*Proceedings of the 2015 Winter Simulation Conference*
*L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.*

# EFFICIENT SIMULATION FOR BRANCHING LINEAR RECURSIONS

Ningyuan Chen
Mariana Olvera-Cravioto

Industrial Engineering and Operations Research
Columbia University
New York, NY 10027, USA

## ABSTRACT

We provide an algorithm for simulating the unique attracting fixed-point of linear branching distributional equations. Such equations appear in the analysis of information ranking algorithms, e.g., PageRank, and in the complexity analysis of divide and conquer algorithms, e.g., Quicksort. The naive simulation approach would be to simulate exactly a suitable number of generations of a weighted branching process, which has exponential complexity in the number of generations being sampled. Instead, we propose an iterative bootstrap algorithm that has linear complexity; we prove its convergence and the consistency of a family of estimators based on our approach.

## 1 INTRODUCTION

The complexity analysis of divide and conquer algorithms such as Quicksort (Rösler 1991, Fill and Janson 2001, Rösler and Rüschendorf 2001) and the more recent analysis of information ranking algorithms on complex graphs (e.g., Google's PageRank) (Volkovich and Litvak 2010, Jelenković and Olvera-Cravioto 2010, Chen, Litvak, and Olvera-Cravioto 2014) motivate the analysis of the stochastic fixed-point equation

$$R \stackrel{\mathscr{D}}{=} Q + \sum_{r=1}^{N} C_r R_r, \tag{1}$$

where $(Q, N, C_1, C_2, \ldots)$ is a real-valued random vector with $N \in \mathbb{N}$, and $\{R_i\}_{i \in \mathbb{N}}$ is a sequence of i.i.d. copies of $R$, independent of $(Q, N, C_1, C_2, \ldots)$. More precisely, the number of comparisons required in Quicksort for sorting an array of length $n$, properly normalized, satisfies in the limit as the array's length grows to infinity a distributional equation of the form in (1). In the context of ranking algorithms, it has been shown that the rank of a randomly chosen node in a large directed graph with $n$ nodes converges in distribution, as the size of the graph grows, to $R$, where $N$ represents the in-degree of the chosen node and the $\{C_i\}_{i \geq 1}$ are functions of the out-degree and node attributes of its neighbors. In the complexity analysis of algorithms, knowing the distribution of $R$ makes it possible to estimate the moments and tail probabilities of the number of operations required to sort a list of numbers, which is important for benchmarking and worst case analysis. In the case of information ranking algorithms, the distribution of $R$ can be used to determine what type of nodes are typically ranked highly, which in turn can be used to design new ranking algorithms capable of identifying pre-specified data attributes.

As further motivation for the study of branching fixed-point equations, we mention the closely related maximum equation

$$R \stackrel{\mathscr{D}}{=} Q \vee \bigvee_{r=1}^{N} C_r R_r, \tag{2}$$

978-1-4673-9743-8/15/$31.00 ©2015 IEEE

with $(Q,N,C_1,C_2,\ldots)$ nonnegative, which has been shown to appear in the analysis of the waiting time distribution in large queueing networks with parallel servers and synchronization requirements (Karpelevich, Kelbert, and Suhov 1994, Olvera-Cravioto and Ruiz-Lacedelli 2014). In this setting, $W = \log R$ represents the waiting time in stationarity of a job, that upon arrival to the network, is split into a number of subtasks requiring simultaneous service from a random subset of servers. Computing the distribution and the moments of $W$ is hence important for evaluating the performance of such systems (e.g., implementations of MapReduce and similar algorithms in today's cloud computing). Due to length limitations, we focus in this paper only on (1), but we mention that the algorithm we provide can easily be adapted to approximately simulate the solutions to (2) (see Remark 2).

Although the study of (1) and (2), and other max-plus branching recursions, has received considerable attention in the recent years (Rösler 1991, Biggins 1998, Fill and Janson 2001, Rösler and Rüschendorf 2001, Aldous and Bandyopadhyay 2005, Alsmeyer, Biggins, and Meiners 2012, Alsmeyer and Meiners 2012, Alsmeyer and Meiners 2013, Jelenković and Olvera-Cravioto 2012b, Jelenković and Olvera-Cravioto 2012a, Jelenković and Olvera-Cravioto 2015), the current literature only provides results on the characterization of the solutions to (1) and (2), their tail asymptotics, and in some instances, their integer moments, which is not always enough for the applications mentioned above. It is therefore of practical importance to have a numerical approach to estimate both the distribution and the general moments of $R$.

As a mathematical observation, we mention that both (1) and (2) are known to have multiple solutions (see e.g. Biggins (1998), Alsmeyer, Biggins, and Meiners (2012), Alsmeyer and Meiners (2012), Alsmeyer and Meiners (2013) and the references therein for the characterization of the solutions). However, in applications we are often interested in the so-called endogenous solution. This endogenous solution is the unique limit under iterations of the distributional recursion

$$R^{(k+1)} \stackrel{\mathscr{D}}{=} \sum_{r=1}^{N} C_r R_r^{(k)} + Q, \tag{3}$$

where $(Q,N,C_1,C_2,\ldots)$ is a real-valued random vector with $N \in \mathbb{N}$, and $\{R_i^{(k)}\}_{i\in\mathbb{N}}$ is a sequence of i.i.d. copies of $R^{(k)}$, independent of $(Q,N,C_1,C_2,\ldots)$, provided one starts with an initial distribution for $R^{(0)}$ with sufficient finite moments (see, e.g., Lemma 4.5 in Jelenković and Olvera-Cravioto (2012a)). Moreover, asymptotics for the tail distribution of the endogenous solution $R$ are available under several different sets of assumptions for $(Q,N,C_1,C_2,\ldots)$ (Jelenković and Olvera-Cravioto 2010, Jelenković and Olvera-Cravioto 2012b, Jelenković and Olvera-Cravioto 2012a, Olvera-Cravioto 2012).

As will be discussed later, the endogenous solution to (1) can be explicitly constructed on a weighted branching process. Thus, drawing some similarities with the analysis of branching processes, and the Galton-Watson process in particular, one could think of using the Laplace transform of $R$ to obtain its distribution. Unfortunately, the presence of the weights $\{C_i\}$ in the Laplace transform

$$\varphi(s) = E\left[\exp(-sR)\right] = E\left[\exp(-sQ)\prod_{i=1}^{N}\varphi(sC_i)\right]$$

makes its inversion problematic, making a simulation approach even more necessary.

