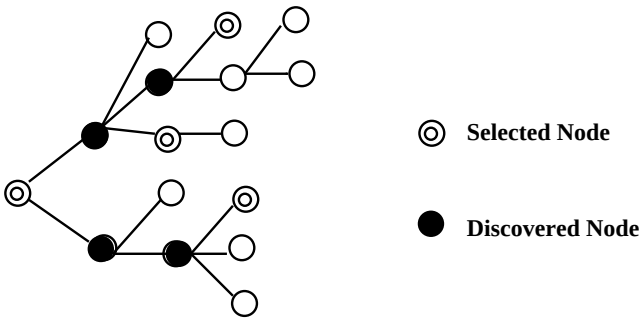

FIGURE 1. Traceroute Algorithm.

To complete the description of the problem, it remains to specify how the nodes of the original tree are selected. In the following, we shall consider two selection criteria:

Uniform model: Nodes are chosen at random on all the nodes of the tree whose depth is less than or equal to N, N being a fixed integer. A node is selected with probability $1 - \exp(-\lambda)$ independently of his depth in the tree. The mean number of nodes involved in the discovery experiment is then $(1 - \exp(-\lambda))(m^{N+1} - 1)/(m - 1)$. (Recall that the mean size of the nth generation is m^n, $n \geq 0$, where $m = \mathbb{E}(G)$, the mean of the offspring variable G.)

To investigate the topology discovery process, we shall consider for a fixed $N > 0$ the N first levels of the original tree $\mathcal{T}$ and count the number of nodes which are discovered, given by

$$R_N = \sum_{n=0}^{N} \sum_{\ell=1}^{Z_{N-n}} \mathbb{1}_{\{\mathcal{N}(\mathcal{T}_n^{N-n,\ell}) \neq 0\}}. \tag{2}$$

In the following, we shall be particularly interested in the quantity

$$\rho_N(\lambda) = \frac{\mathbb{E}(R_N)}{\mathbb{E}(T_N)}, \tag{3}$$

i.e., the ratio of the mean number of discovered nodes to the mean number of nodes in the tree, when the analysis is restricted to the N first levels of the tree. Then the behavior of this ratio when the number N of levels tends to infinity is investigated.

Depth biased model: Nodes at given level n are selected with probability $1-\exp[-(\alpha/m)^n]$ for some $\alpha \in (0,1)$. The mean number of nodes selected at level n is $m^n(1-\exp[-(\alpha/m)^n]) \sim \alpha^n$ and therefore is exponentially decreasing with respect to the depth. The rational behind that is the fact that, for this model, the traceroute procedure will rarely select nodes "far away" from the root node, in contrary to the uniform case where geometric aspects are completely ignored for the selections of the hosts.

By denoting by $R(\alpha)$ the total number of nodes discovered, the efficiency of the traceroute algorithm is measured in this case through the ratio of the mean $\mathbb{E}(R(\alpha))$ to the average number of selected nodes. The limiting behavior when the average number of selected nodes becomes large, i.e. when $\alpha \nearrow 1$, is investigated.

Additionally it is assumed that the root node of the tree is always selected; it is not difficult to show that for both models described above, the root node belongs to the sample tree with a very high probability and then the above assumption is not really restrictive. This implies that a node (n,ℓ) of the tree $\mathcal{T}$ at level n belongs to the sampled tree whenever $\mathcal{N}(\mathcal{T}_{N-n}^{n,\ell})$ is not 0. In other words, a node of the original tree belongs to the discovered tree if at least one of his descendants has been selected. In the following, we shall use the following notation: for a subtree $\mathcal{T}_n^{N-n,\ell}$ rooted at a vertex $(N-n,\ell)$ of the $(N-n)$th generation of the tree $\mathcal{T}$ and with depth n, the quantity $\mathbb{P}(\mathcal{N}(\mathcal{T}_n^{N-n,\ell}) \neq 0)$ is the probability that a least one vertex of the subtree $\mathcal{T}_n^{N-n,\ell}$ is marked and $(N-n,\ell) \in \mathcal{T}$.

Before proceeding to the analysis of the topology discovery process, we prove in the next section a technical result, which is important in the analysis of the speed of the exploration process.

3. A Convergence Result

To prove asymptotic expansions in the following sections, the following proposition will repeatedly be used. Its proof is based on integral representations and Fubini's Theorem instead of complex analysis techniques as it is usually the case in the context of harmonic series. See Robert [15] for a presentation of these methods.

Proposition 1. *Let V be a positive random variable with $\mathbb{E}(V^2) < +\infty$ and h be a non-negative twice differentiable function on $\mathbb{R}_+$ such that $h(0) = 0$. In addition, it is assumed that the function h' is integrable with $h'(0) \neq 0$ and that there exists some constant $K > 0$ such that $|h''(x)| < K$ for all $x \in [0,\infty)$.*

The function $\Psi(h)(x)$ defined by

$$\Psi(h)(x) = \sum_{n=0}^{+\infty} \frac{1}{m^n} \mathbb{E}\left(h\left(xVm^n\right)\right), \quad x \geq 0, \tag{4}$$

is such that

$$\lim_{x\to 0}\frac{\Psi(h)(x)}{x\log_m(1/x)}=\mathbb{E}(V)h'(0).$$

Proof. Since h is non-negative and $|h'|$ integrable with respect to Lebesgue measure on $\mathbb{R}_+$, Fubini's Theorem applied twice shows that $\Psi(h)$ can be expressed as

$$\begin{aligned}\Psi(h)(x)&=\sum_{n=0}^{+\infty}\frac{1}{m^n}\mathbb{E}\left(h\left(xVm^n\right)\right)=\mathbb{E}\left(\sum_{n=0}^{+\infty}\frac{1}{m^n}h\left(xVm^n\right)\right)\\&=\mathbb{E}\left(\sum_{n=0}^{+\infty}\frac{1}{m^n}\int_0^{+\infty}h'(u)\mathbb{1}_{\{u\le xVm^n\}}\,du\right)\\&=\mathbb{E}\left(\int_0^{+\infty}h'(u)\sum_{n=0}^{+\infty}\frac{1}{m^n}\mathbb{1}_{\{u\le xVm^n\}}\,du\right).\end{aligned}\tag{5}$$

The function $\Psi(h)$ is thus well defined.

Since $h'(0)>0$, Fatou's Lemma applied successively gives the relation

$$\begin{aligned}\liminf_{x\to 0}\frac{\Psi(h)(x)}{x}&\ge\sum_{n=0}^{+\infty}\liminf_{x\to 0}\frac{m-1}{m^n}\mathbb{E}\left(\frac{h\left(xVm^n\right)}{x}\right)\\&\ge\sum_{n=0}^{+\infty}\frac{m-1}{m^n}\mathbb{E}\left(\liminf_{x\to 0}\frac{h\left(xVm^n\right)}{x}\right)=\sum_{n=0}^{+\infty}(m-1)\mathbb{E}\left(V\right)h'(0)=+\infty,\end{aligned}$$

therefore the ratio $\Psi(h)(x)/x$ diverges as $x\to 0$.

By using representation (5) of $\Psi(h)$, we have

$$\begin{aligned}(m-1)\Psi(h)(x)&=m\mathbb{E}\left(\int_0^{xV}h'(u)\,du\right)\\&\quad+\mathbb{E}\left(\int_{xV}^{V}\frac{1}{m^{\lfloor\log_m(u/xV)\rfloor}}h'(u)\,du\right)+\mathbb{E}\left(\int_{V}^{+\infty}\frac{1}{m^{\lfloor\log_m(u/xV)\rfloor}}h'(u)\,du\right),\end{aligned}$$

where $\lfloor y\rfloor$ is the integer part of $y\in\mathbb{R}$. One first shows that only the central term of the right hand side plays a role in the asymptotic behavior of $\Psi(h)$ at the first order.

