Large deviations of Markov chains indexed by random trees

AMIR DEMBO,[1] *Stanford University*
PETER MÖRTERS, *University of Bath*
SCOTT SHEFFIELD, *Microsoft Research*

Abstract

Given a finite typed rooted tree T with n vertices, the *empirical subtree measure* is the uniform measure on the n typed subtrees of T formed by taking all descendants of a single vertex. We prove a large deviation principle in n, with explicit rate function, for the empirical subtree measures of multitype Galton-Watson trees conditioned to have exactly n vertices. In the process, we extend the notions of shift-invariance and specific relative entropy—as typically understood for Markov fields on deterministic graphs such as $\mathbb{Z}^d$—to Markov fields on random trees. We also develop single-generation empirical measure large deviation principles for a more general class of random trees including trees sampled uniformly from the set of all trees with n vertices.

Keywords: Tree-indexed Markov chain, branching Markov chain, random tree, Galton-Watson tree, multitype Galton-Watson process, multitype Galton-Watson tree, marked tree, large deviation principle, empirical pair measure, empirical offspring measure, process level.
MSC 2000: Primary 60F10. Secondary 60J80, 05C05.

1. Introduction

The *empirical measures* of Markov fields on large, deterministic subsets Λ of $\mathbb{Z}^d$—and the limit points of these empirical measures—play a central role in statistical physics and the theory of Gibbs measures. The limit points are always shift-invariant, and the rate functions of the empirical measure large deviation principles are generally defined in terms of *specific relative entropy* or *specific free energy*, see, e.g., Chapters 14–16 of [Ge88].

When $\mathbb{Z}^d$ is replaced with a random graph, the large deviation analysis of even the simplest models—say, Ising or Potts models—becomes more difficult. How does one even define "shift-invariance," for example, when the graphs on which the models are defined are random and almost surely possess no translational symmetries? What is the most natural analogue of "specific relative entropy"? For that matter, what is the most useful definition of "empirical measure"?

The purpose of this paper is to answer the above questions for some natural random planar rooted tree models. By *planar* we mean that the offspring of each vertex are implicitly ordered—from left to right; this ordering determines an embedding of the tree in the plane.

Given a finite planar rooted tree T with n vertices with types drawn from a finite type set $\mathcal{X}$, the *empirical subtree measure* ν^T is the uniform measure on the n typed subtrees of T that are formed by taking all descendants of a single vertex of T. We will prove a large deviation principle, with an explicit rate function defined in terms of specific relative entropy on the empirical subtree measures of multitype Galton-Watson trees conditioned to have exactly n vertices.

[1]Research partially supported by NSF grant #DMS-0072331.

The rate function of this large deviation principle will be infinite on measures that lack a natural "shift-invariance" property. We show that every shift-invariant measure can be "extended backwards" to describe the "infinite past" of a sample from the tree. We may also view this *backward tree* construction as a general technique for examining the steady state of a randomly expanding system. It is on these backward tree measures that we will actually define specific relative entropy, as the conditional entropy of the offspring measure at the root *given* its infinite past.

One motivation for pursuing this problem is the study of *tree-indexed Markov chains*, defined as follows. First we sample a tree from some probability measure,and then, given this tree, we run a Markov chain on the vertices of the tree in such a way that the state of a vertex depends only on the state of its parent. The result of this two-step experiment can also be interpreted as a *typed tree*. We always look at probabilities with respect to the whole experiment, or, in the language of random environments, at the *annealed* probabilities. These tree-indexed process are a natural concept of increasing interest in probability and applications (see, e.g., [BP94], [Pe95] and [LPP95]), often as a new way of looking at existing models. Our analysis will show that large deviations results, which are well-known for classical Markov chains, can be extended to Markov chains indexed by random trees.

When we restrict our attention to a single generation of the empirical measure (the "empirical offspring measure") or to an empirical measure recording only the transitions of types along each edge (the "empirical pair measure") we will obtain a generalized large deviation principle for which the classical Markov results (as developed in, e.g., [DZ98] and the references therein) are a special case. These problems turn out to be among the rare problems for which large deviation rates can be stated completely explicitly in a closed form. Indeed, the rates we find in this setting are hardly more complicated than the rates for classical Markov chains. For example, our rate functions are simple enough to allow one to compute the pressure and related macroscopic quantities for Gibbs measures corresponding to a short-range potential with configuration space that is the set of all typed rooted trees of n vertices with types in $\mathcal{X}$. This is in sharp contrast with the large deviation principle for the distance from the root of simple random walk on supercritical Galton-Watson trees, for which no explicit rate function is known, see [DGPZ02].

In another application, from the case of binary trees and uniform distribution of types, we calculate an explicit growth rate for the total number of binary trees of size n (odd) with types in a finite alphabet $\mathcal{X}$, which have an empirical pair measure in a given set of measures. In [KM02] the analogous combinatorial formula for the number of tuples of length n with a given empirical pair measure was used to analyse the tail behaviour of Brownian intersection local times. We hope that the formulas derived here give rise to a similar analysis of the tail behaviour of integrated super-Brownian excursion, as formulas for high moments of intersection local times involve summation over large binary trees, see e.g. [LG99].

There are a number of technical issues that make the analysis of tree-indexed Markov chains more complicated than the analogous work for classical Markov chains. One arises from the fact that, for some models of Galton-Watson trees, the probability of having exactly n vertices is zero for n in an infinite subset of $\mathbb{Z}$. It is therefore necessary to restrict our attention to those n for which the probability is positive and to prove lower bounds on probabilities that apply only for select values of n. Another arises from the possibility of an unbounded number of offspring at a single step, which necessitates the use of a technical "mass exchange" argument in Lemma 3.6.

The precise statements of our results are given in Section 2 beginning with empirical pair and empirical offspring measures and then progressing to the empirical subtree measures. The former results will apply to a larger class of random trees than the latter, which will only be proved for bounded-offspring multitype Galton-Watson trees. The proofs of all of these results are then given in Section 3.

2. Statement of the results

2.1 Large deviations for tree-indexed Markov chains

We start by looking at the situation where the tree is generated independently of the types, in which case our rate functions are particularly simple.

By $\mathcal{T}$ we denote the set of all finite rooted planar trees T, by $V = V(T)$ the set of all vertices and by $E = E(T)$ the set of all edges oriented away from the root, which is always denoted by ρ. We write $|T|$ for the number of vertices in the tree T, with the k-th generation of T being the subset of vertices of T of distance k from its root and the height of T is the largest k such that the k-th generation of T is non-empty.

Suppose that T is any finite tree and we are given an initial probability measure μ on a finite alphabet $\mathcal{X}$ and a Markovian transition kernel $Q\colon \mathcal{X} \times \mathcal{X} \geq 0$. We can obtain a *tree indexed Markov chain* $X : V \to \mathcal{X}$ by choosing $X(\rho)$ according to μ and choosing $X(v)$, for each vertex $v \neq \rho$, using the transition kernel given the value of its parent, independently of everything else. If the tree is chosen randomly, we always consider $X = \{X(v) : v \in T\}$ under the *joint law* of tree and chain. It is sometimes convenient to interpret X as a *typed tree*, considering $X(v)$ as the *type* of the vertex v.

For the generation of the trees themselves we look at the class of *simply generated trees*. These are obtained by conditioning a critical Galton-Watson on its total number of vertices. More precisely, we look at the class of Galton-Watson trees, where the number of children $N(v)$ of each $v \in T$ is an independent random variable, with the same law $p(\,\cdot\,) = \mathbb{P}\{N(v) = \,\cdot\,\}$ for all $v \in T$, such that $0 < p(0) < 1$. We assume that p is critical, i.e. the mean offspring number $\sum_{\ell=0}^{\infty} \ell p(\ell)$ is one, but this assumption is not restrictive: Note that the distribution of T conditioned on $\{|T| = n\}$ is exactly the same as when the offspring law is $p_\theta(\ell) = p(\ell)e^{\theta\ell}/\sum_j p(j)e^{\theta j}$, regardless of the value of $\theta \in \mathbb{R}$. With $0 < p(0) < 1 - p(1)$ there exists a unique θ_* such that $\sum_\ell \ell p_{\theta_*}(\ell) = 1$. Hence all our results hold in the noncritical cases with p_{θ_*} in place of p.