The first observation we make regarding the simulation of $R$, is that when $P(Q=0)<1$ it is enough to be able to approximate $R^{(k)}$ for fixed values of $k$, since both $R^{(k)}$ and $R$ can be constructed in the same probability space in such a way that the difference $|R^{(k)} - R|$ is geometrically small. More precisely, under very general conditions (see Proposition 1 in Section 2), there exist positive constants $K<\infty$ and $c<1$ such that

$$E\left[\left|R^{(k)} - R\right|^{\beta}\right] \leq Kc^{k+1}. \tag{4}$$

Our goal is then to simulate $R^{(k)}$ for a suitably large value of $k$.

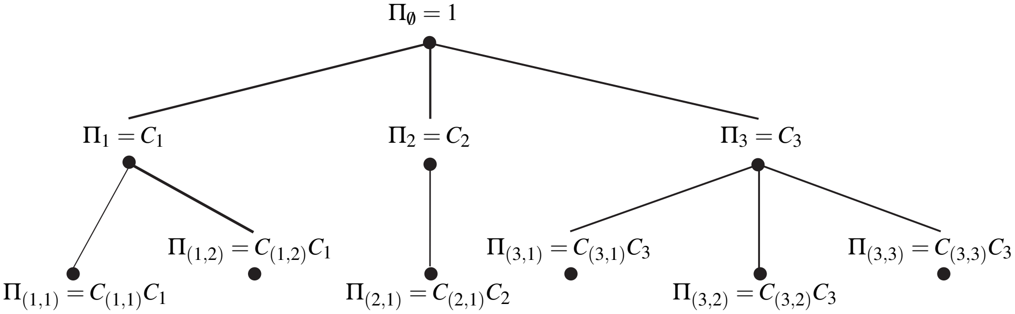

$\Pi_\emptyset = 1$
$\Pi_1 = C_1$
$\Pi_2 = C_2$
$\Pi_3 = C_3$
$\Pi_{(1,1)} = C_{(1,1)}C_1$
$\Pi_{(1,2)} = C_{(1,2)}C_1$
$\Pi_{(2,1)} = C_{(2,1)}C_2$
$\Pi_{(3,1)} = C_{(3,1)}C_3$
$\Pi_{(3,2)} = C_{(3,2)}C_3$
$\Pi_{(3,3)} = C_{(3,3)}C_3$

Figure 1: Weighted branching process

The simulation of $R^{(k)}$ is not that straightforward either, since the naive approach of simulating i.i.d. copies of $(Q,N,C_1,C_2,\dots)$ to construct a single realization of a weighted branching process, up to say $k$ generations, is of order $(E[N])^k$. Considering that in the examples mentioned earlier we typically have $E[N] > 1$ ($N \equiv 2$ for Quicksort, $E[N] \approx 30$ in many information ranking applications, and $E[N]$ in the hundreds for MapReduce implementations), this approach is prohibitive. Instead, we propose in this paper an iterative bootstrap algorithm that outputs a sample pool of observations $\{\hat{R}_i^{(k,m)}\}_{i=1}^m$ whose empirical distribution converges, in the Kantorovich-Rubinstein distance, to that of $R^{(k)}$ as the size of the pool $m \to \infty$. This mode of convergence is equivalent to weak convergence and convergence of the first absolute moments (see, e.g., Villani (2009)). Moreover, the complexity of our proposed algorithm is linear in $k$. This algorithm is known in the statistical physics literature as "population dynamics" (see, e.g., Mezard and Montanari (2009)), where it has been used heuristically for the approximation of belief propagation algorithms.

The paper is organized as follows. Section 2 describes the weighted branching process and the linear recursion. The algorithm itself is given in Section 3, which includes a remark on how to adapt it to the maximum case. Section 4 introduces the Kantorovich-Rubinstein distance and proves the convergence properties of our proposed algorithm. Numerical examples to illustrate the precision of the algorithm are presented in Section 5.

## 2 LINEAR RECURSIONS ON WEIGHTED BRANCHING PROCESSES

As mentioned in the introduction, the endogenous solution to (1) can be explicitly constructed on a weighted branching process. To describe the structure of a weighted branching process, let $\mathbb{N}_+ = \{1,2,3,\dots\}$ be the set of positive integers and let $U = \bigcup_{k=0}^\infty (\mathbb{N}_+)^k$ be the set of all finite sequences $\mathbf{i} = (i_1,i_2,\dots,i_n)$, $n \geq 0$, where by convention $\mathbb{N}_+^0 = \{\emptyset\}$ contains the null sequence $\emptyset$. To ease the exposition, we will use $(\mathbf{i},j) = (i_1,\dots,i_n,j)$ to denote the index concatenation operation.

Next, let $(Q,N,C_1,C_2,\dots)$ be a real-valued vector with $N \in \mathbb{N}$. We will refer to this vector as the generic branching vector. Now let $\{(Q_\mathbf{i},N_\mathbf{i},C_{(\mathbf{i},1)},C_{(\mathbf{i},2)},\dots)\}_{\mathbf{i}\in U}$ be a sequence of i.i.d. copies of the generic branching vector. To construct a weighted branching process we start by defining a tree as follows: let $A_0 = \{\emptyset\}$ denote the root of the tree, and define the $n$th generation according to the recursion

$$A_n = \{(\mathbf{i},i_n) \in U : \mathbf{i} \in A_{n-1}, 1 \leq i_n \leq N_\mathbf{i}\}, \quad n \geq 1.$$

Now, assign to each node $\mathbf{i}$ in the tree a weight $\Pi_\mathbf{i}$ according to the recursion

$$\Pi_\emptyset \equiv 1, \qquad \Pi_{(\mathbf{i},i_n)} = C_{(\mathbf{i},i_n)}\Pi_\mathbf{i}, \qquad n \geq 1,$$

see Figure 1. If $P(N < \infty) = 1$ and $C_i \equiv 1$ for all $i \geq 1$, the weighted branching process reduces to a Galton-Watson process.

For a weighted branching process with generic branching vector $(Q, N, C_1, C_2, \ldots)$, define the process $\{R^{(k)} : k \geq 0\}$ as follows:

$$R^{(k)} = \sum_{j=0}^{k} \sum_{\mathbf{i} \in A_j} Q_{\mathbf{i}} \Pi_{\mathbf{i}}, \quad k \geq 0. \tag{5}$$

By focusing on the branching vector belonging to the root node, i.e., $(Q_\emptyset, N_\emptyset, C_1, C_2, \ldots)$ we can see that the process $\{R^{(k)}\}$ satisfies the distributional equations

$$\begin{aligned} R^{(0)} &= Q_\emptyset \stackrel{\mathscr{D}}{=} Q \\ R^{(k)} &= Q_\emptyset + \sum_{r=1}^{N_\emptyset} C_r \left( \sum_{j=1}^{k} \sum_{(r,\mathbf{i}) \in A_j} Q_{(r,\mathbf{i})} \Pi_{(r,\mathbf{i})} / C_r \right) \stackrel{\mathscr{D}}{=} Q + \sum_{r=1}^{N} C_r R_r^{(k-1)}, \qquad k \geq 1, \end{aligned} \tag{6}$$

where $R_r^{(k-1)}$ are i.i.d. copies of $R^{(k-1)}$, all independent of $(Q, N, C_1, C_2, \ldots)$. Here and throughout the paper the convention is that $XY/Y \equiv 1$ if $Y = 0$. Moreover, if we define

$$R = \sum_{j=0}^{\infty} \sum_{\mathbf{i} \in A_j} Q_{\mathbf{i}} \Pi_{\mathbf{i}}, \tag{7}$$

we have the following result. We use $x \vee y$ to denote the maximum of $x$ and $y$.