For the first term, note that, if $\|h''\|_\infty$ is the L_∞ norm of h'',

$$\begin{aligned}&\left|\frac{1}{x}\mathbb{E}\left(\int_0^{xV}h'(u)\,du\right)\right|\\&\qquad\le\frac{1}{x}\mathbb{E}\left(\int_0^{xV}\left(h'(0)+u\|h''\|_\infty\right)\,du\right)\le h'(0)\mathbb{E}(V)+\frac{x}{2}\mathbb{E}(V^2)\|h''\|_\infty\end{aligned}$$

For $u\ge V$, one has

$$xm^{\lfloor\log_m(u/xV)\rfloor}\ge xm^{\lfloor\log_m(1/x)\rfloor}\ge xm^{\log_m(1/x)-1}=m^{-1},$$

and hence,

$$\frac{1}{x}\mathbb{E}\left(\int_V^{+\infty}\frac{1}{m^{\lfloor\log_m(u/xV)\rfloor}}|h'(u)|\,du\right)\le m\int_0^{+\infty}|h'(u)|\,du.$$

By gathering these estimations, it follows that the following equivalence

$$\frac{\Psi(h)(x)}{x} \sim \mathbb{E}\left(\int_{xV}^{V} \frac{1}{xm^{\lfloor \log_m(u/xV)\rfloor}} h'(u)\,du\right) = \mathbb{E}\left(V\int_{xV}^{V} m^{\{\log_m(u/xV)\}}\frac{h'(u)}{u}\,du\right),$$

holds as $x \to 0$, with $\{y\} = y - \lfloor y \rfloor$, the fractional value of $y \in \mathbb{R}$. The above equivalence can be rewritten as

$$\frac{\Psi(h)(x)}{x} \sim \mathbb{E}\left(V\int_{xV}^{V} m^{\{\log_m(u/xV)\}}\frac{h'(u)-h'(0)}{u}\,du\right) + h'(0)\mathbb{E}\left(V\int_{xV}^{V}\frac{m^{\{\log_m(u/xV)\}}}{u}\,du\right)$$

Due to the boundedness of h'' and the integrability of V^2, the first term in the right hand side of the above equation is bounded as x goes to 0. Hence, only the second term has to be considered. For $x < 1$, we have

$$\int_{xV}^{V}\frac{m^{\{\log_m(u/xV)\}}}{u}\,du = \int_{1}^{1/x}\frac{m^{\{\log_m(u)\}}}{u}\,du$$
$$= \sum_{k\geq 0: m^k \leq 1/x}\int_{m^k}^{m^{k+1}}\frac{m^{\{\log_m(u)\}}}{u}\,du + O(1) = (m-1)\lfloor -\log_m(x)\rfloor + O(1)$$

and the result follows. □

Asymptotic behavior of algorithms with an underlying tree structure has been extensively investigated, see Flajolet *et al.* [6], Mohamed and Robert [10] and Mahmoud [8] for a general presentation. By using the terminology of Flajolet *et al.* [6], for non-negative sequences (λ_n) and (μ_n), a series like

$$G(x) = \sum_{n\geq 0}\lambda_n g(\mu_n x), \tag{6}$$

for some function g is defined as an *harmonic sum*. Because of the integration of the random variable V and given that one wants the weakest assumptions on this random variable, series (4) could be seen as a special case of harmonic sums. The fact that the sequences (λ_n) and (μ_n) are specific in Expression (4) is not a real restriction, see Robert [15].

Flajolet *et al.* [6] derives the asymptotic expansion of $G(x)$ when x goes to 0 or $+\infty$ by using Mellin transform techniques. For $s \in \mathbb{C}$, if $h^*(s)$ is the Mellin transform of h, i.e. for s in some vertical strip of $\mathbb{C}$,

$$h^*(s) = \int_0^{+\infty} h(x)\,x^{s-1}\,dx,$$

it is easy to check that the Mellin transform of $\Psi(h)$ is given by

$$\Psi(h)^*(s) = \frac{1}{1-m^{-(s+1)}}\mathbb{E}\left(V^{-s}\right)h^*(s).$$

Following the methods of Flajolet *et al.* [6], to derive the asymptotic behavior of $\Psi(h)(x)$ as x goes to infinity, one has to identify the first singularity of $\Psi(h)^*$ on

the right of the maximal vertical strip where it is defined. In particular, some conditions on the finiteness of some fractional moments of the random variable V have to be assumed (as well as growth conditions on h^*). From this point of view, our approach is minimal since only the finiteness of $\mathbb{E}(V^2)$ and differentiability conditions on h are assumed. It turns out that it is important as it will be seen in the following sections, since in practice little is known on the fractional moments of the corresponding variable V.

4. The Exploration Rate in the Uniform Model

In this section, nodes are selected at random with uniform probability in the tree with depth less than N. The variable R_N is the size of the underlying subtree (or sampled tree) containing the selected nodes. The asymptotic behavior of $\rho_N(\lambda) = \mathbb{E}(R_N)/\mathbb{E}(T_N)$, the fraction of discovered nodes, when N tends to infinity is investigated. In the second part of this section, the ratio $\mathrm{var}(R_N)/\mathbb{E}(T_N)$ is analyzed.

4.1. First Order Asymptotics.

In the uniform case, the limiting behavior of the ratio $\rho_N(\lambda)$ when N tends to infinity is given by the following result.

Theorem 1. *The ratio of the average size R_N of the sampled tree to the total average size of the tree $\mathbb{E}(T_N)$ satisfies the relation*

$$\rho(\lambda) \stackrel{def.}{=} \lim_{N\to+\infty} \rho_N(\lambda) = \sum_{n=0}^{+\infty} \frac{m-1}{m^{n+1}} \left(1 - \mathbb{E}\left(\exp\left(-\lambda \sum_{i=0}^{n} Z_i\right)\right)\right). \tag{7}$$

If additionally the condition $\mathbb{E}\left(G^2\right) < +\infty$ holds then

$$\lim_{\lambda\to 0} \frac{\rho(\lambda)}{\lambda \log_m(1/\lambda)} = 1. \tag{8}$$

Relation (8) shows that the rate of increase of the discovery process is infinite near the origin. This implies that with only a few selected nodes one has the impression of rapidly discovering the whole network.

Proof. By conditioning on the tree, the conditional probability that node $(N-n,\ell)$ does not belong to the sampled tree is

$$\mathbb{P}\left(\mathcal{N}\left(\mathcal{T}_n^{N-n,\ell}\right) \neq 0 \middle| \mathcal{T}\right) = 1 - \exp\left(-\lambda T_n^{N-n,\ell}\right).$$

By summing-up these relations, one obtains that the expected value of R_N, i.e., the average number of nodes in the sampled tree, is given by

$$\begin{aligned}\mathbb{E}(R_N) &= \sum_{n=0}^{N} \mathbb{E}(Z_{N-n})\left(1 - \mathbb{E}\left(\exp\left(-\lambda T_n\right)\right)\right) \\ &= \sum_{n=0}^{N} m^{N-n}\left(1 - \mathbb{E}\left(\exp\left(-\lambda \sum_{i=0}^{n} Z_i\right)\right)\right).\end{aligned}$$

The limit when $N \to \infty$ of the ratio $\rho_N(\lambda)$ is then given by

$$\rho(\lambda) \stackrel{\text{def.}}{=} \lim_{N\to+\infty} \rho_N(\lambda) = \sum_{n=0}^{+\infty} \frac{m-1}{m^{n+1}} \left(1 - \mathbb{E}\left(\exp\left(-\lambda \sum_{i=0}^{n} Z_i\right)\right)\right)$$

since $\mathbb{E}(T_N) \sim m^{N+1}/(m-1)$ for large N. This proves the first equality stated in Theorem 1.