We allow offspring laws p with unbounded support, but we need to restrict attention to those laws, which have superexponential decay at infinity, i.e. we require that $\ell^{-1} \log p(\ell) \to -\infty$. We also allow that the support of p is a proper subset of $\mathbb{N}$, and assume throughout the paper that all statements conditioned on the event $\{|T| = n\}$ are made only for those values of n where this event has positive probability. In particular, our large deviation approximation of probabilities hold for those values of n where $\mathbb{P}\{|T| = n\} > 0$. For the general structure of the set of admissible values, see the proof of Lemma 3.1.

With each finite tree and sample X we associate a probability measure on $\mathcal{X} \times \mathcal{X}$, the *empirical pair measure* L_X, by

$$L_X(a,b) = \frac{1}{|E|} \sum_{e \in E} \delta_{(X(e_1),X(e_2))}(a,b), \quad \text{for } a, b \in \mathcal{X}, \tag{2.1}$$

where e_1, e_2 are the beginning and end vertex of the edge $e \in E$ (so e_1 is closer to ρ than e_2). The empirical pair measure is the natural object to study in this context, as it is a sufficient statistic for the estimation of the transition kernel Q.

Our first result is a large deviation principle for L_X, conditional upon the event $\{|T| = n\}$ with n chosen such that the latter has positive probability. For its formulation recall the definition of the relative entropy $H(\cdot\,\|\,\cdot)$ from [DZ98, (2.1.5)] and Cramér's rate function, as in [DZ98, (2.1.26)],

$$I_p(x) = \sup_{\lambda \in \mathbb{R}} \Big\{ \lambda x - \log \Big[\sum_{n=0}^{\infty} p(n) e^{\lambda n} \Big] \Big\}. \tag{2.2}$$

Theorem 2.1. *Suppose that T is a Galton-Watson tree, with offspring law $p(\cdot)$ such that $0 < p(0) < 1 - p(1)$, $\sum_\ell \ell p(\ell) = 1$ and $\ell^{-1}\log p(\ell) \to -\infty$. Let X be a Markov chain indexed by T with arbitrary initial distribution and an irreducible Markovian transition kernel Q. Then, for $n \to \infty$, the empirical pair measure L_X, conditioned on $\{|T| = n\}$ satisfies a large deviation principle in the space of probability vectors on $\mathcal{X} \times \mathcal{X}$ with speed n and the convex, good rate function*

$$I(\mu) = \begin{cases} H(\mu \,\|\, \mu_1 \otimes Q) + \displaystyle\sum_{a \in \mathcal{X}} \mu_2(a)\, I_p\Big(\frac{\mu_1(a)}{\mu_2(a)}\Big) & \text{if } \mu_1 \ll \mu_2, \\ \infty & \text{otherwise,} \end{cases} \tag{2.3}$$

where μ_1 and μ_2 are the first and second marginal of μ and $\mu_1 \otimes Q(a,b) = Q\{b \,|\, a\}\mu_1(a)$.

Remark: The representation (2.3) of $I(\cdot)$ provides the interpretation of the large deviations of L_X as the result of two independent contributions: when $\mu_1 = \mu_2$ we have only the term $H(\mu \,\|\, \mu_1 \otimes Q)$ which is the rate function for the large deviation principle of empirical pair measures of the Markov chain with kernel Q, see e.g. [DZ98, Section 3.1.3]. The hard constraint of $\mu_1 = \mu_2$ of the Markov chain setting is replaced here by the additional term $\sum_a \mu_2(a) I_p(\mu_1(a)/\mu_2(a))$ which reflects the large deviations contribution due to the geometry of the tree T. In loose terms, while the event $\{L_X \approx \mu\}$ for measures μ satisfying $\mu_1 = \mu_2$ is still concentrated on the typical trees, in the case of measures with $\mu_1 \neq \mu_2$ the event is concentrated on an exponentially shrinking set of trees, with an exponential rate given by the second term. $\diamond$

Examples: The class of Galton-Watson trees conditioned on the total size appears in the combinatorial literature, see e.g. [MM78], under the name *simply generated trees* and is surveyed in [Al91]. We look at some interesting examples.

• Choose the offspring law $p(\,\cdot\,)$ such that $p(k) = 1 - p(0) = 1/k$. In this case $\mathbb{P}\{|T| = n\} > 0$ if and only if $n - 1$ is divisible by k. The law of T conditional on $\{|T| = n\}$ is exactly the same as sampling the tree uniformly from the collection of all possible k-ary trees with n vertices. We have that $I_p(x) = (x/k)\log x + (1 - x/k)\log((1 - x/k)/(1 - 1/k))$, leading to the good rate function

$$I(\mu) = \begin{cases} H(\mu \,\|\, \mu_1 \otimes Q) + \frac{k-1}{k} H\big(\frac{1}{k-1}(k\mu_2 - \mu_1) \,\|\, \mu_2\big) + \frac{1}{k} H(\mu_1 \,\|\, \mu_2) & \text{if } k\mu_2 \geq \mu_1, \\ \infty & \text{otherwise,} \end{cases} \tag{2.4}$$

for the large deviation principle of L_X.

• Choose the offspring law $p(\,\cdot\,)$ as the standard Poisson distribution, $p(\ell) = e^{-\ell}/\ell!$ for $\ell = 0, 1, 2, \ldots$. Now $\mathbb{P}\{|T| = n\} > 0$ for all $n \geq 1$ and the law of T conditioned on $\{|T| = n\}$ is that of a tree chosen uniformly from all unordered trees with n vertices. We have $I_p(x) = 1 - x + x\log x$, and get a large deviations rate of

$$I(\mu) = H(\mu \,\|\, \mu_1 \otimes Q) + H(\mu_1 \,\|\mu_2), \tag{2.5}$$

for the large deviation principle of L_X.

• Choose the offspring law $p(\,\cdot\,)$ as $p(0) = p(1) = \cdots = p(k) = 1/(k+1)$. Note that this law is only critical if $k = 2$, and recall our remark on the noncritical case preceeding the theorem. Again $\mathbb{P}\{|T| = n\} > 0$ for all $n \geq 1$, and now the law of T conditional on $\{|T| = n\}$ is the same as sampling the tree uniformly from the collection of all ordered trees with n vertices and offspring number bounded by k. $\diamond$

2.2 Large deviations for multitype Galton-Watson trees

In order to prove the results of the previous section, it is necessary to pass to a larger class of typed trees, which are also of independent interest. Indeed one can go much beyond the previous setting and consider trees and types chosen *simultaneously* according to a *multitype Galton-Watson tree*.

We write $\mathcal{X}^* = \bigcup_{n=0}^{\infty} \{n\} \times \mathcal{X}^n$ and equip it with the discrete topology. Note that the offspring of any vertex $v \in T$ is characterized by an element of $\mathcal{X}^*$ and that there is an element $(0, \emptyset)$ in $\mathcal{X}^*$ symbolizing absence of offspring. For each typed tree X and each vertex v we denote by

$$C(v) = (N(v), X_1(v), \ldots, X_{N(v)}(v)) \in \mathcal{X}^*$$

the number and types of the children of v, ordered from left to right.