**Proposition 1** Let $\beta \geq 1$ be such that $E[|Q|^\beta] < \infty$ and $E\left[ \left( \sum_{i=1}^{N} |C_i| \right)^\beta \right] < \infty$. In addition, assume either (i) $(\rho_1 \vee \rho_\beta) < 1$ , or (ii) $\beta = 2$, $\rho_1 = 1$, $\rho_\beta < 1$ and $E[Q] = 0$. Then, there exist constants $K_\beta > 0$ and $0 < c_\beta < 1$ such that for $R^{(k)}$ and $R$ defined according to (5) and (7), respectively, we have

$$\sup_{k \geq 0} E\left[ |R^{(k)}|^\beta \right] \leq K_\beta < \infty \qquad \text{and} \qquad E\left[ |R^{(k)} - R|^\beta \right] \leq K_\beta c_\beta^{k+1}.$$

*Proof.* For the case $\rho_1 \vee \rho_\beta < 1$, Lemma 4.4 in Jelenković and Olvera-Cravioto (2012a) gives that for $W_n = \sum_{\mathbf{i} \in A_n} Q_{\mathbf{i}} \Pi_{\mathbf{i}}$ and some finite constant $H_\beta$ we have

$$E\left[ |W_n|^\beta \right] \leq H_\beta (\rho_1 \vee \rho_\beta)^n.$$

Let $c_\beta = \rho_1 \vee \rho_\beta$. Minkowski's inequality then gives

$$\left|\left| R^{(k)} \right|\right|_\beta \leq \sum_{n=0}^{k} ||W_n||_\beta \leq \sum_{n=0}^{\infty} \left( H_\beta c_\beta^n \right)^{1/\beta} = \left( \frac{H_\beta}{1 - c_\beta^{1/\beta}} \right)^{1/\beta} \triangleq (K_\beta)^{1/\beta} < \infty.$$

Similarly,

$$\left|\left| R^{(k)} - R \right|\right|_\beta \leq \sum_{n=k+1}^{\infty} ||W_n||_\beta \leq \sum_{n=k+1}^{\infty} \left( H_\beta c_\beta^n \right)^{1/\beta} = c_\beta^{(k+1)/\beta} \left( \frac{H_\beta}{1 - (\rho_1 \vee \rho_\beta)^{1/\beta}} \right)^{1/\beta} = \left( K_\beta c_\beta^{k+1} \right)^{1/\beta}.$$

For the case $\beta = 2$, $\rho_1 = 1$, $\rho_\beta < 1$ and $E[Q] = 0$ we have that

$$E\left[ W_n^2 \right] = E\left[ \left( \sum_{r=1}^{N_\emptyset} C_r W_{n-1,r} \right)^2 \right] = E\left[ \sum_{r=1}^{N_\emptyset} C_r^2 (W_{n-1,r})^2 + \sum_{1 \leq r \neq s \leq N_\emptyset} C_r C_s W_{n-1,r} W_{n-1,s} \right],$$

where $W_{n-1,r} = \sum_{(r,\mathbf{i})\in A_n} Q_{(r,\mathbf{i})}\Pi_{(r,\mathbf{i})}/C_r$, and the $\{W_{n-1,r}\}_{r\geq 1}$ are i.i.d. copies of $W_{n-1}$, independent of $(N_\emptyset, C_1, C_2, \dots)$. Since $E[W_n] = 0$ for all $n \geq 0$, it follows that

$$E[W_n^2] = \rho_2 E[W_{n-1}^2] = \rho_2^n E[W_0^2] = \mathrm{Var}(Q)\rho_2^n.$$

The two results now follow from the same arguments used above with $H_2 = \mathrm{Var}(Q)$ and $c_2 = \rho_2$. □

It follows from the previous result that under the conditions of Proposition 1, $R^{(k)}$ converges to $R$ both almost surely and in $L^\beta$-norm. Similarly, if we ignore the $Q$ in the generic branching vector, assume that $C_i \geq 0$ for all $i$, and define the process

$$W^{(k)} = \sum_{\mathbf{i}\in A_k} \Pi_{\mathbf{i}} = \sum_{r=1}^{N_\emptyset} C_r \left( \sum_{(r,\mathbf{i})\in A_k} \Pi_{(r,\mathbf{i})}/C_r \right) \stackrel{\mathscr{D}}{=} \sum_{r=1}^{N} C_r W_r^{(k-1)},$$

where the $\{W_r^{(k-1)}\}_{r\geq 1}$ are i.i.d. copies of $W^{(k-1)}$ independent of $(N, C_1, C_2, \dots)$, then it can be shown that $\{W^{(k)}/\rho_1^k : k \geq 0\}$ defines a nonnegative martingale which converges almost surely to the endogenous solution of the stochastic fixed-point equation

$$W \stackrel{\mathscr{D}}{=} \sum_{i=1}^{N} \frac{C_i}{\rho_1} W_i,$$

where the $\{W_i\}_{i\geq 1}$ are i.i.d. copies of $W$, independent of $(N, C_1, C_2, \dots)$. We refer to this equation as the homogeneous case.

As mentioned in the introduction, our objective is to generate a sample of $R^{(k)}$ for values of $k$ sufficiently large to suitably approximate $R$. Our proposed algorithm can also be used to simulate $W^{(k)}$, but due to space limitations we will omit the details.