We now study the behavior of $\rho(\lambda)$ when λ goes to 0. Since $\mathbb{E}(G^2) < +\infty$, Kesten-Stigum's Theorem ensures the existence of a random variable W such that $\mathbb{P}(W > 0) = 1$ (because of the assumption on the distribution of G) and $\mathbb{E}(W) = 1$ (See Lyons and Peres [7]) and that, almost surely,

$$\lim_{n\to+\infty} \frac{Z_n}{m^n} = W. \tag{9}$$

Let us define

$$f(\lambda) \stackrel{\text{def.}}{=} \sum_{n=0}^{+\infty} \frac{m-1}{m^{n+1}} \left(1 - \mathbb{E}\left(\exp\left(-\lambda W \frac{m^{n+1}-1}{m-1}\right)\right)\right).$$

Then,

$$\begin{aligned} &\frac{|\rho(\lambda) - f(\lambda)|}{\lambda} \\ &\quad \leq \sum_{n=0}^{+\infty} \frac{m-1}{\lambda m^{n+1}} \left| \mathbb{E}\left(\exp\left(-\lambda \sum_{i=0}^{n} Z_i\right)\right) - \mathbb{E}\left(\exp\left(-\lambda W \frac{m^{n+1}-1}{m-1}\right)\right)\right|. \end{aligned} \tag{10}$$

Since W is integrable, Lebesgue's dominated convergence Theorem gives that

$$\begin{aligned} &\lim_{\lambda\to 0} \frac{1}{\lambda} \mathbb{E}\left(\exp\left(-\lambda \sum_{i=0}^{n} Z_i\right) - \exp\left(-\lambda W \frac{m^{n+1}-1}{m-1}\right)\right) \\ &\qquad = \mathbb{E}\left(\sum_{i=0}^{n} Z_i - W \frac{m^{n+1}-1}{m-1}\right) = 0. \end{aligned} \tag{11}$$

We have

$$\begin{aligned} &\frac{1}{m^{n+1}\lambda} \left| \mathbb{E}\left(\exp\left(-\lambda \sum_{i=0}^{n} Z_i\right)\right) - \mathbb{E}\left(\exp\left(-\lambda W \frac{m^{n+1}-1}{m-1}\right)\right)\right| \\ &\qquad \leq \frac{1}{m^{n+1}} \sum_{i=0}^{n} \mathbb{E}|Z_i - m^i W|. \end{aligned}$$

From Athreya and Ney [1, Theorem 1, page 54], for $n \geq 1$, there exists a sequence (W^i) of i.i.d. random variables with the same distribution as W such that

$$Z_n - m^n W = \sum_{i=1}^{Z_n} (1 - W^i). \tag{12}$$

By using Cauchy-Shwartz's Inequality, we obtain

$$\begin{aligned} \mathbb{E}\left(|Z_n - m^n W|\right)| &\leq \sqrt{\mathbb{E}\left((Z_n - m^n W)^2\right)} \\ &= \mathrm{Var}(1-W)\sqrt{\mathbb{E}(Z_n)} \\ &= \mathrm{Var}(1-W) m^{n/2}. \end{aligned} \tag{13}$$

From the above inequality, we deduce that

$$\frac{1}{m^{n+1}\lambda}\left|\mathbb{E}\left(\exp\left(-\lambda\sum_{i=0}^{n}Z_i\right)\right)-\mathbb{E}\left(\exp\left(-\lambda W\frac{m^{n+1}-1}{m-1}\right)\right)\right| \leq \frac{\mathrm{Var}(1-W)}{\sqrt{m}-1}\frac{1}{m^{(n+1)/2}}.$$

Relation (11) and Lebesgue's Theorem then imply that

$$\lim_{\lambda\to 0}\frac{\rho(\lambda)-f(\lambda)}{\lambda}=0.$$

Hence, up to an expression which is of the order of $o(\lambda)$, the behavior at 0 of $\rho(\lambda)$ is equivalent to the behavior of $f(\lambda)$ as λ becomes small.

By using Proposition 1, we have by taking $h(u)=1-e^{-u}$ and $V=Wm/(m-1)$,

$$\sum_{n=0}^{+\infty}\frac{m-1}{m^{n+1}}\left(1-\mathbb{E}\left(\exp\left(-xVm^n\right)\right)\right)=\Psi(h)(x)\sim x\log_m(1/x)$$

as $x\to 0$. To conclude the proof, we note that

$$\lim_{x\to 0}\frac{\Psi(h)(x)-f(x)}{x\log_m x}=0$$

and the result follows. □

4.2. **Second Order Properties.** The results obtained in the previous section show that the size of the sampled tree is of the same order of magnitude as the original tree when the probability of selecting a node is fixed. When this probability is very small (i.e., for small λ), the speed of the discovery process is even very fast. In this section, we evaluate the second moment of the random variable R_N in order to estimate the dispersion of the size of the sampled tree around the mean value.

In the rest of this section, we use the following notation: If (n,ℓ) and (n',ℓ') are two nodes of the tree, the relation $(n',\ell')<(n,\ell)$ indicates that the nodes are distinct and that (n',ℓ') is a node of the sub-tree whose root is (n,ℓ).

Proposition 2 (Asymptotic behavior of the variance)**.** *When the size N of the original tree goes to infinity, the variance of the size of the sampled tree is such that*

(1) *If the random variable G is not deterministic and $\mathbb{E}(G^2)<+\infty$, then*

$$\rho_2^{(1)}(\lambda)\stackrel{\mathrm{def}}{=}\lim_{N\to\infty}\frac{\mathrm{Var}(R_N)}{\mathbb{E}(T_N)^2}=\frac{\mathrm{Var}(G)}{m^2-m}\rho(\lambda)^2 \tag{14}$$

where $\rho(\lambda)$ is defined by Equation (7)*.*

(2) *If $G\equiv m$ almost surely, then*

$$\rho_2^{(2)}(\lambda)\stackrel{\mathrm{def}}{=}\lim_{N\to\infty}\frac{\mathrm{Var}(R_N)}{\mathbb{E}(T_N)} = \frac{m-1}{m}\sum_{n=1}^{+\infty}\frac{1}{m^n}\left[\mathbb{E}\left(e^{-\lambda T_n}\right)\left(1-\mathbb{E}\left(e^{-\lambda T_n}\right)\right)\right. \left.+2\sum_{k=0}^{n-1}\mathbb{E}\left(e^{-\lambda T_{n-k-1}}Z_{n-k}\mathbb{E}\left(e^{-\lambda T_k}\right)^{Z_{n-k}-1}\mathbb{E}\left(e^{-\lambda T_k}\left(1-e^{-\lambda T_k}\right)\right)\right)\right]. \tag{15}$$

where $T_n=(m^{n+1}-1)/(m-1)$.

It is worth noting that the case of a deterministic offspring distribution is degenerate in the sense that the standard deviation of the size of the tree discovered by means of the traceroute algorithm does not scale with the size of the tree (and the discovered tree). The coefficient of variation of the random variables R_N tends to 0 when N goes to infinity.

Proof. Using Representation (2) for the size of the sampled tree, one obtains by conditioning on the tree the relation

$$R_N - \mathbb{E}(R_N) = A_{N,1} + A_{N,2} + A_{N,3},$$

where

$$\begin{aligned}
A_{N,1} &= \sum_{n=0}^{N} \sum_{\ell=1}^{Z_{N-n}} \Delta_n^\ell, \\
A_{N,2} &= \sum_{n=0}^{N} (Z_{N-n} - \mathbb{E}(Z_{N-n}))\mathbb{P}\left(\mathcal{N}(\mathcal{T}_n) \neq 0\right), \\
A_{N,3} &= \sum_{n=0}^{N} \sum_{\ell=1}^{Z_{N-n}} \left(1 - \exp(-\lambda T_n^{N-n,\ell}) - \mathbb{P}(\mathcal{N}(\mathcal{T}_n) \neq 0)\right)
\end{aligned}$$

with $\Delta_n^\ell = \mathbb{1}_{\{\mathcal{N}(\mathcal{T}_n^{N-n,\ell})\neq 0\}} - (1 - \exp(-\lambda T_n^{N-n,\ell}))$. Note that if distinct nodes (n,l) and (n',l') cannot be compared with the relation $''<''$ then, conditionally on the tree $\mathcal{T}$, the corresponding random variables Δ_n^ℓ and $\Delta_{n'}^{\ell'}$ are *centered* and *independent*. In addition, note that $A_{N,1} = R_N - \mathbb{E}(R_N \mid \mathcal{T})$.

To study the variance of the random variable R_N, we separately consider the second moments of the terms $A_{N,1}$, $A_{N,2}$ and $A_{N,3}$. Of course, the terms $A_{N,2}$ and $A_{N,3}$ are non null if and only if the variable G is not deterministic.