We now describe the joint law of a tree T and tree-indexed chain X, which defines a multitype Galton-Watson tree. The ingredients are a probability measure μ on $\mathcal{X}$, serving as the initial distribution, and an offspring transition kernel $\mathbb{Q}$ from $\mathcal{X}$ to $\mathcal{X}^*$. We define the law $\mathbb{P}$ of a tree-indexed process X by the following rules:

- The root ρ carries a random type $X(\rho)$ chosen according to the probability measure μ on $\mathcal{X}$.
- For each vertex with type $a \in \mathcal{X}$ the offspring number and types are given independently of everything else, by the offspring law $\mathbb{Q}\{\,\cdot \mid a\}$ on $\mathcal{X}^*$. We write

$$\mathbb{Q}\{\,\cdot \mid a\} = \mathbb{Q}\{(N, X_1, \ldots, X_N) \in \cdot \mid a\},$$

  i.e. we have a random number N of offspring particles with types $X_1, \ldots, X_N$.

We assume that the exponential moments $\mathbb{Q}\{\, e^{\eta N} \mid a\} < \infty$, for all $a \in \mathcal{X}$ and $\eta > 0$. We also need a weak form of irreducibility assumption. Denote, for every $c = (n, a_1, \ldots, a_n) \in \mathcal{X}^*$ and $a \in \mathcal{X}$, the *multiplicity* of the symbol a in c by

$$m(a, c) = \sum_{i=1}^{n} 1_{\{a_i = a\}}. \tag{2.6}$$

Define the matrix A with index set $\mathcal{X} \times \mathcal{X}$ and nonnegative entries by

$$A(a, b) = \sum_{c \in \mathcal{X}^*} \mathbb{Q}\{c \mid b\} m(a, c), \text{ for } a, b \in \mathcal{X}, \tag{2.7}$$

i.e. $A(a, b)$ are the expected number of offspring of type a of a vertex of type b. With $A^*(a, b) = \sum_{k=1}^{\infty} A^k(a, b) \in [0, \infty]$ we say that the matrix A is *weakly irreducible* if $\mathcal{X}$ can be partitioned into a non empty set $\mathcal{X}_r$ of *recurrent states* and a disjoint set $\mathcal{X}_t$ of *transient states* such that

- $A^*(a, b) > 0$ whenever $b \in \mathcal{X}_r$, while
- $A^*(a, b) = 0$ whenever $b \in \mathcal{X}_t$ and either $a = b$ or $a \in \mathcal{X}_r$.

For example, any *irreducible* matrix A has A^* strictly positive, hence is also weakly irreducible with $\mathcal{X}_r = \mathcal{X}$. The multitype Galton-Watson tree is called weakly irreducible (or irreducible) if the matrix A is weakly irreducible (or irreducible, respectively) and the number $\sum_{a \in \mathcal{X}_t} m(a, c)$ of transient offspring is uniformly bounded under $\mathbb{Q}$.

Note that a weakly irreducible matrix has $A(a, b) = 0$ whenever $b \in \mathcal{X}_t$ and $a \in \mathcal{X}_r$. Moreover $\mathcal{X}_t$ may be ordered such that $A(a, b) = 0$ when $a \geq b$ are both in $\mathcal{X}_t$. Consequently, the non-zero eigenvalues of a weakly irreducible matrix A are exactly those of the irreducible matrix obtained by its restriction to $\mathcal{X}_r$. Recall that, by the Perron-Frobenius theorem, see e.g. [DZ98, Theorem 3.1.1], the largest eigenvalue of an irreducible matrix is real and positive. Obviously, the same applies to weakly irreducible matrices. The multitype Galton-Watson tree is called *critical* if this eigenvalue is 1 for the matrix A.

In our new setup the empirical pair measure L_X defined in (2.1) is *not* a sufficient statistic for the estimation of $\mathbb{Q}$ and therefore it is useful to replace the empirical pair measure by a more inclusive object. To each sample chain X we associate the *empirical offspring measure* M_X on $\mathcal{X} \times \mathcal{X}^*$, which is defined by

$$M_X(a,c) = \frac{1}{|T|} \sum_{v \in V} \delta_{(X(v),C(v))}(a,c). \tag{2.8}$$

Our second main result is a large deviation principle for M_X if X is a multitype Galton-Watson tree. For its formulation denote, for every probability measure ν on $\mathcal{X} \times \mathcal{X}^*$, by ν_1 the $\mathcal{X}$-marginal of ν. We call ν *shift-invariant* if

$$\nu_1(a) = \sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*} m(a,c)\nu(b,c) \text{ for all } a \in \mathcal{X}. \tag{2.9}$$

We denote by $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ the space of probability measures ν on $\mathcal{X} \times \mathcal{X}^*$ with $\int n\,\nu(da\,,dc) < \infty$, using the convention $c = (n, a_1, \ldots, a_n)$. We endow this space with the smallest topology which makes the functionals $\nu \mapsto \int f(b,c)\,\nu(db\,,dc)$ continuous, for $f : \mathcal{X} \times \mathcal{X}^* \to \mathbb{R}$ either bounded, or $f(b,c) = m(a,c)1_{b_0}(b)$ for some $a, b_0 \in \mathcal{X}$. Define the function J on $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ by

$$J(\nu) = \begin{cases} H(\nu \,\|\, \nu_1 \otimes \mathbb{Q}) & \text{if } \nu \text{ is shift-invariant,} \\ \infty & \text{otherwise.} \end{cases} \tag{2.10}$$

In general, the topology on $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ is stronger than the weak topology, making the function J lower semicontinuous, as shown in Lemma 3.4.

Theorem 2.2. *Suppose that X is a weakly irreducible, critical multitype Galton-Watson tree with an offspring law whose exponential moments are all finite, conditioned to have exactly n vertices. Then, for $n \to \infty$, the empirical offspring measure M_X satisfies a large deviation principle in $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ with speed n and the convex, good rate function J.*

Examples:

• The situation of Theorem 2.1 corresponds to offspring kernels $\mathbb{Q}\{\cdot\,|a\}$ choosing offspring numbers according to the law $p(\cdot)$ and then choosing the offspring types independently, according to the marginal law $Q\{\cdot\,|\,a\}$ on $\mathcal{X}$. Consequently, Theorem 2.1 follows by contraction from Theorem 2.2, see Section 3.4 for more details. As its proof reveals, Theorem 2.1 applies even when the law of offspring numbers $p(\cdot\,|\,a)$ depends on the type of the parent, provided the matrix $Q\{b\,|\,a\}\sum_\ell \ell p(\ell\,|\,a)$ is weakly irreducible, with largest eigenvalue one (then, of course, $I_{p(\cdot\,|\,a)}$ replaces I_p in (2.3)).

• For a more concrete example contained in our framework, we suppose that individuals in a population may have two genetic types, a and b. Individuals of type a (resp. b) breed offspring according to the law p_a (resp. p_b), typically of the same type, but independently, mutations occur with a small probability $p > 0$. Let n_a and n_b denote the positive mean offspring number of p_a and p_b and $\eta = n_a/n_b$ representing the genetic advantage of type a. Assuming $p_a(0) + p_b(0) > 0$ and criticality, namely $n_a n_b(1-2p)+1 = (n_a+n_b)(1-p) < 2$, in a large family of size n the probability that the ratio of the numbers of individuals of type a and b in the population is close to $x \in [0,1]$ is approximately equal to $\exp(-nI(x))$ for

$$I(x) = \inf\Big\{\frac{x}{x+1} H\big(\nu_a \,\big\|\, q_a\big) + \frac{1}{x+1} H\big(\nu_b \,\big\|\, q_b\big)\Big\}, \tag{2.11}$$

where $q_a(n,m) = p_a(n+m)\,\binom{n+m}{m}\,p^m(1-p)^n$ and $q_b(n,m) = p_b(n+m)\,\binom{n+m}{m}\,p^n(1-p)^m$ and the infimum is over all probability measures ν_a, ν_b on $\mathbb{N} \times \mathbb{N}$ satisfying

$$x = \sum_{n,m=0}^{\infty} nx\nu_a(n,m) + n\nu_b(n,m) \text{ and } 1 = \sum_{n,m=0}^{\infty} mx\nu_a(n,m) + m\nu_b(n,m). \tag{2.12}$$

This rate function is zero exactly at the typical ratio, which is given by the solution $x > 0$ of the equation $x/(1+x) = (x\eta(1-p)+p)/(x\eta+1)$. Our result gives the probability of a significant deviation from this ratio, the precise rate is depending of course on the exact offspring laws of particles of either genetic type, represented by p_a, p_b. ◇

2.3 A large-deviation principle on the process level

We conclude with the extension to a *process level* large deviation principle. This will also allow us to make the concept of *shift-invariance* more transparent.