## 3 THE ALGORITHM

Note that based on (5), one way to simulate $R^{(k)}$ would be to simulate a weighted branching process starting from the root and up to the $k$ generation and then add all the weights $Q_{\mathbf{i}}\Pi_{\mathbf{i}}$ for $\mathbf{i} \in \bigcup_{j=0}^{k} A_j$. Alternatively, we could generate a large enough pool of i.i.d. copies of $Q$ which would represent the $Q_{\mathbf{i}}$ for $\mathbf{i} \in A_k$, and use them to generate a pool of i.i.d. observations of $R^{(1)}$ by setting

$$R_i^{(1)} = \sum_{r=1}^{N_i} C_{(i,r)} R_r^{(0)} + Q_i,$$

where $\{(Q_i, N_i, C_{(i,1)}, C_{(i,2)}, \dots)\}_{i\geq 1}$ are i.i.d. copies of the generic branching vector, independent of everything else, and the $R_r^{(0)}$ are the $Q$'s generated in the previous step. We can continue this process until we get to the root node. On average, we would need $(E[N])^k$ i.i.d. copies of $Q$ for the first pool of observations, $(E[N])^{k-1}$ copies of the generic branching vector for the second pool, and in general, $(E[N])^{k-j}$ for the $j$th step. This approach is equivalent to simulating the weighted branching process starting from the $k$th generation and going up to the root, and is the result of iterating (3).

Our proposed algorithm is based on this "leaves to root" approach, but to avoid the need for a geometric number of "leaves", we will resample from the initial pool to obtain a pool of the same size of observations of $R^{(1)}$. In general, for the $j$th generation we will sample from the pool obtained in the previous step of (approximate) observations of $R^{(j-1)}$ to obtain conditionally independent (approximate) copies of $R^{(j)}$. In other words, to obtain a pool of approximate copies of $R^{(j)}$ we bootstrap from the pool previously obtained of approximate copies of $R^{(j-1)}$. The approximation lies in the fact that we are not sampling from

$R^{(j-1)}$ itself, but from a finite sample of conditionally independent observations that are only approximately distributed as $R^{(j-1)}$. The algorithm is described below.

Let $(Q,N,C_1,C_2,\ldots)$ denote the generic branching vector defining the weighted branching process. Let $k$ be the depth of the recursion that we want to simulate, i.e., the algorithm will produce a sample of random variables approximately distributed as $R^{(k)}$. Choose $m \in \mathbb{N}_+$ to be the bootstrap sample size. For each $0 \le j \le k$, the algorithm outputs $\mathscr{P}^{(j,m)} \triangleq \left(\hat{R}_1^{(j,m)}, \hat{R}_2^{(j,m)}, \ldots, \hat{R}_m^{(j,m)}\right)$, which we refer to as the sample pool at level $j$.

1. *Initialize*: Set $j=0$. Simulate a sequence $\{Q_i\}_{i=1}^m$ of i.i.d. copies of $Q$ and let $\hat{R}_i^{(0,m)} = Q_i$ for $i=1,\ldots,m$. Output $\mathscr{P}^{(0,m)} = \left(\hat{R}_1^{(0,m)}, \hat{R}_2^{(0,m)}, \ldots, \hat{R}_m^{(0,m)}\right)$ and update $j=1$.
2. While $j \le k$:
   (a) Simulate a sequence $\{(Q_i,N_i,C_{(i,1)},C_{(i,2)},\ldots)\}_{i=1}^m$ of i.i.d. copies of the generic branching vector, independent of everything else.
   (b) Let
   $$\hat{R}_i^{(j,m)} = Q_i + \sum_{r=1}^{N_i} C_{(i,r)} \hat{R}_{(i,r)}^{(j-1,m)}, \qquad i=1,\ldots,m, \tag{8}$$
   where the $\hat{R}_{(i,r)}^{(j-1,m)}$ are sampled uniformly with replacement from the pool $\mathscr{P}^{(j-1,m)}$.
   (c) Output $\mathscr{P}^{(j,m)} = \left(\hat{R}_1^{(j,m)}, \hat{R}_2^{(j,m)}, \ldots, \hat{R}_m^{(j,m)}\right)$ and update $j=j+1$.

**Remark 2** To simulate an approximation for the endogenous solution to the maximum equation (2), given by $R = \bigvee_{j=0}^{\infty} \bigvee_{\mathbf{i} \in A_j} Q_{\mathbf{i}} \Pi_{\mathbf{i}}$, simply replace (8) with

$$\hat{R}_i^{(j,m)} = Q_i \vee \bigvee_{r=1}^{N_i} C_{(i,r)} \hat{R}_{(i,r)}^{(j-1,m)}, \qquad i=1,\ldots,m.$$

Bootstrapping refers broadly to any method that relies on random sampling with replacement (Efron and Tibshirani 1993). For example, bootstrapping can be used to estimate the variance of an estimator, by constructing samples of the estimator from a number of resamples of the original dataset with replacement. With the same idea, our algorithm draws samples uniformly with replacement from the previous bootstrap sample pool. Therefore, the $\hat{R}_{(i,r)}^{(j-1,m)}$ on the right-hand side of (8) are only conditionally independent given $\mathscr{P}^{(j-1,m)}$. Hence, the samples in $\mathscr{P}^{(j,m)}$ are identically distributed but not independent for $j \ge 1$.

As we mentioned earlier, the distribution of the $\{\hat{R}_i^{(j,m)}\}$ in $\mathscr{P}^{(j,m)}$ are only approximately distributed as $R^{(j)}$, with the exception of the $\{\hat{R}_i^{(0,m)}\}$ which are exact. The first thing that we need to prove is that the distribution of the observations in $\mathscr{P}^{(j,m)}$ does indeed converge to that of $R^{(j)}$. Intuitively, this should be the case since the empirical distribution of the $\{\hat{R}_i^{(0,m)}\}$ is the empirical distribution of $m$ i.i.d. observations of $R^{(0)}$, and therefore should be close to the true distribution of $R^{(0)}$ for suitably large $m$. Similarly, since the $\{\hat{R}_i^{(1,m)}\}$ are constructed by sampling from the empirical distribution of $\mathscr{P}^{(0,m)}$, which is close to the true distribution of $R^{(0)}$, then their empirical distribution should be close to the empirical distribution of $R^{(1)}$, which in turn should be close to the true distribution of $R^{(1)}$. Inductively, provided the approximation is good in step $j-1$, we can expect the empirical distribution of $\mathscr{P}^{(j,m)}$ to be close to the true distribution of $R^{(j)}$. In the following section we make the mode of the convergence precise by considering the Kantorovich-Rubinstein distance between the empirical distribution of $\mathscr{P}^{(j,m)}$ and the true distribution of $R^{(j)}$.

The second technical aspect of our proposed algorithm is the lack of independence among the observations in $\mathscr{P}^{(k,m)}$, since a natural estimator for quantities of the form $E[h(R^{(k)})]$ would be to use

$$\frac{1}{m}\sum_{i=1}^{m} h(\hat{R}_i^{(k,m)}). \tag{9}$$

Hence, we also provide a result establishing the consistency of estimators of the form in (9) for a suitable family of functions $h$.