The second moment of $A_{N,1}$. It is shown that the second moment of $A_{N,1}$ is of the order of m^N. By using the independence in the selection of nodes in the tree and the fact that the random variables Δ_n^ℓ are centered conditionally on $\mathcal{T}$, we have the identity

$$\mathbb{E}(A_{N,1}^2 \mid \mathcal{T}) = \sum_{(n,\ell)\in\mathcal{T}_N} \mathrm{Var}(\Delta_n^\ell \mid \mathcal{T}) + 2 \sum_{\substack{(n,\ell),(n',\ell')\in\mathcal{T}_N \\ (n',\ell')<(n,\ell)}} \mathbb{E}\left(\Delta_n^\ell \Delta_{n'}^{\ell'} \mid \mathcal{T}\right).$$

Conditioning on the state of the tree, when $(n',\ell') < (n,\ell)$, one has the identity

$$\mathbb{E}\left(\Delta_n^\ell \Delta_{n'}^{\ell'} \mid \mathcal{T}\right) = \exp\left(-\lambda T_n^{N-n,\ell}\right)\left(1 - \exp\left(-\lambda T_{n'}^{N-n',\ell'}\right)\right).$$

By symmetry, the above computations yield the following relation for the second moment $\mathbb{E}(A_{N,1}^2)$

$$\begin{aligned}
U_N &\overset{\text{def.}}{=} \mathbb{E}(A_{N,1}^2) - \sum_{n=0}^{N} \mathbb{E}(Z_{N-n})\mathrm{Var}(\Delta_n^1) \\
&= 2\mathbb{E}\left(\sum_{n=0}^{N} \mathbb{E}(Z_{N-n}) \exp\left(-\lambda T_n^{N-n,1}\right) \sum_{\substack{(N-n',l')\in\mathcal{T}_N \\ (N-n',l')<(N-n,1)}} \left(1 - \exp\left(-\lambda T_{n'}^{N-n',\ell'}\right)\right)\right),
\end{aligned}$$

where $\mathrm{Var}(\Delta_n)$ is the variance of the random variable $\mathbb{1}_{\{\mathcal{N}(\mathcal{T}_n)\neq 0\}} - \mathbb{P}(\mathcal{N}(\mathcal{T}_n) \neq 0)$.

For two nodes of the tree such that $(N-n', l') < (N-n, 1)$, Equation (1) gives the relation

$$T_n \stackrel{\text{dist.}}{=} \sum_{k=0}^{n-n'-1} \widetilde{Z}_k + \sum_{\ell'=1}^{\widetilde{Z}_{n-n'}} T_{n'}^{N-n',\ell'},$$

where $(\widetilde{Z}_k, k \geq 0)$ denotes another independent Galton-Watson process independent of $(Z_n, n \geq 0)$ with the same offspring distribution. By using this relation, we have

$$\begin{aligned}
&\mathbb{E}\left(\exp\left(-\lambda T_n^{N-n,1}\right) \sum_{\substack{(N-n',l')\in\mathcal{T}_N \\ (N-n',l')<(N-n,1)}} \left(1-\exp\left(-\lambda T_{n'}^{N-n',\ell'}\right)\right)\right) \\
&\qquad = \mathbb{E}\left(\sum_{n'=0}^{n-1} \sum_{\ell'=1}^{Z_{n-n'}} \exp\left(-\lambda T_n^{N-n,\ell}\right)\left(1-\exp\left(-\lambda T_{n'}^{N-n',\ell'}\right)\right)\right) \\
&\qquad\qquad = \mathbb{E}\left(\sum_{n'=0}^{n-1} \exp\left(-\lambda \sum_{k=0}^{n-n'-1} Z_k\right) \mathbb{E}(V_{n-n'})\right),
\end{aligned}$$

where

$$V_{n-n'} = \sum_{\ell'=1}^{Z_{n-n'}} \exp\left(-\lambda \sum_{\ell''=1}^{Z_{n-n'}} T_{n'}^{N-n',\ell''}\right)\left(1-\exp\left(-\lambda T_{n'}^{N-n',\ell'}\right)\right).$$

By using the independence of the different trees $\mathcal{T}_{n'}^{N-n',\ell'}$ for $\ell = 1, \ldots, Z_{n-n'}$, we have

$$\begin{aligned}
&\mathbb{E}(V_{n-n'} \mid Z_0, \ldots, Z_{n-n'-1}) \\
&\quad = \mathbb{E}\left(Z_{n-n'}\left(\mathbb{E}\left(e^{-\lambda T_{n'}}\right)\right)^{Z_{n-n'}-1} \mid Z_0, \ldots, Z_{n-n'-1}\right) \mathbb{E}\left(e^{-\lambda T_{n'}}\left(1-e^{-\lambda T_{n'}}\right)\right).
\end{aligned}$$

It follows that by using the above expression for U_N, one obtains

$$\begin{aligned}
U_N = 2\sum_{n=0}^{N} \mathbb{E}(Z_{N-n}) \sum_{n'=0}^{n-1} \mathbb{E}\Bigg(&\exp\left(-\lambda \sum_{k=0}^{n-n'-1} Z_k\right) \\
&Z_{n-n'} \mathbb{E}\left(\exp\left(-\lambda T_{n'}\right)\right)^{Z_{n-n'}-1} \mathbb{E}\left(\exp\left(-\lambda T_{n'}\right)\left(1-\exp\left(-\lambda T_{n'}\right)\right)\right)\Bigg).
\end{aligned}$$

Dividing by $\mathbb{E}(T_N)$, we have

$$\begin{aligned}
\frac{U_N}{\mathbb{E}(T_N)} = \frac{2(m-1)}{m-1/m^n} \sum_{n=0}^{N} \frac{1}{m^n} \sum_{n'=0}^{n-1} \mathbb{E}\Bigg(&\exp\left(-\lambda \sum_{k=0}^{n-n'-1} Z_k\right) \\
&Z_{n-n'} \mathbb{E}\left(\exp\left(-\lambda T_{n'}\right)\right)^{Z_{n-n'}-1} \mathbb{E}\left(\exp\left(-\lambda T_{n'}\right)\left(1-\exp\left(-\lambda T_{n'}\right)\right)\right)\Bigg).
\end{aligned}$$

By letting N go to infinity, we finally obtain the relation for the second moment of the random variable $A_{N,1}$

$$\lim_{N\to\infty}\frac{\mathbb{E}(A_{N,1}^2)}{\mathbb{E}(T_N)}=\frac{m-1}{m}\sum_{n=1}^{+\infty}\frac{1}{m^n}\left[\mathbb{E}\left(e^{-\lambda T_n}\right)\left(1-\mathbb{E}\left(e^{-\lambda T_n}\right)\right)\right.$$
$$\left.+2\sum_{k=0}^{n-1}\mathbb{E}\left(e^{-\lambda T_{n-k-1}}Z_{n-k}\mathbb{E}\left(e^{-\lambda T_k}\right)^{Z_{n-k}-1}\mathbb{E}\left(e^{-\lambda T_k}\left(1-e^{-\lambda T_k}\right)\right)\right)\right]. \tag{16}$$

The second moment of $A_{N,2}$. We have

$$\frac{A_{N,2}}{m^N}=\sum_{n=0}^{N}\frac{(Z_{N-n}-\mathbb{E}(Z_{N-n}))}{m^{N-n}}\frac{(1-\mathbb{E}(e^{-\lambda T_n}))}{m^n}$$
$$=\sum_{n=0}^{N}\left(\frac{Z_{N-n}}{m^{N-n}}-1\right)\frac{(1-\mathbb{E}(e^{-\lambda T_n}))}{m^n}.$$

If $\|H\|_2=\sqrt{\mathbb{E}(H^2)}$ for some random variable H, then we have

$$\left\|\frac{A_{N,2}}{m^N}-(W-1)\sum_{n=0}^{N}\frac{(1-\mathbb{E}(e^{-\lambda T_n}))}{m^n}\right\|_2\leq\sum_{n=0}^{N}\left\|W-\frac{Z_{N-n}}{m^{N-n}}\right\|_2\frac{(1-\mathbb{E}(e^{-\lambda T_n}))}{m^n},$$

where W is defined by Equation (9). Athreya and Ney [1, Theorem 2, page 9] gives that the sequence $(\|W-Z_n/m^n\|_2)$ converges to 0. This implies

$$\lim_{N\to+\infty}\frac{\mathbb{E}(A_{N,2}^2)}{\mathbb{E}(T_N)^2}=\frac{(m-1)\mathrm{Var}(G)}{m^3}\left(\sum_{n=0}^{+\infty}\frac{(1-\mathbb{E}(e^{-\lambda T_n}))}{m^n}\right)^2. \tag{17}$$

since $\mathbb{E}\left((1-W)^2\right)=\mathrm{Var}(G)/m(m-1)$.