For the rest of this section we assume that the offspring numbers generated by the kernel $\mathbb{Q}$ are uniformly bounded by some $N_0 \in \mathbb{N}$. We denote by $\bar{\mathcal{X}}$ the set of all finite or infinite rooted, planar trees such that every vertex has at most N_0 offspring, with types from the finite alphabet $\mathcal{X}$ attached to the vertices. Recall that the fact that the trees are embedded in the plane imposes an ordering (say from left to right) on the children of each vertex. The laws of multitype Galton-Watson trees are probability measures on $\bar{\mathcal{X}}$. We equip $\bar{\mathcal{X}}$ with the topology generated by the functions $f : \bar{\mathcal{X}} \to \mathbb{R}$ depending only on a finite number of generations.

If $v \in V$ is a vertex of a tree T and $X \in \bar{\mathcal{X}}$ a sample chain on this tree, we denote by X^v the sample chain obtained from the subtree of T consisting of v and all successors of v. To each *finite* sample chain X we associate a probability measure T_X on $\bar{\mathcal{X}}$, the *empirical subtree measure*, defined by

$$T_X(x) = \frac{1}{|T|} \sum_{v \in V} \delta_{X^v}(x), \text{ for } x \in \bar{\mathcal{X}}. \tag{2.13}$$

Note that for fixed X the measure T_X lives on the sample space $\bar{\mathcal{X}}$ of the actual process, which justifies the terminology of a process level large-deviation principle.

Before formulating a large deviation principle for the random variable T_X we need the concept of a *backward tree* associated to a class of multitype Galton-Watson trees. We denote by $N[k]$ the number of vertices in generation k, and in particular by $N = N[1]$ the number of children of the root in T. Suppose that μ is a probability measure on $\bar{\mathcal{X}}$ with $\int N d\mu = 1$. Then we can define a *shifted* probability measure $S(\mu)$ on $\bar{\mathcal{X}}$ by

$$S(\mu)(\Gamma) = \int d\mu(X) \sum_{i=1}^{N} 1_{\{X^{v_i} \in \Gamma\}}, \text{ for any Borel set } \Gamma \subset \bar{\mathcal{X}} \,, \tag{2.14}$$

where $v_1, \dots, v_N$ are the children of the root. We define

$$\mu \text{ is } \textit{shift-invariant} \qquad \Longleftrightarrow \qquad S(\mu) = \mu. \tag{2.15}$$

Intuitively, one obtains the shifted tree by choosing a child of the root and selecting the subtree consisting of its progeny. Clearly this operation leads to trees which are less deep than the original tree, and for shift-invariant trees this effect is compensated by weighting the trees according to the number of offspring of the root. It is important to note that shift-invariance can be defined only for *measures on trees*, not for the individual trees themselves.

To any shift-invariant measure μ on $\bar{\mathcal{X}}$ we can associate a *backward tree measure* μ^* in the following way. Suppose that X is a sample chain on a (finite or infinite) tree of height at least k, and mark a vertex in generation k of X as the *centre* of the tree. Denote by $\mathcal{X}[k]$ the set of all objects (x, ζ) (typed tree x with centre at ζ) arising in this way, endowed with the canonical topology inherited from $\bar{\mathcal{X}}$. For $k \geq l$ there are canonical projections $p_{kl} : \mathcal{X}[k] \to \mathcal{X}[l]$ obtained by keeping the same centre and removing all vertices from the tree whose last common ancestor with the centre lived before generation $k-l$. Note that the root of the projected tree $p_{kl}X$ is the ancestor of the centre in generation $k-l$, see Figure 1.

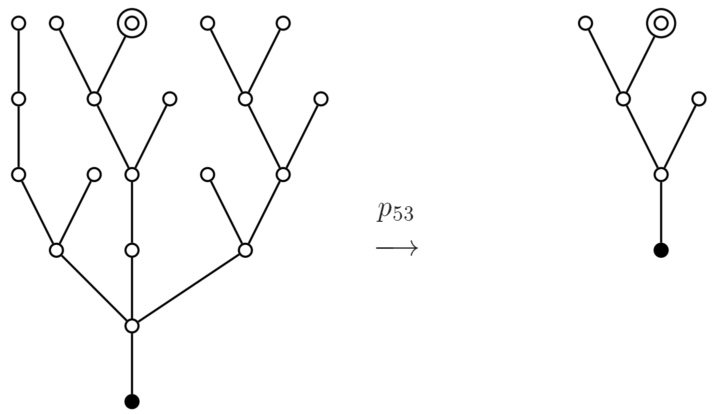

FIGURE 1. The projection p_{53}, the double circle is the centre and the filled circle the root.

The spaces $\mathcal{X}[l]$ and projections $p_{kl}, k \geq l$ form a projective system. Hence there exists a projective limit space $\underline{\mathcal{X}}$, the space of *backward trees*, and canonical projections $p_k : \underline{\mathcal{X}} \to \mathcal{X}[k]$. See [DZ98, Appendix B] for more information about projective limits.

Intuitively, the backward trees have a centre, which may have offspring itself, and an infinite ancestral line, the *spine* or *backbone*, from whose vertices offspring is emanating.

If μ is a shift-invariant measure then we can associate a measure μ_k on $\mathcal{X}[k]$ by

$$\mu_k(\Gamma) = \int d\mu(X) \sum_{i=1}^{N[k]} 1_{\{(X,v_i)\in\Gamma\}}, \text{ for any Borel set } \Gamma \subset \mathcal{X}[k], \tag{2.16}$$

where $v_1, \ldots, v_{N[k]}$ are the vertices in generation k of X.

Shift-invariance of μ ensures that all μ_k are probability measures and that $\mu_l = \mu_k \circ p_{kl}^{-1}$ for all $k \geq l$. Hence, by Kolmogorov's extension theorem, there exists a unique probability measure μ^* on $\underline{\mathcal{X}}$ such that $\mu^* \circ p_k^{-1} = \mu_k$. This is the backward tree measure μ^* associated to μ.

For each $k \geq 1$ we denote by $\mathfrak{p}_{1,k} : \mathcal{X}[k] \to \mathcal{X}[k]$ the projection obtained by removing all vertices of distance at least $k+2$ from the root and all those of distance $k+1$ from the root whose parent is to the right of the centre. Similarly, we denote by $\mathfrak{p}_{0,k} : \mathcal{X}[k] \to \mathcal{X}[k]$ the projection which in addition to all the vertices removed by $\mathfrak{p}_{1,k}$ also removes all children of the centre. Note that $p_{kl} \circ \mathfrak{p}_{0,k} = \mathfrak{p}_{0,l} \circ p_{kl}$ and $p_{kl} \circ \mathfrak{p}_{1,k} = \mathfrak{p}_{1,l} \circ p_{kl}$ for all $k \geq l$. Hence, the projective limits $\mathfrak{p}_1 : \underline{\mathcal{X}} \to \underline{\mathcal{X}}$ and $\mathfrak{p}_0 : \underline{\mathcal{X}} \to \underline{\mathcal{X}}$ of $\mathfrak{p}_{1,k}$ and $\mathfrak{p}_{0,k}$, respectively, are well defined with $p_k \circ \mathfrak{p}_0 = \mathfrak{p}_{0,k} \circ p_k$ and $p_k \circ \mathfrak{p}_1 = \mathfrak{p}_{1,k} \circ p_k$ for all $k \geq 1$.