We conclude this section by pointing out that the complexity of the algorithm described above is of order $km$, while the naive Monte Carlo approach described earlier, which consists on sampling $m$ i.i.d. copies of a weighted branching process up to the $k$th generation, has order $(E[N])^k m$. This is a huge gain in efficiency.

## 4 CONVERGENCE AND CONSISTENCY

In order to show that our proposed algorithm does indeed produce observations that are approximately distributed as $R^{(k)}$ for any fixed $k$, we will show that the empirical distribution function of the observations in $\mathscr{P}^{(k,m)}$ , i.e.,

$$\hat{F}_{k,m}(x) = \frac{1}{m}\sum_{i=1}^{m} 1(\hat{R}_i^{(k,m)} \le x)$$

converges as $m \to \infty$ to the true distribution function of $R^{(k)}$, which we will denote by $F_k$. We will show this by using the Kantorovich-Rubinstein distance, which is a metric on the space of probability measures. In particular, convergence in this sense is equivalent to weak convergence plus convergence of the first absolute moments.

**Definition 1** let $M(\mu,\nu)$ denote the set of joint probability measures on $\mathbb{R}\times\mathbb{R}$ with marginals $\mu$ and $\nu$. then, the Kantorovich-Rubinstein distance between $\mu$ and $\nu$ is given by

$$d_1(\mu,\nu) = \inf_{\pi\in M(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x-y|\, d\pi(x,y).$$

We point out that $d_1$ is only strictly speaking a distance when both $\mu$ and $\nu$ have finite first absolute moments. Moreover, it is well known that

$$d_1(\mu,\nu) = \int_0^1 |F^{-1}(u) - G^{-1}(u)| du = \int_{-\infty}^{\infty} |F(x) - G(x)| dx. \tag{10}$$

where $F$ and $G$ are the cumulative distribution functions of $\mu$ and $\nu$, respectively, and $f^{-1}(t) = \inf\{x \in \mathbb{R} : f(x) \ge t\}$ denotes the pseudo-inverse of $f$. It follows that the optimal coupling of two real random variables $X$ and $Y$ is given by $(X,Y) = (F^{-1}(U), G^{-1}(U))$, where $U$ is uniformly distributed in $[0,1]$.

**Remark 3** The Kantorovich-Rubinstein distance is also known as the Wasserstein metric of order 1. In general, both the Kantorovich-Rubinstein distance and the more general Wasserstein metric of order $p$ can be defined in any metric space; we restrict our definition in this paper to the real line since that is all we need. We refer the interested reader to (Villani 2009) for more details.

With some abuse of notation, for two distribution functions $F$ and $G$ we use $d_1(F,G)$ to denote the Kantorovich-Rubinstein distance between their corresponding probability measures.

The following proposition shows that for i.i.d. samples, the expected value of the Kantorovich-Rubinstein distance between the empirical distribution function and the true distribution converges to zero.

**Proposition 4** Let $\{X_i\}_{i\ge1}$ be a sequence of i.i.d. random variables with common distribution $F$. Let $F_n$ denote the empirical distribution function of a sample of size $n$. Then, provided there exists $\alpha \in (1,2)$

such that $E\left[|X_1|^{\alpha}\right] < \infty$, we have that

$$E\left[d_1(F_n, F)\right] \leq n^{-1+1/\alpha}\left(\frac{2\alpha}{\alpha-1} + \frac{2}{2-\alpha}\right)E[|X_1|^{\alpha}].$$

Proposition 4 can be proved following the same arguments used in the proof of Theorem 2.2 in del Barrio, Giné, and Matrán (1999) by setting $M = 1$, and thus we omit it.

We now give the main theorem of the paper, which establishes the convergence of the expected Kantorovich-Rubinstein distance between $\hat{F}_{k,m}$ and $F_k$. Its proof is based on induction and the explicit representation (10). Recall that $\rho_\beta = E\left[\sum_{i=1}^{N} |C_i|^{\beta}\right]$.

**Theorem 5** Suppose that the conditions of Proposition 1 are satisfied for some $\beta > 1$. Then, for any $\alpha \in (1,2)$ with $\alpha \leq \beta$, there exists a constant $K_\alpha < \infty$ such that

$$E\left[d_1(\hat{F}_{k,m}, F_k)\right] \leq K_\alpha m^{-1+1/\alpha}\sum_{i=0}^{k}\rho_1^i. \tag{11}$$

*Proof.* By Proposition 1 there exists a constant $H_\alpha$ such that

$$H_\alpha = \sup_{k\geq 0} E\left[|R^{(k)}|^{\alpha}\right] \leq \sup_{k\geq 0}\left(E\left[|R^{(k)}|^{\beta}\right]\right)^{\alpha/\beta} < \infty.$$

Set $K_\alpha = H_\alpha\left(\frac{2\alpha}{\alpha-1} + \frac{2}{2-\alpha}\right)$. We will give a proof by induction.

For $j = 0$, we have that $\hat{F}_{0,m}(x) = \frac{1}{m}\sum_{i=1}^{m} 1(Q_i \leq x)$, where $\{Q_i\}_{i\geq 1}$ is a sequence of i.i.d. copies of $Q$. It follows that $\hat{F}_{0,m}$ is the empirical distribution function of $R^{(0)}$, and by Proposition 4 we have that

$$E\left[d_1(\hat{F}_{0,m}, F_0)\right] \leq K_\alpha m^{-1+1/\alpha}.$$

Now suppose that (11) holds for $j-1$. Let $\{U_r^i\}_{i,r\geq 1}$ be a sequence of i.i.d. Uniform$(0,1)$ random variables, independent of everything else. Let $\{(Q_i, N_i, C_{(i,1)}, C_{(i,2)}, \dots)\}_{i\geq 1}$ be a sequence of i.i.d. copies of the generic branching vector, also independent of everything else. Recall that $F_{j-1}$ is the distribution function of $R^{(j-1)}$ and define the random variables

$$\hat{R}_i^{(j,m)} = \sum_{r=1}^{N_i} C_{(i,r)}\hat{F}_{j-1,m}^{-1}(U_r^i) + Q_i \qquad \text{and} \qquad R_i^{(j)} = \sum_{r=1}^{N_i} C_{(i,r)}F_{j-1}^{-1}(U_r^i) + Q_i$$

for each $i = 1, 2, \dots, m$. Now use these random variables to define

$$\hat{F}_{j,m}(x) = \frac{1}{m}\sum_{i=1}^{m} 1(\hat{R}_i^{(j,m)} \leq x) \qquad \text{and} \qquad F_{j,m}(x) = \frac{1}{m}\sum_{i=1}^{m} 1(R_i^{(j)} \leq x).$$

Note that $F_{j,m}$ is an empirical distribution function of i.i.d. copies of $R^{(j)}$, which has been carefully coupled with the function $\hat{F}_{j,m}$ produced by the algorithm.