4.3. **Second moment of $A_{N,3}$.** Clearly

$$\|A_{N,3}\|_2\leq\sum_{n=0}^{N}\left\|\sum_{\ell=1}^{Z_{N-n}}\exp(-\lambda T_n^{N-n,\ell})-\mathbb{E}(\exp(-\lambda T_n))\right\|_2,$$

and since conditionally on Z_{N-n}, the random variables $\exp(-\lambda T_n^{N-n,\ell})$ for $\ell=1,\ldots,Z_{N-n}$ are independent and identically distributed with mean $\mathbb{E}(\exp(-\lambda T_n))$, we then have

$$\left\|\sum_{\ell=1}^{Z_{N-n}}\exp(-\lambda T_n^{N-n,\ell})-\mathbb{E}(\exp(-\lambda T_n))\right\|_2^2=\mathbb{E}(Z_{N-n})\mathrm{Var}(\exp(-\lambda T_n))$$
$$\leq m^{(N-n)}.$$

It follows that

$$\|A_{N,3}\|_2\leq\frac{m^{(N+1)/2}}{\sqrt{m}-1}.$$

and then

$$\limsup_{N\to\infty}\frac{\mathbb{E}(A_{N,3}^2)}{m^N}\leq\frac{m}{(\sqrt{m}-1)^2}. \tag{18}$$

Since $R_N - \mathbb{E}(R_N) = A_{N,1} + A_{N,2} + A_{N,3}$, Relations (16), (17) and (18) then imply that

(1) When G is non-deterministic, the expression $A_{N,2}$ dominates in $R_N - \mathbb{E}(R_N)$ so that $\mathrm{Var}(R_N)/\mathbb{E}(T_N)^2$ is converging to the right hand side of Equation (17).
(2) If $G \equiv m$, the term $A_{N,3}$ vanishes so that $\mathrm{Var}(R_N)/\mathbb{E}(T_N)$ is converging to the right hand side of Equation (16).

Equations (14) and (15) are established. □

As for the first moment of R_N, we turn now to the analysis of the behavior of of the second order characteristics defined by Equations (14) and (15) when λ is in the neighborhood of 0. For the non deterministic case, we have from Proposition 1

$$\lim_{\lambda \to 0} \frac{\rho_2^{(1)}(\lambda)}{(\lambda \log_m(1/\lambda))^2} = \frac{\mathrm{Var}(G)}{(m^2 - m)}.$$

In Proposition 2, the expression of $\rho_2^{(2)}(\lambda)$ is defined a priori only for a deterministic offspring distribution, but can be extended to any offspring distribution by using the right hand side of Equation (15). In the following, we study the behavior of $\rho_2^{(2)}(\lambda)$ for an arbitrary offspring distribution.

Lemma 1 (Asymptotic Behavior of $\lambda \to \rho_2^{(2)}(\lambda)$ at 0)**.** *Provided that the random variable G has a finite second moment, the function $\rho_2^{(2)}(\lambda)$ defined by Equation* (15) *is such that*

$$\lim_{\lambda \searrow 0} \frac{\rho_2^{(2)}(\lambda)}{\lambda (\log_m \lambda)^2} = 1. \tag{19}$$

Proof. Define

$$f_a(\lambda) \stackrel{\text{def.}}{=} \sum_{n=1}^{+\infty} \frac{m-1}{m^n} \left(\mathbb{E}\left(e^{-\lambda T_n}\right) \left(1 - \mathbb{E}\left(e^{-\lambda T_n}\right)\right)\right)$$

and

$$f_b(\lambda) \stackrel{\text{def.}}{=} \sum_{n=1}^{+\infty} \frac{m-1}{m^n} \sum_{k=0}^{n-1} \mathbb{E}\left(e^{-\lambda T_{n-k-1}} Z_{n-k} \mathbb{E}\left(e^{-\lambda T_k}\right)^{Z_{n-k}-1}\right) \mathbb{E}\left(e^{-\lambda T_k}\left(1 - e^{-\lambda T_k}\right)\right).$$

Equation (15) gives that $m\rho_2^{(2)}(\lambda) = 2f_b(\lambda) + f_a(\lambda)$.

Asymptotic behavior of f_a. With similar arguments as in the proof of Theorem 1 the asymptotic behavior of $f_a(\lambda)$ when λ goes to 0 is equivalent to the asymptotic behavior of

$$\sum_{n=1}^{+\infty} \frac{m-1}{m^n} \left(\mathbb{E}\left(\exp\left(-\lambda \frac{W m^{n+1}}{m-1}\right)\right) \left(1 - \mathbb{E}\left(\exp\left(-\lambda \frac{W m^{n+1}}{m-1}\right)\right)\right)\right).$$

If W_1 and W_2 are two independent random variables with the same distribution as W, the above series can be rewritten as

$$\sum_{n=1}^{+\infty} \frac{m-1}{m^n} \mathbb{E}\left(\exp\left(-\lambda \frac{W_1 m^{n+1}}{m-1}\right)\left(1-\mathbb{E}\left(\exp\left(-\lambda \frac{W_2 m^{n+1}}{m-1}\right)\right)\right)\right)$$
$$= (m-1)\sum_{n=1}^{+\infty}\left(\frac{1}{m^n}\mathbb{E}(h(\lambda(W_1+W_2)m^n/(m-1))) - \mathbb{E}(h(\lambda W_1 m^n/(m-1)))\right),$$

with $h(u) = 1 - e^{-u}$. Consequently,

$$\lim_{\lambda \to 0} \frac{f_a(\lambda)}{-\lambda \log_m \lambda} = m \tag{20}$$

by Proposition 1.

Asymptotic behavior of f_b. Let us fix some $\varepsilon > 0$ and assume that $\lambda < \varepsilon$. The function $f_b(\lambda)$ can be rewritten as

$$f_b(\lambda) = \sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} \frac{m-1}{m^n} S(n;\lambda) + \sum_{n=\lfloor \log_m(\varepsilon/\lambda)\rfloor+1}^{\infty} \frac{m-1}{m^n} S(n;\lambda) \tag{21}$$

where

$$S(n;\lambda) = \sum_{k=0}^{n-1} \mathbb{E}\left(e^{-\lambda T_{n-k-1}} Z_{n-k} \mathbb{E}\left(e^{-\lambda T_k}\right)^{Z_{n-k}-1}\right) \mathbb{E}\left(e^{-\lambda T_k}\left(1-e^{-\lambda T_k}\right)\right)$$

Since for $x \geq 0$, $e^{-x}(1-e^{-x}) \leq x$, we easily deduce that for all $n \geq 1$

$$S(n;\lambda) \leq \sum_{k=0}^{n-1} \mathbb{E}\left(Z_{n-k}\right) \mathbb{E}\left(\lambda T_k\right) \leq \frac{n\lambda m^{n+1}}{m-1}$$

and then

$$\sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} \frac{m-1}{m^n} S(n;\lambda) \leq \frac{m}{2}\lambda \log_m(\varepsilon/\lambda)(\log_m(\varepsilon/\lambda)+1).$$

The second term in the right hand side of Equation (21) can be written as

$$\sum_{n=\lfloor \log_m(\varepsilon/\lambda)\rfloor+1}^{\infty} \frac{m-1}{m^n} S(\lfloor \log_m(\varepsilon/\lambda)\rfloor + 1;\lambda) + \sum_{n=\lfloor \log_m(\varepsilon/\lambda)\rfloor+1}^{\infty} \frac{m-1}{m^n}\left(S(n;\lambda) - S(\lfloor \log_m(\varepsilon/\lambda)\rfloor + 1;\lambda)\right). \tag{22}$$

By using the fact that for $x > 0$ and $\alpha > 0$, $xe^{-\alpha x} \leq 1/\alpha$, we get that

$$S(\lfloor \log_m(\varepsilon/\lambda)\rfloor + 1;\lambda) \leq \frac{1}{\mathbb{E}(e^{-\lambda T_k})} \sum_{k=0}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} \frac{\lambda \mathbb{E}(T_k)}{-\log \mathbb{E}(e^{-\lambda T_k})}.$$