Heuristically, $\mathfrak{p}_1$ is the projection obtained by removing all vertices of the backward tree further from the root than the centre except the children of the centre and those of the vertices to the right of the centre whose distance from the root is the same as the centre, with $\mathfrak{p}_0$ removing also the children of the centre, see Figure 2 (again trees are growing upwards, and centres are marked by a double circle).

If $\mathbb{Q}$ is an offspring transition kernel, we define $\mu^* \circ \mathfrak{p}_0^{-1} \otimes \mathbb{Q}$ as the probability measure generated by starting with a backward tree sampled according to $\mu^* \circ \mathfrak{p}_0^{-1}$ and adding independently offspring according to $\mathbb{Q}$ to the centre.

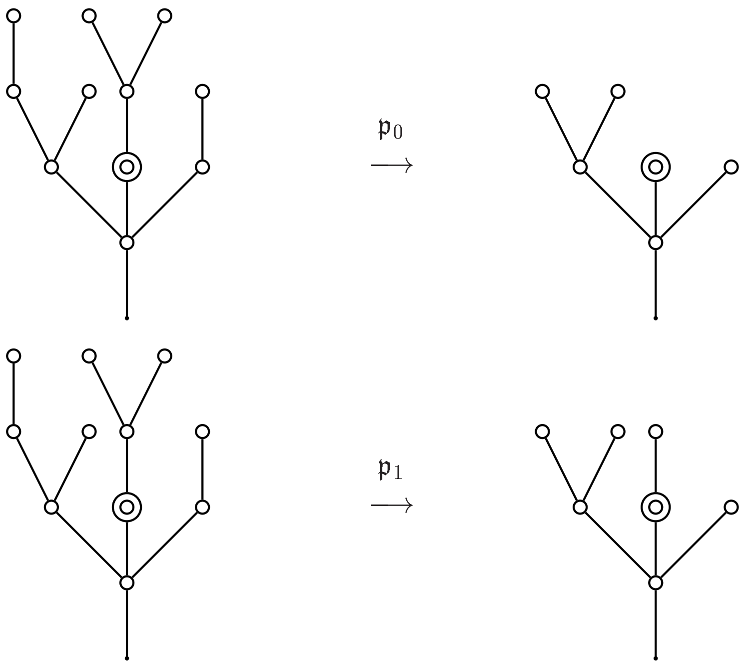

FIGURE 2. The projections $\mathfrak{p}_0$ and $\mathfrak{p}_1$.

Let $\mathcal{M}(\bar{\mathcal{X}})$ be the set of probability measures on $\bar{\mathcal{X}}$. Define the function K on $\mathcal{M}(\bar{\mathcal{X}})$ by

$$K(\mu) = \begin{cases} H\big(\mu^* \circ \mathfrak{p}_1^{-1} \,\big\|\, \mu^* \circ \mathfrak{p}_0^{-1} \otimes \mathbb{Q}\big) & \text{if } \mu \text{ is shift-invariant,} \\ \infty & \text{otherwise.} \end{cases} \tag{2.17}$$

We equip $\mathcal{M}(\bar{\mathcal{X}})$ with the smallest topology which makes the functionals $\mu \mapsto \int f \, d\mu$ continuous, for each continuous and bounded $f : \bar{\mathcal{X}} \to \mathbb{R}$.

Theorem 2.3. *Suppose that X is an irreducible, critical multitype Galton-Watson tree with uniformly bounded offspring sizes, conditioned to have exactly n vertices. Then, for $n \to \infty$, the empirical subtree measure T_X satisfies a large deviation principle in $\mathcal{M}(\bar{\mathcal{X}})$ with speed n and the convex, good rate function K.*

Example: To give an example, which is *not* covered by the previous theorems, suppose we have a multitype Galton-Watson tree with two types, say a and b, which satisfies our assumptions. Let $X_a \subseteq X$ denote the smallest subtree containing the root and all vertices of type a, and $M_a \subset \bar{\mathcal{X}}$ denote the open set of all typed trees containing at least one vertex of type a. Note that $v \in X_a$ if and only if $X^v \in M_a$. Hence, $|X_a| = nT_X(M_a)$ and by Theorem 2.3 we have, for $x \in [0,1]$, that

$$\limsup_{n\to\infty} \tfrac{1}{n} \log \mathbb{P}\big\{|X_a| \le nx \,\big|\, |T| = n\big\} \le -\inf\big\{K(\mu) \,:\, \mu \in \mathcal{M}(\bar{\mathcal{X}}),\, \mu(M_a) \le x\big\}.$$

While this rate seems difficult to evaluate more explicitly, it is easy to check that $K(\cdot)$ has a unique minimiser given by the law of the multitype Galton-Watson tree X with transition kernel $\mathbb{Q}$ and initial distribution given by the normed principal right eigenvector of the associated matrix A, recall (2.7). Denoting by p the probability that such a tree contains at least one vertex of type a, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big\{\tfrac{1}{n}|X_a| \le p - \varepsilon \,\big|\, |T| = n\big\} = 0.$$

While this may not be difficult to guess, it appears to be difficult to prove directly. $\diamond$

3. Proof of the large deviation principles

We first give a brief overview of this section, which contains the proofs of our results. First we need to establish the fact that for a critical multitype Galton-Watson tree our conditioning events $\{|T| = n\}$ decay with an exponential rate zero over the set of admissible values of n. The proof of this fact, well-known for single-type Galton-Watson trees, requires a careful analysis of the lattice structure of the set $S = \{n \in \mathbb{N} : \mathbb{P}\{|T| = n\} > 0\}$ in the multitype case, and is of some independent interest. This result is proved in Section 3.1.

Equipped with this result, in Section 3.2 the upper bound of Theorem 2.2 is derived. Exponential tightness is established in the topology on $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ using the moment conditions imposed on $\mathbb{Q}$. Based on the exponential Chebyshev inequality we first represent the upper bound in a variational form, and then solve the variational problem. Nonstandard arguments arise in the proof from the fact that we endow $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ with a topology, which is stronger than the weak topology of measures. This is necessary in order to make the set of shift-invariant measures a closed set in our topology.

The lower bound, proved in Section 3.3, is based on a change of measure technique. As we allow for potentially unbounded offspring numbers intricate approximation arguments are needed to show that this change of measure provides sufficient freedom to represent a sufficiently large class of offspring measures. This will be achieved by means of an intricate mass exchange argument, see Lemma 3.6. In Section 3.4 we prove Theorem 2.1 by contraction from Theorem 2.2.

Finally, in Section 3.5 we prove Theorem 2.3. For this purpose we first extend Theorem 2.2 from one-generation offspring measures to k-generation offspring measures, see Lemma 3.8. This extension is based on expanding the statespace and needs crucially the fact that in Theorem 2.2 we are only requiring *weak* irreducibility. The step from k-generation offspring measures to empirical subtree measures is then based on the Dawson-Gärtner Theorem.

3.1 On the rate of decay of $\mathbb{P}\{|T| = n\}$.

An important role in our proofs is played by the fact that for critical multitype Galton-Watson trees the probability $\mathbb{P}\{|T| = n\}$ decays only subexponentially on the set S of integers n where the probability is positive. We exclude the trivial case when S fails to be infinite from our consideration (in particular, we assume throughout that $\mu(\mathcal{X}_r) > 0$).