By the triangle inequality and Proposition 4 we have that

$$E\left[d_1(\hat{F}_{j,m}, F_j)\right] \leq E\left[d_1(\hat{F}_{j,m}, F_{j,m})\right] + E\left[d_1(F_{j,m}, F_j)\right] \leq E\left[d_1(\hat{F}_{j,m}, F_{j,m})\right] + K_\alpha m^{-1+1/\alpha}.$$

To analyze the remaining expectation note that

$$\begin{aligned} E\left[d_1(\hat{F}_{j,m}, F_{j,m})\right] &= E\left[\int_{-\infty}^{\infty}|\hat{F}_{j,m}(x) - F_{j,m}(x)|dx\right] \leq \frac{1}{m}\sum_{i=1}^{m} E\left[\int_{-\infty}^{\infty}\left|1(\hat{R}_i^{(j,m)} \leq x) - 1(R_i^{(j)} \leq x)\right|dx\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} E\left[\left|\hat{R}_i^{(j,m)} - R_i^{(j)}\right|\right] = \frac{1}{m}\sum_{i=1}^{m} E\left[\left|\sum_{r=1}^{N_i} C_{(i,r)}(\hat{F}_{j-1,m}^{-1}(U_r^i) - F_{j-1}^{-1}(U_r^i))\right|\right] \\ &\leq E\left[\sum_{r=1}^{N}|C_r|\right]E\left[d_1(\hat{F}_{j-1,m}, F_{j-1})\right], \end{aligned}$$

where in the last step we used the fact that $(N_i, C_{(i,1)}, C_{(i,2)}, \dots)$ is independent of $\left\{U_r^i\right\}_{r\geq 1}$ and of $\hat{F}_{j-1,m}$, combined with the explicit representation of the Kantorovich-Rubinstein distance given in (10). The induction hypothesis now gives

$$E\left[d_1(\hat{F}_{j,m}, F_j)\right] \leq \rho_1 E\left[d_1(\hat{F}_{j-1,m}, F_{j-1})\right] + K_\alpha m^{-1+1/\alpha} \leq K_\alpha m^{-1+1/\alpha} \rho_1 \sum_{i=0}^{j-1} \rho_1^i + K_\alpha m^{-1+1/\alpha}$$
$$= K_\alpha m^{-1+1/\alpha} \sum_{i=0}^{j} \rho_1^i.$$

This completes the proof. □

Note that the proof of Theorem 5 implies that $\hat{R}_i^{(j,m)} \to R_i^{(j)} = \sum_{r=1}^{N_i} C_{(i,r)} F_{j-1}^{-1}(U_r^i) + Q_i \stackrel{\mathscr{D}}{=} R^{(j)}$ in $L^1$-norm for all fixed $j \in \mathbb{N}$, and hence in distribution. In other words,

$$P\left(\hat{R}_i^{(k,m)} \leq x\right) \to F_k(x) \qquad \text{as } m \to \infty, \tag{12}$$

for all $i = 1, 2, \dots, m$, and for any continuity point of $F_k$. This also implies that

$$E\left[\hat{F}_{k,m}(x)\right] = P\left(\hat{R}_1^{(k,m)} \leq x\right) \to F_k(x) \qquad \text{as } m \to \infty, \tag{13}$$

for all continuity points of $F_k$.

Since our algorithm produces a pool $\mathscr{P}^{(k,m)}$ of $m$ random variables approximately distributed according to $F_k$, it makes sense to use it for estimating expectations related to $R^{(k)}$. In particular, we are interested in estimators of the form in (9). The problem with this kind of estimators is that the random variables in $\mathscr{P}^{(k,m)}$ are only conditionally independent given $\hat{F}_{k-1,m}$.

**Definition 2** We say that $\Theta_n$ is a consistent estimator for $\theta$ if $\Theta_n \xrightarrow{P} \theta$ as $n \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability.

Our second theorem shows the consistency of estimators of the form in (9) for a broad class of functions.

**Theorem 6** Suppose that the conditions of Proposition 1 are satisfied for some $\beta > 1$. Suppose $h : \mathbb{R} \to \mathbb{R}$ is continuous and $|h(x)| \leq C(1 + |x|)$ for all $x \in \mathbb{R}$ and some constant $C > 0$. Then, the estimator

$$\frac{1}{m} \sum_{i=1}^{m} h(\hat{R}_i^{(k,m)}) = \int_{\mathbb{R}} h(x) d\hat{F}_{k,m}(x),$$

where $\mathscr{P}^{(k,m)} = \left(\hat{R}_1^{(k,m)}, \hat{R}_2^{(k,m)}, \dots, \hat{R}_m^{(k,m)}\right)$, is a consistent estimator for $E[h(R^{(k)})]$.

*Proof.* For any $M > 0$, define $h_M(x)$ as

$$h_M(x) = h(-M) 1(x \leq -M) + h(x) 1(-M < x \leq M) + h(M) 1(x > M),$$

and note that $h_M$ is uniformly continuous. We then have

$$\left| \int_{\mathbb{R}} h(x) d\hat{F}_{k,m}(x) - \int_{\mathbb{R}} h(x) dF_k(x) \right| \leq 2C \int_{|x|>M} (1 + |x|) dF_k(x) + 2C \int_{|x|>M} (1 + |x|) d\hat{F}_{k,m}(x)$$
$$+ \left| \int_{\mathbb{R}} h_M(x) d\hat{F}_{k,m}(x) - \int_{\mathbb{R}} h_M(x) dF_k(x) \right|. \tag{14}$$

Fix $\varepsilon > 0$ and choose $M_\varepsilon > 0$ such that $E\left[(|R^{(k)}|+1)1(|R^{(k)}| > M_\varepsilon)\right] \leq \varepsilon/(4C)$ and such that $-M_\varepsilon$ and $M_\varepsilon$ are continuity points of $F_k$. Define $(\hat{R}^{(k,m)}, R^{(k)}) = (\hat{F}_{k,m}^{-1}(U), F_k^{-1}(U))$, where $U$ is a uniform $[0,1]$ random variable independent of $\mathscr{P}^{(k,m)}$. Next, note that $g(x) = 1+|x|$ is Lipschitz continuous with Lipschitz constant one and therefore