The relation $\mathbb{E}\left(\exp\left(-\lambda T_k\right)\right) \geq \exp\left(-\lambda \mathbb{E}(T_k)\right) \geq \exp(-m\varepsilon/(m-1))$ holds by Jensen's Inequality under the condition that $k \leq \lfloor \log_m(\varepsilon/\lambda)\rfloor$. In addition,

$$\mathbb{E}(T_n^2) \leq \left(\sum_{i=0}^{n} \sqrt{\mathbb{E}(Z_i^2)}\right)^2 \leq \frac{m^{2n}}{(m-1)^2}\left(\frac{\sigma^2}{m-1}+1\right),$$

where σ^2 is the variance of the random variable G, so that for $k \leq \lfloor \log_m(\varepsilon/\lambda) \rfloor$

$$\frac{\lambda \mathbb{E}(T_k^2)}{\mathbb{E}(T_k)} \leq \frac{\lambda m^k m^k}{(m-1)(mm^k-1)} \left(\frac{\sigma^2}{m-1} + 1 \right) \leq \frac{\varepsilon}{(m-1)^2} \left(\frac{\sigma^2}{m-1} + 1 \right). \tag{23}$$

Since for $x \geq 0$, $e^{-x} \leq 1 - x + x^2/2$, we have

$$\mathbb{E}(e^{-\lambda T_k}) \leq 1 - \lambda \mathbb{E}(T_k) + \frac{\mathbb{E}\left((\lambda T_k)^2\right)}{2} \leq 1 - \lambda \mathbb{E}(T_k) \left(1 - \frac{\varepsilon}{(m-1)^2} \left(\frac{\sigma^2}{m-1} + 1 \right) \right).$$

and then, for $k \leq \lfloor \log_m(\varepsilon/\lambda) \rfloor$,

$$\begin{aligned} \frac{\lambda \mathbb{E}(T_k)}{-\log \mathbb{E}(e^{-\lambda T_k})} &\leq \frac{\lambda \mathbb{E}(T_k)}{1 - \mathbb{E}(e^{-\lambda T_k})} \\ &\leq \left(1 - \frac{\varepsilon}{(m-1)^2} \left(\frac{\sigma^2}{m-1} + 1 \right) \right)^{-1} \stackrel{\text{def.}}{=} \kappa(\varepsilon) \end{aligned} \tag{24}$$

as long as

$$\varepsilon < (m-1)^2 \Big/ \left(\frac{\sigma^2}{m-1} + 1 \right) \stackrel{\text{def.}}{=} \varepsilon_1.$$

It follows that for $\varepsilon < \varepsilon_1$,

$$S(\lfloor \log_m(\varepsilon/\lambda) \rfloor + 1; \lambda) \leq (1 + \log_m(\varepsilon/\lambda)) e^{m\varepsilon/(m-1)} \kappa(\varepsilon).$$

and therefore,

$$\sum_{n=\lfloor \log_m(\varepsilon/\lambda) \rfloor + 1}^{\infty} \frac{m-1}{m^n} S(\lfloor \log_m(\varepsilon/\lambda) \rfloor + 1; \lambda) \leq \frac{\lambda}{\varepsilon} \log_m(\varepsilon/\lambda) e^{m\varepsilon/(m-1)} \kappa(\varepsilon),$$

which is $o(\lambda(\log \lambda)^2)$ when $\lambda \to 0$. In addition, the second term in the right hand side of Equation (22) can be rewritten as

$$\sum_{k=\lfloor \log_m(\varepsilon/\lambda) \rfloor + 1}^{\infty} \frac{m-1}{m^k} \sum_{n=1}^{\infty} \frac{1}{m^n} \mathbb{E}\left(e^{-\lambda T_{n-1}} Z_n \mathbb{E}(e^{-\lambda T_k})^{Z_n - 1} \right) \mathbb{E}\left(e^{-\lambda T_k} \left(1 - e^{-\lambda T_k} \right) \right).$$

We first note that

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{1}{m^n} \mathbb{E}\left(e^{-\lambda T_{n-1}} Z_n \mathbb{E}(e^{-\lambda T_k})^{Z_n - 1} \right) = \\ \sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda) \rfloor} \frac{1}{m^n} \mathbb{E}\left(e^{-\lambda T_{n-1}} Z_n \mathbb{E}(e^{-\lambda T_k})^{Z_n - 1} \right) \\ + \sum_{n=\lfloor \log_m(\varepsilon/\lambda) \rfloor + 1}^{\infty} \frac{1}{m^n} \mathbb{E}\left(e^{-\lambda T_{n-1}} Z_n \mathbb{E}(e^{-\lambda T_k})^{Z_n - 1} \right). \end{aligned}$$

The first term in the right hand side of the above equation is less than or equal to the quantity $\log_m(\varepsilon/\lambda)$ since $\mathbb{E}(Z_n) = m^n$. The second term can be upper bounded

as

$$\sum_{n=\lfloor \log_m(\varepsilon/\lambda)\rfloor+1}^{\infty} \frac{1}{m^n}\mathbb{E}\left(e^{-\lambda T_{n-1}} Z_n \mathbb{E}(e^{-\lambda T_k})^{Z_n-1}\right)$$
$$\leq \sum_{n=\lfloor \log_m(\varepsilon/\lambda)\rfloor+1}^{\infty} \frac{1}{m^n}\mathbb{E}\left(Z_n e^{-\lambda Z_n}\right) \leq \frac{1}{(m-1)\varepsilon},$$

where we have used the fact that $T_k \geq 1$ for all $k \geq 0$ and $xe^{-\lambda x} \leq 1/(e\lambda)$ for all $x > 0$. It follows that the second term in the right hand side of Equation (22) is upper bounded by the quantity

$$\frac{\lambda}{\varepsilon}\left(\log_m(\varepsilon/\lambda) + \frac{1}{(m-1)\varepsilon}\right),$$

which is $o(\lambda(\log_m \lambda)^2)$ when $\lambda \to 0$.

By using the above inequalities, we come up with the conclusion that for every $\varepsilon > 0$,

$$\limsup_{\lambda\to 0} \frac{f_b(\lambda)}{\lambda(\log_m \lambda)^2} \leq \frac{m}{2}. \tag{25}$$

For establishing a lower bound for $f_b(\lambda)$, we introduce the size-biased Galton-Watson branching process. The sequence of random variables (Z_n/m^n) being a positive martingale, it induces a probability distribution $\widetilde{\mathbb{P}}$ such that, for any $n \geq 1$ and any random variable Y measurable with respect to the random variables $Z_1, \ldots, Z_n$,

$$\int Y d\widetilde{\mathbb{P}} = \mathbb{E}\left(Y\frac{Z_n}{m^n}\right).$$

It is known, see Lyons and Peres [7], that under the probability $\widetilde{\mathbb{P}}$, the sequence (Z_n) as the same distribution as a branching process with immigration $(\widetilde{Z}_n)$ where the number of children has the same distribution as G and the number of new immigrants is distributed as $\widetilde{G}$ such that $\mathbb{P}(\widetilde{G} = n) = n\mathbb{P}(G = n)/m$. If $\widetilde{Z}_0 = 1$, it is easy to check that

$$\widetilde{\mathbb{E}}\left(\widetilde{Z}_n\right) = m^n + \frac{m^n - 1}{m(m-1)}\mathbb{E}(G^2).$$