Lemma 3.1. *Suppose T is the random tree generated by a weakly irreducible, critical multitype Galton-Watson tree with finite second moment. Then*

$$\lim_{\substack{n\to\infty \\ n \in S}} \frac{1}{n} \log \mathbb{P}\{|T| = n\} = 0.$$

Proof. Recall that the number of children of any given $v \in T$ with types in $\mathcal{X}_t$ is uniformly bounded. Moreover, if $X(u) \in \mathcal{X}_t$ for some $u \in T$ then there are only types from $\mathcal{X}_t$ in the sample chain X^u consisting of u and all successors of u, and the height of the corresponding subtree T^u is uniformly bounded (by the size of $\mathcal{X}_t$). Let $G(v) = \sum_i |T^{u_i}|$ where the sum extends over the children $u_1, u_2, \ldots$ of v such that $X(u_i) \in \mathcal{X}_t$. Hence $G(v)$ is also uniformly bounded, say by $N_1 < \infty$. For $c \in \mathcal{X}^*$ let $c|\mathcal{X}_r$ be the natural restriction of c to $\mathcal{X}_r^*$. For each $b \in \mathcal{X}_r$, $c \in \mathcal{X}_r^*$ and $g \in \{0, \ldots, N_1\}$ let $\mathbb{Q}_r\{(c,g)\,|\,b\}$ denote the probability induced by $\mathbb{Q}$ that given $X(v) = b$ we have $C(v)|\mathcal{X}_r^* = c$ and $G(v) = g$. Then, for each $c_r \in \mathcal{X}_r^*$,

$$\sum_{g=0}^{N_1} \mathbb{Q}_r\{(c_r, g)\,|\,b\} = \sum_{\substack{c \in \mathcal{X}^* \\ c|\mathcal{X}_r = c_r}} \mathbb{Q}\{c\,|\,b\}\,. \tag{3.1}$$

Hence $\mathbb{Q}_r$ is a transition probability measure from $\mathcal{X}_r$ to $\mathcal{X}_r^* \times \{0,\dots,N_1\}$ such that

$$A_r(a,b) = \sum_{c\in\mathcal{X}_r^*}\sum_{g=0}^{N_1} m(a,c)\mathbb{Q}_r\{(c,g)\,|\,b\}$$

is exactly the restriction of the matrix A to $\mathcal{X}_r$. In particular, since A is weakly irreducible and critical, it follows that A_r is irreducible and critical on $\mathcal{X}_r$. Further, $\mathbb{Q}_r$ constructs the restriction of the multitype Galton-Watson tree X to $\mathcal{X}_r$ with $G(v)$ keeping track of the number of vertices with types in $\mathcal{X}_t$ that have been omitted as a result of being in T^u for some child u of v such that $X(u) \in \mathcal{X}_t$. Thus, fix a type $a \in \mathcal{X}_r$ and construct a multitype Galton-Watson tree with law $\mathbb{P}$, for $\mu = \delta_a$ as follows: Start at size $n = 0$ with one *active* vertex ρ of type a. At each future step choose an active vertex v uniformly from all active vertices, independently of everything else, provide it with offspring $C(v)$ according to $\mathbb{Q}_r\{\cdot\,|\,X(v)\}$, adding $G(v)+1$ to the current tree size n, deactivating v and activating its offspring. When there are no active vertices left, the process terminates, producing the restriction to $\mathcal{X}_r$ of a typed tree of law $\mathbb{P}$ and size n for $\mu = \delta_a$.

Let $p_{a,b}(n)$ be the probability that when the size is n we have exactly one active vertex, which is of type b. For any $a_1, a_2, a_3 \in \mathcal{X}_r$ and positive integers n_1, n_2 we have

$$p_{a_1,a_2}(n_1)p_{a_2,a_3}(n_2) \le p_{a_1,a_3}(n_1+n_2). \tag{3.2}$$

Indeed, $p_{a_1,a_2}(n_1)p_{a_2,a_3}(n_2)$ is the probability of having exactly one active vertex when the size is n_1 and again when the size is n_1+n_2, having types a_2 and a_3, respectively.

Since the restricted multitype Galton-Watson tree is irreducible, starting with $a \in \mathcal{X}_r$ active vertices of each type appear with positive probability and our procedure allows each active vertex to eventually remain the only active vertex with positive probability. Hence for any $a_1, a_2 \in \mathcal{X}_r$, there exists n such that $p_{a_1,a_2}(n) > 0$. Together with (3.2) this suffices to make the structure of the sets

$$S_{a,b} = \big\{n \in \mathbb{N} \,:\, p_{a,b}(n) > 0\big\} \tag{3.3}$$

for $a, b \in \mathcal{X}_r$, analogous to that of the sets $\big\{n \in \mathbb{N} \,:\, (P^n)_{a,b} > 0\big\}$ for a finite state irreducible Markov chain with transition matrix P. Namely, there exists a period $d =$gcd $S_{a,a}$, independent of $a \in \mathcal{X}_r$, and $k_{a,b} \in \{0,\dots,d-1\}$ such that $S_{a,b} \subset k_{a,b} + d\mathbb{N}$ with $|(k_{a,b}+d\mathbb{N}) \setminus S_{a,b}| < \infty$, see for example the proof in [Du96, Lemmas 5.5.3, 5.5.4 and 5.5.6]. Analogously to the theory of d-periodic finite state irreducible Markov chains, (3.2) and subadditivity imply the existence of $I < \infty$ such that, for all $a, b \in \mathcal{X}_r$,

$$\lim_{l\to\infty} -\frac{1}{ld}\log p_{a,b}(k_{a,b}+ld) = I. \tag{3.4}$$

(Indeed, one can take first $a = b \in \mathcal{X}_r$ showing existence of limits $I_{a,a} < \infty$, then show that $I_{a,a} \le I_{b,b}$ for all $a, b \in \mathcal{X}_r$, hence for each such a and b the limit $I_{a,b}$ exists and is equal to $I_{a,a}$ by a sandwich argument). Now let

$$p_a(n) = \mathbb{P}\big\{|T| = n\,|\,X(\rho) = a\big\}, \qquad \text{and} \qquad S_a = \{n \,:\, p_a(n) > 0\}. \tag{3.5}$$

Define

$$\mathcal{X}_g = \big\{b \,:\, \mathbb{Q}_r\{((0,\emptyset),g)\,|\,b\} > 0\big\}, \tag{3.6}$$

noting that the latter set is nonempty for some g, as otherwise no finite trees are possible. The event $\{|T| = n\}$ corresponds to *one* active vertex from $\mathcal{X}_g$ at size $n-1-g$ producing g omitted vertices of types from $\mathcal{X}_t$ and no offspring with type in $\mathcal{X}_r$. Summing over the possible types of this vertex we get

$$p_a(n) = \sum_{g=0}^{N_1}\sum_{b\in\mathcal{X}_g} p_{a,b}(n-1-g)\mathbb{Q}_r\{((0,\emptyset),g)\,|\,b\},$$

implying (recall (3.3)) that $S_a = \{n : n-1-g \in S_{a,b}$ for some $g \geq 0$ and $b \in \mathcal{X}_g\}$ and, for any $a \in \mathcal{X}_r$,

$$\lim_{\substack{n\to\infty \\ n\in S_a}} -\frac{1}{n} \log p_a(n) = I.$$

Now suppose for contradiction that $I > 0$. Then, for $a \in \mathcal{X}_r$ and all $n \in S_a$ with $n \geq n_0$, we have $p_a(n) \leq \exp(-nI/2)$. As $p_a(n) = 0$ for all $n \notin S_a$, this implies that

$$\mathbb{P}\big\{|T| \geq n \,\big|\, X(\rho) = a\big\} \leq \frac{\exp(-nI/2)}{1 - \exp(-I/2)} \text{ for all } n \geq n_0.$$

But this probability is at least as large as the corresponding probability for the restriction of T to vertices whose type is in $\mathcal{X}_r$. The latter is an irreducible, critical multitype Galton-Watson tree, so by the corollary in [AN72, p.191] under the hypothesis of finite second moment this probability is bounded below by a constant multiple of $1/n$, which is a contradiction. Hence, $I = 0$ and the result of the lemma follows since by the weak irreducibility of X we have that $p_a(n) = 0$ for all $n \geq n_0$ and $a \in \mathcal{X}_t$. $\square$