$$\begin{aligned}
\int_{|x|>M_\varepsilon} (1+|x|)d\hat{F}_{k,m}(x) &= (1+M_\varepsilon)\left(\hat{F}_{k,m}(-M_\varepsilon)+1-\hat{F}_{k,m}(M_\varepsilon)\right) + \int_{x<-M_\varepsilon} \hat{F}_{k,m}(x)\,dx + \int_{x>M_\varepsilon}(1-\hat{F}_{k,m}(x))dx \\
&\leq (1+M_\varepsilon)\left(\hat{F}_{k,m}(-M_\varepsilon)+1-\hat{F}_{k,m}(M_\varepsilon)\right) + d_1(\hat{F}_{k,m}, F_k) \\
&\quad + \int_{x<-M_\varepsilon} F_k(x)\,dx + \int_{x>M_\varepsilon}(1-F_k(x))dx \\
&= (1+M_\varepsilon)\left(\hat{F}_{k,m}(-M_\varepsilon) - F_k(-M_\varepsilon) + F_k(M_\varepsilon) - \hat{F}_{k,m}(M_\varepsilon)\right) + d_1(\hat{F}_{k,m}, F_k) \\
&\quad + E\left[(|R^{(k)}|+1)1(|R^{(k)}| > M_\varepsilon)\right].
\end{aligned}$$

Finally, since $h_{M_\varepsilon}$ is bounded and uniformly continuous, then $\omega(\delta) = \sup\{|h_{M_\varepsilon}(x) - h_{M_\varepsilon}(y)| : |x-y| \leq \delta\}$ converges to zero as $\delta \to 0$. Hence, for any $\gamma > 0$,

$$\begin{aligned}
\left|\int_{\mathbb{R}} h_{M_\varepsilon}(x)d\hat{F}_{k,m}(x) - \int_{\mathbb{R}} h_{M_\varepsilon}(x)dF_k(x)\right| &\leq E\left[\left|h_{M_\varepsilon}(\hat{R}^{(k,m)}) - h_{M_\varepsilon}(R^{(k)})\right| \Big| \hat{F}_{k,m}\right] \\
&\leq \omega(m^{-\gamma}) + K_\varepsilon E\left[1\left(|\hat{R}^{(k,m)} - R^{(k)}| > m^{-\gamma}\right) \Big| \hat{F}_{k,m}\right] \\
&\leq \omega(m^{-\gamma}) + K_\varepsilon m^\gamma d_1(\hat{F}_{k,m}, F_k),
\end{aligned}$$

where $2K_\varepsilon = \sup\{|h_{M_\varepsilon}(x)| : x \in \mathbb{R}\}$. Choose $0 < \gamma < 1 - 1/\alpha$ for the $\alpha \in (1,2)$ in Theorem 5 and combine the previous estimates to obtain

$$\begin{aligned}
E\left[\left|\int_{\mathbb{R}} h(x)d\hat{F}_{k,m}(dx) - \int_{\mathbb{R}} h(x)dF_k(dx)\right|\right] &\leq 2C(1+M_\varepsilon)\left(E[\hat{F}_{k,m}(-M_\varepsilon)] - F_k(-M_\varepsilon) + F_k(M_\varepsilon) - E[\hat{F}_{k,m}(M_\varepsilon)]\right) \\
&\quad + \varepsilon + \omega(m^{-\gamma}) + (2C + K_\varepsilon m^\gamma) E\left[d_1(\hat{F}_{k,m}, F_k)\right].
\end{aligned}$$

Since $E[\hat{F}_{k,m}(-M_\varepsilon)] \to F_k(-M_\varepsilon)$ and $E[\hat{F}_{k,m}(M_\varepsilon)] \to F_k(M_\varepsilon)$ by (13), and $m^\gamma E\left[d_1(\hat{F}_{k,m}, F_k)\right] \to 0$ by Theorem 5, it follows that

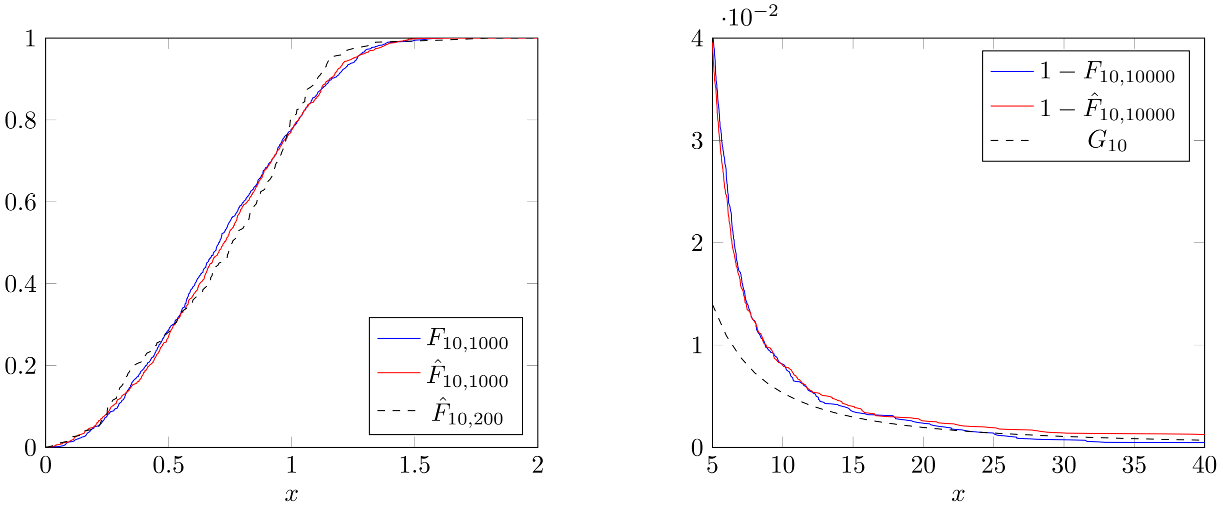
$$\limsup_{m\to\infty} E\left[\left|\int_{\mathbb{R}} h(x)d\hat{F}_{k,m}(dx) - \int_{\mathbb{R}} h(x)dF_k(dx)\right|\right] \leq \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, the convergence in $L^1$, and therefore in probability, follows. □

## 5 NUMERICAL EXAMPLES

This last section of the paper gives a numerical example to illustrate the performance of our algorithm. Consider a generic branching vector $(Q, N, C_1, C_2, \dots)$ where the $\{C_i\}_{i\geq 1}$ are i.i.d. and independent of $N$ and $Q$, with $N$ also independent of $Q$.