If $\widetilde{T}_n = \widetilde{Z}_0 + \widetilde{Z}_1 + \cdots + \widetilde{Z}_n$, we have by Jensen inequality

$$\mathbb{E}\left(e^{-\lambda T_{n-k-1}}\ \frac{Z_{n-k}}{m^{n-k}}\mathbb{E}\left(e^{-\lambda T_k}\right)^{Z_{n-k}-1}\right) = \widetilde{\mathbb{E}}\left(e^{-\lambda \widetilde{T}_{n-k-1}}\mathbb{E}\left(e^{-\lambda T_k}\right)^{\widetilde{Z}_{n-k}-1}\right)$$
$$\geq \widetilde{\mathbb{E}}\left(e^{-\lambda \widetilde{T}_{n-k}}\mathbb{E}\left(e^{-\lambda T_k}\right)^{\widetilde{T}_{n-k}}\right)$$
$$\geq \exp\left(-\lambda(1+\mathbb{E}(T_k))\widetilde{\mathbb{E}}(\widetilde{T}_{n-k})\right)$$
$$\geq \exp\left(-\frac{\lambda m}{m-1}\left(m^{n-k} + \frac{m^{n+1}}{(m-1)}\right)\left(1+\frac{g_2}{m}\right)\right)$$

since

$$\widetilde{\mathbb{E}}(\widetilde{T}_n) \leq \frac{m^{n+1}}{m-1}\left(1+\frac{g_2}{m}\right),$$

where $g_2 = \mathbb{E}(G^2)$. In addition, by using the fact that $e^{-x}(1-e^{-x}) \geq x - 2x^2$ holds for $x > 0$, we have

$$\sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} \frac{m-1}{m^n} S(n;\lambda) \geq \sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} (m-1)$$
$$\times \sum_{k=0}^{n-1} \frac{1}{m^k} \exp\left(-\frac{\lambda m}{m-1}\left(m^{n-k} + \frac{m^{n+1}}{(m-1)}\right)\left(1+\frac{g_2}{m}\right)\right) \mathbb{E}\left(\lambda T_k - 2(\lambda T_k)^2\right)$$
$$\geq \exp\left(-\frac{\varepsilon m}{(m-1)}\left(1+\frac{m}{(m-1)}\right)\left(1+\frac{g_2}{m}\right)\right)$$
$$\sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} (m-1) \sum_{k=0}^{n-1} \frac{\lambda \mathbb{E}(T_k)}{m^k}\left(1 - \frac{\lambda \mathbb{E}(T_k^2)}{\mathbb{E}(T_k)}\right).$$

By using Inequality (23) and Definition (24), we have, for $\varepsilon < \varepsilon_1$,

$$\sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} (m-1) \sum_{k=0}^{n-1} \frac{\lambda \mathbb{E}(T_k)}{m^k}\left(1 - \frac{\lambda \mathbb{E}(T_k^2)}{\mathbb{E}(T_k)}\right) \geq \frac{\lambda}{\kappa(\varepsilon)} \sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} \sum_{k=0}^{n-1} \frac{m^{k+1}-1}{m^k}$$

Since

$$\sum_{n=1}^{\lfloor \log_m(\varepsilon/\lambda)\rfloor} \sum_{k=0}^{n-1} \frac{m^{k+1}-1}{m^k} = \frac{m}{2}\lfloor \log_m(\varepsilon/\lambda)\rfloor(\lfloor \log_m(\varepsilon/\lambda)\rfloor + 1) + \frac{m\lfloor \log_m(\varepsilon/\lambda)\rfloor}{m-1}$$
$$- \frac{m}{(m-1)^2}\left(\frac{1}{m^{\lfloor \log_m(\varepsilon/\lambda)\rfloor}} - 1\right)$$

and since we already know that the second term in the right hand side of Equation (21) is $o(\lambda(\log_m(\lambda))^2$ when $\lambda \to 0$, we then deduce that for all $\varepsilon \in (0, \varepsilon_1)$

$$\liminf_{\lambda\to 0} \frac{f_b(\lambda)}{\lambda(\log_m \lambda)^2} \geq \frac{m}{2\kappa(\varepsilon)} \exp\left(-\frac{\varepsilon m}{(m-1)}\left(1+\frac{m}{(m-1)}\right)\left(1+\frac{g_2}{m}\right)\right)$$

and hence,

$$\liminf_{\lambda\to 0} \frac{f_b(\lambda)}{\lambda(\log_m \lambda)^2} \geq \frac{m}{2}. \tag{26}$$

Combining Equations (20), (25) and (26), Equation (19) follows. □

Proposition 3. *The functions $\rho_2^{(1)}(\lambda)$ and $\rho_2^{(2)}(\lambda)$ are such that*

$$\lim_{\lambda\to 0} \frac{\rho_2^{(1)}(\lambda)}{(\lambda \log_m \lambda)^2} = \frac{1}{m^2-m}\mathrm{Var}(G), \tag{27}$$

$$\lim_{\lambda\to 0} \frac{\rho_2^{(2)}(\lambda)}{\lambda(\log_m \lambda)^2} = 1. \tag{28}$$

where G is the random variable describing the offspring of a node.

From Theorem 1, we observe that the size of the sampled tree scales with the size of the original tree. The same phenomenon is true for the squared coefficient of variation of the size of the sampled tree if an only if the offspring distribution is not deterministic as shown by Proposition 2. In the case of a deterministic offspring distribution, when $\lambda \to 0$, the squared coefficient of variation is approximately equal

to $1/(\lambda\mathbb{E}(T_N))$ for large N. The quantity $\lambda\mathbb{E}(T_N)$ is precisely the mean number of selected points. This indicates that the distribution of the random variable R_N is concentrated around its mean value. There is almost no randomness in the discovered tree.

5. The Depth Biased Model

In this section, it is assumed that conditionally on the tree, for $n \geq 0$, a node at depth n is chosen with probability $(1 - \exp(-(\alpha/m)^n))$ for some $\alpha \in [0, 1)$. The mean number of selected nodes at depth n in the tree is equal to $m^n(1 - \exp(-(\alpha/m)^n)) \sim \alpha^n$ and the total number of selected nodes in the whole tree $\mathcal{N}(\mathcal{T})$ is such that

$$\frac{1}{1-\alpha} - \frac{1}{2(1-\alpha^2/m)} \leq \mathbb{E}(\mathcal{N}(\mathcal{T})) = \sum_{n=0}^{\infty} m^n(1 - e^{-(\alpha/m)^n}) \leq \frac{1}{1-\alpha},$$

in particular the mean number of selected nodes $\mathcal{N}(\mathcal{T}) \sim 1/(1-\alpha)$ when $\alpha \to 1$. The behavior of the size $R(\alpha)$ of the sampled tree is used to estimate the speed of the exploration process, when the number of selected nodes becomes large. We first give the expression of the mean value $\mathbb{E}(R(\alpha))$ of the size of the sampled tree.

Lemma 2. *The mean value of the size of the sampled tree in the depth biased model is given by*

$$\mathbb{E}(R(\alpha)) = \sum_{n=0}^{\infty} m^n \left(1 - \mathbb{E}\left(\exp\left(-\left(\frac{\alpha}{m}\right)^n \sum_{i=0}^{\infty} \alpha^i \frac{Z_i}{m^i}\right)\right)\right). \tag{29}$$

Proof. As in the previous section, for $n \geq 0$ and $1 \leq \ell \leq Z_n$, the symbol $\mathcal{T}^{n,\ell}$ denotes the sub-tree of $\mathcal{T}$ whose root is (n, ℓ). The node (n, ℓ) is in the sampled tree if $\mathcal{N}(\mathcal{T}^{n,\ell}) \neq 0$. Since nodes at a given depth are selected independently one of each other, we have

$$\mathbb{P}\left(\mathcal{N}(\mathcal{T}^{n,\ell}) \neq 0 \mid \mathcal{T}, (n,\ell) \in \mathcal{T}\right) = 1 - \exp\left(-\left(\frac{\alpha}{m}\right)^n \sum_{i=0}^{\infty} \alpha^i \frac{Z_i^{(n,\ell)}}{m^i}\right),$$

where $Z_i^{(n,\ell)}$ is the number of descendants of (n, ℓ) at generation i and where we have used the fact that the sub-tree $\mathcal{T}^{n,\ell}$ has the same offspring distribution as the original tree $\mathcal{T}$. Hence,

$$\mathbb{P}\left(\mathcal{N}(\mathcal{T}^{n,\ell}) \neq 0 \mid (n,\ell) \in \mathcal{T}\right) = 1 - \mathbb{E}\left(\exp\left(-\left(\frac{\alpha}{m}\right)^n \sum_{i=0}^{\infty} \alpha^i \frac{Z_i}{m^i}\right)\right).$$

It follows that the size of the sampled tree given by

$$R(\alpha) = \sum_{n=0}^{+\infty} \sum_{\ell=1}^{Z_n} \mathbb{1}_{\{\mathcal{N}(\mathcal{T}^{\ell,n}) \neq 0\}} \tag{30}$$

and its mean value is, by using the independence in the selection of nodes,

$$\mathbb{E}(R(\alpha)) = \sum_{n=0}^{+\infty} \mathbb{E}(Z_n)\mathbb{P}\left(\mathcal{N}(\mathcal{T}^{n,1}) \neq 0\right),$$

Equation (29) follows. □

The growth rate of the exploration process is defined by the ratio

$$\frac{\mathbb{E}\left(R(\alpha)\right)}{\mathbb{E}(\mathcal{N}(\mathcal{T}))} = \frac{1}{\eta(\alpha)} \sum_{n=0}^{+\infty} (1-\alpha) m^n \left(1 - \mathbb{E}\left(\exp\left(-\left(\frac{\alpha}{m}\right)^n \sum_{i=0}^{\infty} \alpha^i \frac{Z_i}{m^i}\right)\right)\right),$$

where $\eta(\alpha) = (1-\alpha)\mathbb{E}(\mathcal{N}(\mathcal{T})) \to 1$ when $\alpha \to 1$.