3.2 Proof of the upper bound in Theorem 2.2

Given a bounded function $\tilde{g} : \mathcal{X} \times \mathcal{X}^* \to \mathbb{R}$ we define the function

$$U_{\tilde{g}}(a) = \log \sum_{c\in\mathcal{X}^*} \mathbb{Q}\{c \,|\, a\} e^{\tilde{g}(a,c)}, \tag{3.7}$$

for $a \in \mathcal{X}$. We use $\tilde{g}$ to define a new multitype Galton-Watson tree as follows:

- The type of the root ρ is $a \in \mathcal{X}$ with probability
$$\mu_{\tilde{g}}(a) = \frac{e^{U_{\tilde{g}}(a)}\mu(a)}{\int e^{U_{\tilde{g}}(b)}\mu(db)}. \tag{3.8}$$
- for each vertex with type $a \in \mathcal{X}$ the offspring number and types are given independently of everything else, by the offspring law $\tilde{\mathbb{Q}}\{\,\cdot \,|\, a\}$ given by
$$\tilde{\mathbb{Q}}\{c \,|\, a\} = \exp\big(\tilde{g}(a,c) - U_{\tilde{g}}(a)\big)\mathbb{Q}\{c \,|\, a\}. \tag{3.9}$$

We denote the transformed law by $\tilde{\mathbb{P}}$ and make the simple observation that $\tilde{\mathbb{P}}$ is absolutely continuous with respect to $\mathbb{P}$, as for each finite $X \in \tilde{\mathcal{X}}$,

$$\frac{d\tilde{\mathbb{P}}}{d\mathbb{P}}(X) = \frac{e^{U_{\tilde{g}}(X(\rho))}}{\int e^{U_{\tilde{g}}(b)}\mu(db)} \prod_{v\in V} \exp\Big[\tilde{g}(X(v), C(v)) - U_{\tilde{g}}(X(v))\Big] \tag{3.10}$$

$$= \frac{1}{\int e^{U_{\tilde{g}}(a)}\mu(da)} \prod_{v\in V} \exp\Big[\tilde{g}(X(v), C(v)) - \sum_{j=1}^{N(v)} U_{\tilde{g}}(X_j(v))\Big], \tag{3.11}$$

recalling that $C(v) = \big(N(v), X_1(v), \ldots, X_N(v)\big)$.

We begin by establishing exponential tightness of the family of laws of M_X on the space $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$.

Lemma 3.2. *For every $\alpha > 0$ there exists a compact $K_\alpha \subset \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ with*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{M_X \notin K_\alpha \,\big|\, |T| = n\big\} \leq -\alpha.$$

Proof. Recall that $\mathbb{Q}\{e^{\eta N}\,|\,a\} < \infty$ for all $\eta > 0$. Hence, given $l \in \mathbb{N}$, we may choose $k(l) \in \mathbb{N}$ so large that

$$\mathbb{Q}\big\{\, \exp(l^2 N 1_{\{N>k(l)\}}) \,\big|\, a\big\} < 2 \text{ for all } a \in \mathcal{X}.$$

Using the exponential Chebyshev inequality,

$$\begin{aligned}\mathbb{P}\Big\{ \int_{\{N>k(l)\}} N\, dM_X \ge \tfrac{1}{l},\, |T| = n\Big\} &\le e^{-ln}\, \mathbb{E}\Big\{ \exp\big(l^2 n \int_{\{N>k(l)\}} N\, dM_X\big),\, |T| = n\Big\} \\ &= e^{-ln}\mathbb{E}\Big\{ \prod_{v\in T} \exp\big(l^2 1_{\{N(v)>k(l)\}} N(v)\big),\, |T| = n\Big\} \\ &\le e^{-ln}\Big(\sup_{a\in\mathcal{X}} \mathbb{Q}\big\{\, \exp(l^2 N 1_{\{N>k(l)\}}) \,\big|\, a\big\}\Big)^n \le e^{-n(l-\log 2)}.\end{aligned}$$

Now choose $M > \alpha + \log 2$. Define the set

$$\Gamma_M = \Big\{ \nu \in \mathcal{M}(\mathcal{X}\times\mathcal{X}^*) \,:\, \int_{\{N>k(l)\}} N\, d\nu < \tfrac{1}{l}, \text{ for all } l \ge M\Big\}.$$

As $\{N \le k(l)\} \subset \mathcal{X}\times\mathcal{X}^*$ is compact, the set Γ_M is pre-compact in the weak topology, by Prohorov's criterion. Moreover, since $m(a,c) \le N$, it is easy to see by truncation that for every weakly convergent sequence $\nu_n \to \nu$ with $\nu_n \in \Gamma_M$, we also have $\lim_{n\to\infty} \int m(a,c)\nu_n(b,dc) = \int m(a,c)\nu(b,dc)$. Hence, Γ_M is even pre-compact in the stronger topology we are using on the space $\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$. As

$$\mathbb{P}\big\{ M_X \notin \Gamma_M \,\big|\, |T| = n\big\} \le \frac{1}{\mathbb{P}\{|T| = n\}} \frac{1}{1-e^{-1}} \exp(-n(M - \log 2)),$$

we can use Lemma 3.1 to infer that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{ M_X \notin K_\alpha \,\big|\, |T| = n\big\} \le -\alpha,$$

for the closure K_α of Γ_M as required for the proof. $\square$

Next we derive an upper bound in a variational formulation. Denote by $\mathcal{C}$ the space of bounded functions on $\mathcal{X}\times\mathcal{X}^*$ and define for each $\nu \in \mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$,

$$\widehat{J}(\nu) = \sup_{g\in\mathcal{C}} \Big\{ \int \Big[g(b,c) - \sum_{j=1}^n U_g(a_j)\Big]\, \nu(db\,,dc)\Big\}, \tag{3.12}$$

where $c = (n, a_1, \dots, a_n)$.

Lemma 3.3. *For each closed set $F \subset \mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{ M_X \in F \,\big|\, |T| = n\big\} \le -\inf_{\nu\in F} \widehat{J}(\nu).$$

Proof. Fix $\tilde g \in \mathcal{C}$ bounded by some $M > 0$, then also $\int e^{U_{\tilde g}(a)}\mu(da) \le e^M$. Define $h : \mathcal{X}\times\mathcal{X}^* \to \mathbb{R}$ by $h(b,c) = \tilde g(b,c) - \sum_{i=1}^n U_{\tilde g}(a_i)$, where as usual $c = (n, a_1, \dots, a_n)$, and observe that, by (3.11),

$$\begin{aligned}e^M &\ge \tilde{\mathbb{P}}\{|T| = n\} \int e^{U_{\tilde g}(a)}\mu(da) = \mathbb{E}\Big\{ \prod_{v\in V} \exp\Big[\tilde g(X(v), C(v)) - \sum_{j=1}^{N(v)} U_{\tilde g}(X_j(v))\Big] 1_{\{|T|=n\}}\Big\} \\ &= \mathbb{E}\big\{ e^{n\langle h, M_X\rangle}\, 1_{\{|T|=n\}}\big\}.\end{aligned}$$

Together with Lemma 3.1 this shows that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{E}\Big\{ e^{n\langle h, M_X\rangle} \,\Big|\, |T| = n\Big\} \le 0. \tag{3.13}$$

In view of (3.10) the same bound (3.13) applies for $h:\mathcal{X}\times\mathcal{X}^*\to\mathbb{R}$ of the form $h(b,c)=\tilde g(b,c)-U_{\tilde g}(b)$.