Figure 2a plots the empirical cumulative distribution function of 1000 samples of $R^{(10}$, i.e., $F_{10,1000}$ in our notation, versus the functions $\hat{F}_{10,200}$ and $\hat{F}_{10,1000}$ produced by our algorithm, for the case where the $C_i$ are uniformly distributed in $[0, 0.2]$, $Q$ uniformly distributed in $[0,1]$ and $N$ is a Poisson random variable with mean 3. Note that we cannot compare our results with the true distribution $F_{10}$ since it is not available in closed form. Computing $F_{10,1000}$ required 883.3 seconds using Python with an Intel i7-4700MQ 2.40 GHz processor and 8 GB of memory, while computing $\hat{F}_{10,1000}$ required only 2.1 seconds. We point out



(a) The functions $F_{10,1000}(x)$, $\hat{F}_{10,200}(x)$ and $\hat{F}_{10,1000}(x)$.

(b) The functions $1 - F_{10,10000}(x)$, $1 - \hat{F}_{10,10000}(x)$ and $G_{10}(x)$, where $G_{10}$ is evaluated only at integer values of $x$ and linearly interpolated in between.

Figure 2: Numerical examples.

that in applications to information ranking algorithms $E[N]$ can be in the thirties range, which would make the difference in computation time even more impressive.

Our second example plots the tail distribution of the empirical cumulative distribution function of $R^{(10)}$ for 10,000 samples versus the tail of $\hat{F}_{10,10000}$ for an example where $N$ is a zeta random varialbe with a probability mass function $P(N = k) \propto k^{-2.5}$, $Q$ is an exponential random variable with mean 1, and the $C_i$ have a uniform distribution in $[0, 0.5]$. In this case the exact asymptotics for $P(R^{(k)} > x)$ as $x \to \infty$ are given by

$$P(R^{(k)} > x) \sim \frac{(E[C_1]E[Q])^\alpha}{(1-\rho_1)^\alpha} \sum_{j=0}^{k} \rho_\alpha^j (1 - \rho_1^{k-j})^\alpha P(N > x),$$

where $P(N > x) = x^{-\alpha} L(x)$ is regularly varying (see Lemma 5.1 in Jelenković and Olvera-Cravioto (2010)), which reduces for the specific distributions we have chosen to

$$G_{10}(x) \triangleq \frac{(0.25)^{2.5}}{(1-(0.49))^{2.5}} \sum_{j=0}^{10} (0.07)^j (1 - (0.49)^{10-j})^{2.5} P(N > x) = (0.365)\, P(N > x).$$

Figure 2b plots the complementary distributions of $F_{10,10000}$, $\hat{F}_{10,10000}$ and compares them to $G$. We can see that the tails of both $F_{10,10000}$ and $\hat{F}_{10,10000}$ approach the asymptotic roughly at the same time.

## REFERENCES

Aldous, D., and A. Bandyopadhyay. 2005. "A Survey of Max-Type Recursive Distributional Equation". *Annals of Applied Probability* 15 (2): 1047–1110.

Alsmeyer, G., J. Biggins, and M. Meiners. 2012. "The Functional Equation of the Smoothing Transform". *Ann. Probab.* 40 (5): 2069–2105.

Alsmeyer, G., and M. Meiners. 2012. "Fixed Points of Inhomogeneous Smoothing Transforms". *J. Differ. Equ. Appl.* 18 (8): 1287–1304.

Alsmeyer, G., and M. Meiners. 2013. "Fixed points of the smoothing transform: Two-sided solutions". *Probab. Theory Rel.* 155 (1-2): 165–199.

Biggins, J. 1998. "Lindley-type equations in the branching random walk". *Stochastic Process. Appl.* 75:105–133.

Chen, N., N. Litvak, and M. Olvera-Cravioto. 2014. "Ranking algorithms on directed configuration networks". *ArXiv:1409.7443*:1–39.

del Barrio, E., E. Giné, and C. Matrán. 1999. "Central limit theorems for the Wasserstein distance between the empirical and the true distributions". *Annals of Probability*:1009–1071.

Efron, B., and R. J. Tibshirani. 1993. *An introductin to the bootstrap*.

Fill, J., and S. Janson. 2001. "Approximating the limiting Quicksort distribution". *Random Structures Algorithms* 19 (3-4): 376–406.

Jelenković, P., and M. Olvera-Cravioto. 2010. "Information ranking and power laws on trees". *Adv. Appl. Prob.* 42 (4): 1057–1093.

Jelenković, P., and M. Olvera-Cravioto. 2012a. "Implicit Renewal Theorem for Trees with General Weights". *Stochastic Process. Appl.* 122 (9): 3209–3238.

Jelenković, P., and M. Olvera-Cravioto. 2012b. "Implicit Renewal Theory and Power Tails on Trees". *Adv. Appl. Prob.* 44 (2): 528–561.

Jelenković, P., and M. Olvera-Cravioto. 2015. "Maximums on Trees". *Stochastic Process. Appl.* 125:217–232.

Karpelevich, F. I., M. Y. Kelbert, and Y. M. Suhov. 1994. "Higher-order Lindley equations". *Stochastic Processes and their Applications* 53 (1): 65–96.

Mezard, M., and A. Montanari. 2009. *Information, physics, and computation*. Oxford University Press.

Olvera-Cravioto, M. 2012. "Tail behavior of solutions of linear recursions on trees". *Stochastic Process. Appl.* 122 (4): 1777–1807.

Olvera-Cravioto, M., and O. Ruiz-Lacedelli. 2014. "Parallel queues with synchronization". *arXiv:1501.00186*.

Rösler, U. 1991. "A limit theorem for "Quicksort"". *RAIRO Theor. Inform. Appl.* 25:85–100.

Rösler, U., and L. Rüschendorf. 2001. "The contraction method for recursive algorithms". *Algorithmica* 29 (1-2): 3–33.

Villani, C. 2009. *Optimal transport, old and new*. New York: Springer.

Volkovich, Y., and N. Litvak. 2010. "Asymptotic Analysis for Personalized Web Search". *Adv. Appl. Prob.* 42 (2): 577–604.

## AUTHOR BIOGRAPHIES

**NINGYUAN CHEN** is a Ph.D. student in the IEOR Department at Columbia University. He obtained his B.S. in Mathematics from Peking University and a M.S. in Operations Research from Columbia University. His research focuses on two areas: (1) Applied Probability, in particular, random graphs and information ranking algorithms, and (2) strategic behavior of customers under market microstructure, with interfaces of dynamic pricing, stochastic modeling and optimization. His email address is nc2462@columbia.edu.

**MARIANA OLVERA-CRAVIOTO** is an Associate Professor in the IEOR Department at Columbia University. She obtained her Ph.D. in Management Science and Engineering from Stanford University and holds an M.S. in Statistics from the same university. Her research interests are in Applied Probability, in particular, the analysis of information ranking algorithms, queueing theory, random graphs, weighted branching processes and heavy-tailed asymptotics in general. She serves in the editorial board of Stochastic Models and QUESTA. Her e-mail address is molvera@ieor.columbia.edu.