Theorem 2. *If $\mathbb{E}(G^2) < +\infty$, as $\alpha \nearrow 1$, the following limit relation holds*

$$\lim_{\alpha \to 1} \frac{\mathbb{E}\left(R(\alpha)\right)}{\mathbb{E}(\mathcal{N}(\mathcal{T}))^2} = 1.$$

Proof. Let us first introduce the function

$$H(\alpha) = \sum_{n=0}^{+\infty} (1-\alpha) m^n \left(1 - \mathbb{E}\left(\exp\left(-\left(\frac{\alpha}{m}\right)^n \frac{W}{1-\alpha}\right)\right)\right),$$

where W is defined by Equation (9). We have

$$\begin{aligned}
\left| H(\alpha) - \eta(\alpha) \frac{\mathbb{E}\left(R(\alpha)\right)}{\mathbb{E}(\mathcal{N}(\mathcal{T}))} \right| &\leq (1-\alpha) \sum_{n=0}^{\infty} \alpha^n \left| \sum_{i=0}^{\infty} \alpha^i \mathbb{E}\left(\frac{Z_i}{m^i} - W\right) \right| \qquad (31) \\
&\leq \sum_{i=0}^{\infty} \alpha^i \mathbb{E}\left(\left|\frac{Z_i}{m^i} - W\right|\right) \qquad (32) \\
&\leq \mathrm{Var}(W) \sum_{i=0}^{\infty} \left(\frac{\alpha}{\sqrt{m}}\right)^i = \frac{\mathrm{Var}(W)}{1 - \frac{\alpha}{\sqrt{m}}},
\end{aligned}$$

where we have used Inequality (13) in the last step.

Let us define the family of non-negative random variables $\mathcal{H}_\alpha$, $0 < \alpha < 1$ by

$$\mathcal{H}_\alpha = \sum_{n=0}^{+\infty} (1-\alpha)^2 m^n \left(1 - \exp\left(-\left(\frac{\alpha}{m}\right)^n \frac{W}{1-\alpha}\right)\right).$$

We have $(1-\alpha)H(\alpha) = \mathbb{E}(\mathcal{H}_\alpha)$.

Let us fix some $\varepsilon > 0$. Since $W > 0$ a.s., we can define the quantity

$$n(W,\alpha) = \max\left(\left\lceil \log_{m/\alpha}\left(\frac{W}{(\varepsilon(1-\alpha))}\right)\right\rceil, 0\right).$$

For $n \geq n(W,\alpha)$,

$$\left(\frac{\alpha}{m}\right)^n \frac{W}{1-\alpha} < \varepsilon.$$

By using the fact that for $x \geq 0$, $1 - e^{-x} \geq x - x^2/2$, we have

$$\mathcal{H}_\alpha \geq \sum_{n=n(W,\alpha)}^{+\infty} (1-\alpha)^2 m^n \left(1 - \exp\left(-\left(\frac{\alpha}{m}\right)^n \frac{W}{1-\alpha}\right)\right) \geq \alpha^{n(W,\alpha)} W(1-\varepsilon).$$

Since the above inequality is valid for all $\varepsilon > 0$ and $\alpha^{n(W,\alpha)}$ converges to 1 as $\alpha \nearrow 1$, it follows that $\liminf_{\alpha \to 1} \mathcal{H}_\alpha \geq W$ a.s.

Since $1 - e^{-x} \leq x$ for $x \geq 0$, we have

$$\mathcal{H}_\alpha \leq W \quad \text{a.s.}$$

and then $\limsup_{\alpha\to 1}\mathcal{H}_\alpha \le W$ a.s. Hence, $\limsup_{\alpha\to 1}\mathcal{H}_\alpha = W$ a.s. Since the family $(\mathcal{H}_\alpha)$ is non negative and bounded by W, which is integrable, we have

$$\lim_{\alpha\to 1}\mathbb{E}(\mathcal{H}_\alpha) = \mathbb{E}(W) = 1$$

and the result follows by using Inequality (32). □

When $\alpha < 1$ the selected node are closed to the root and only a small fraction of the whole is discovered. When $\alpha \nearrow 1$, we can select nodes deeper in the tree but roughly only one node is selected in average at each level. The above result indicates that the average size of the discovered tree grows as the square of the average number of selected nodes.

References

1. Krishna B. Athreya and Peter E. Ney, *Branching processes*, Springer-Verlag, New York, 1972, Die Grundlehren der mathematischen Wissenschaften, Band 196.
2. Youssef Azzana, Fabrice Guillemin, and Philippe Robert, *A stochastic model for topology discovery of tree networks*, Proceedings of ITC'19 (Beijing), 2005.
3. Caida, *Skitter project*, url: http://www.caida.org/tools/measurement/skitter.
4. C. Christophi and H. Mahmoud, *The oscillatory distribution of distances in random tries*, The Annals of Applied Probability **15** (2005), 1536–1564.
5. L. Dall'Astra, I. Alvarez-Hameli, A. Barrat, A. Vásquez, and A. Vespignani, *A statistical approach to the traceroute exploration of networks: theory and simulations*, Available at arXiv:cond-mat/0406404, June 2004.
6. Philippe Flajolet, Xavier Gourdon, and Philippe Dumas, *Mellin transforms and asymptotics: harmonic sums*, Theoretical Computer Science **144** (1995), no. 1-2, 3–58, Special volume on mathematical analysis of algorithms.
7. R. Lyons and Y. Peres, *Probability on trees and networks*, Preprint, 2005.
8. Hosam M. Mahmoud, *Evolution of random search trees*, Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons Inc., New York, 1992, A Wiley-Interscience Publication.
9. Hosam M. Mahmoud and Ralph Neininger, *Distribution of distances in random binary search trees*, The Annals of Applied Probability **13** (2003), no. 1, 253–276.
10. Hanène Mohamed and Philippe Robert, *A probabilistic analysis of some tree algorithms*, Annals of Applied Probability **15** (2005), no. 4, 2445–2471. MR 2187300
11. Jacques Neveu, *Arbres et processus de Galton-Watson*, Annales de l'institut Henri Poincaré, Série B **22** (1986), 199–207.
12. Alois Panholzer, *Distribution of the Steiner distance in generalized M-ary search trees*, Combinatorics, Probability and Computing **13** (2004), no. 4-5, 717–733.
13. Alois Panholzer and Helmut Prodinger, *Spanning tree size in random binary search trees*, The Annals of Applied Probability **14** (2004), no. 2, 718–733.
14. Dimes project, url: http://www.netdimes.org/.
15. Philippe Robert, *On the asymptotic behavior of some algorithms*, Random Structures and Algorithms **27** (2005), no. 2, 235–250.

(F. Guillemin) Orange Labs, 2, Avenue Pierre Marzin, F-22300 Lannion
E-mail address: Fabrice.Guillemin@orange-ftgroup.com

(Ph. Robert) INRIA-Rocquencourt, RAP project, Domaine de Voluceau, 78153 Le Chesnay, France
E-mail address: Philippe.Robert@inria.fr
URL: http://www-rocq.inria.fr/~robert