Now fix $\varepsilon>0$, and let $\widehat J_\varepsilon(\nu)=\min\{\widehat J(\nu),\varepsilon^{-1}\}-\varepsilon$. Suppose first that $\nu\in F$ is shift-invariant. Then, for any $\tilde g\in\mathcal{C}$,

$$\int\sum_{j=1}^n U_{\tilde g}(a_j)\,\nu(db\,,dc)=\sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*}\sum_{a\in\mathcal{X}}m(a,c)\nu(b,c)U_{\tilde g}(a)=\sum_{a\in\mathcal{X}}U_{\tilde g}(a)\nu_1(a)=\int U_{\tilde g}(b)\,\nu_1(db). \tag{3.14}$$

Choose $\tilde g_\nu\in\mathcal{C}$ such that $h_\nu(b,c)=\tilde g_\nu(b,c)-U_{\tilde g}(b)$ satisfies

$$\langle h_\nu,\nu\rangle:=\int h_\nu(b,c)\nu(db\,,dc)=\int\Big[\tilde g_\nu(b,c)-\sum_{j=1}^n U_{\tilde g_\nu}(a_j)\Big]\,\nu(db\,,dc)\geq\widehat J_\varepsilon(\nu)\,. \tag{3.15}$$

Since h_ν is bounded, the mapping $\langle h_\nu,\cdot\rangle$ is continuous in $\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$. Hence there exists an open neighbourhood B_ν of ν such that

$$\inf_{\mu\in B_\nu}\langle h_\nu,\mu\rangle\geq\langle h_\nu,\nu\rangle-\varepsilon\geq\widehat J_\varepsilon(\nu)-\varepsilon\,.$$

Using the exponential Chebyshev inequality and the remark following (3.13) we obtain that,

$$\begin{aligned}&\limsup_{n\to\infty}\frac1n\log\mathbb{P}\big\{M_X\in B_\nu\,\big|\,|T|=n\big\}\\&\qquad\leq\limsup_{n\to\infty}\frac1n\log\mathbb{E}\big\{e^{n\langle h_\nu,M_X\rangle}\,\big|\,|T|=n\big\}-\widehat J_\varepsilon(\nu)+\varepsilon\leq-\inf_{\nu\in F}\widehat J_\varepsilon(\nu)+\varepsilon.\end{aligned} \tag{3.16}$$

Now suppose that ν fails to be shift-invariant. Assume first that there exists $a\in\mathcal{X}$ such that

$$\nu_1(a)<\sum_{(b,c)}m(a,c)\,\nu(b,c). \tag{3.17}$$

Recall that the mappings $\nu\mapsto\sum_{b,c}m(a,c)\,\nu(b,c)$ are continuous in our topology. Hence there exist $\delta>0$ and a small open neighbourhood $B_\nu\subset\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$ such that

$$\tilde\nu_1(a)<\sum_{(b,c)}m(a,c)\,\tilde\nu(b,c)-\delta,\ \text{ for all }\tilde\nu\in B_\nu. \tag{3.18}$$

Let $\tilde g\in\mathcal{C}$ be defined by $\tilde g(b,c)=-(\delta\varepsilon)^{-1}1_a(b)$ and $h(b,c)=\tilde g(b,c)-\sum_{j=1}^n U_{\tilde g}(a_j)$. Note that, by the definition (3.7), we have $U_{\tilde g}(b)=\tilde g(b,c)$ for all b and this vanishes unless $b=a$. Hence, by (3.18), for every $\tilde\nu\in B_\nu$ we have that $\int h\,d\tilde\nu>\varepsilon^{-1}$. Then, using the exponential Chebyshev inequality and (3.13),

$$\begin{aligned}&\limsup_{n\to\infty}\frac1n\log\mathbb{P}\big\{M_X\in B_\nu\,\big|\,|T|=n\big\}\\&\qquad\leq\limsup_{n\to\infty}\frac1n\log\mathbb{E}\big\{e^{n\langle h,M_X\rangle}\,\big|\,|T|=n\big\}-\varepsilon^{-1}\leq-\varepsilon^{-1}\leq-\inf_{\nu\in F}\widehat J_\varepsilon(\nu).\end{aligned} \tag{3.19}$$

In case the opposite inequality holds in (3.17) the same argument leads to (3.19) if $\tilde g$ is defined as $\tilde g(b,c)=(\delta\varepsilon)^{-1}1_a(b)$.

Now we use Lemma 3.2 to choose a compact set K_α (for $\alpha=\varepsilon^{-1}$) with

$$\limsup_{n\to\infty}\frac1n\log\mathbb{P}\big\{M_X\not\in K_\alpha\,\big|\,|T|=n\big\}\leq-\varepsilon^{-1}. \tag{3.20}$$

The set $K_\alpha \cap F$ is compact and hence it may be covered by finitely many of the sets $B_{\nu_1}, \dots, B_{\nu_m}$, with $\nu_i \in F$ for $i = 1, \dots, m$. Hence,

$$\mathbb{P}\big\{M_X \in F \,\big|\, |T| = n\big\} \le \sum_{i=1}^m \mathbb{P}\big\{M_X \in B_{\nu_i} \,\big|\, |T| = n\big\} + \mathbb{P}\big\{M_X \notin K_\alpha \,\big|\, |T| = n\big\}.$$

Using (3.16) and (3.19) we obtain, for small enough $\varepsilon > 0$, that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{M_X \in F \,\big|\, |T| = n\big\} \le \max_{i=1}^m \limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\big\{M_X \in B_{\nu_i} \,\big|\, |T| = n\big\} \le -\inf_{\nu\in F} \widehat{J}_\varepsilon(\nu) + \varepsilon.$$

Taking $\varepsilon \downarrow 0$ gives the required statement. □

We next show that the convex rate function J may replace the function $\widehat{J}$ of (3.12) in the upper bound of Lemma 3.3.

Lemma 3.4. *The function $J(\cdot)$ is convex and lower semicontinuous on $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$. Moreover, $J(\nu) \le \widehat{J}(\nu)$ for any $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$.*

Proof. We start by proving the inequality $J(\nu) \le \widehat{J}(\nu)$. To this end, suppose first that $\nu \not\ll \nu_1 \otimes \mathbb{Q}$. Then, there exists $(a', c') \in \mathcal{X} \times \mathcal{X}^*$ with $\nu(a', c') > 0$ and $\mathbb{Q}\{c' \mid a'\} = 0$. Consequently, recalling (3.7), we have $U_{\tilde{g}} = 0$ for $\tilde{g}(b,c) = R1_{(a',c')}(b,c)$ and any R. Considering such $\tilde{g}$ in (3.12) with $R \uparrow \infty$ we see that $\widehat{J}(\nu) = \infty$ in this case.

Suppose now that ν fails to be shift-invariant, in which case there exists $a \in \mathcal{X}$ such that $\nu_1(a) \ne \sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*} m(a,c)\nu(b,c)$. Choose $\tilde{g}(b,c) = R1_a(b)$, for which $U_{\tilde{g}}(b) = R1_a(b)$ and

$$\int \Big[\tilde{g}(b,c) - \sum_{j=1}^n U_{\tilde{g}}(a_j)\Big]\, \nu(db, dc) = R\Big(\nu_1(a) - \int m(a,c)\, \nu(db,dc)\Big) \longrightarrow \infty,$$

for $|R| \uparrow \infty$, with the sign of R chosen so that the right hand side is positive.

Finally suppose that ν is shift-invariant and $\nu \ll \nu_1 \otimes \mathbb{Q}$. By the variational characterisation of the relative entropy, see e.g. [DZ98, Lemma 6.2.13], the definition of U_g, Jensen's inequality, and (3.14),

$$\begin{aligned} H(\nu \,\|\, \nu_1 \otimes \mathbb{Q}) &= \sup_{g\in\mathcal{C}} \Big\{ \int g\, d\nu - \log \iint e^{g(a,c)}\, \mathbb{Q}\{dc \,|\, a\}\nu_1(da) \Big\} \\ &= \sup_{g\in\mathcal{C}} \Big\{ \int g\, d\nu - \log \int e^{U_g(a)}\, \nu_1(da) \Big\} \\ &\le \sup_{g\in\mathcal{C}} \Big\{ \int g\, d\nu - \int U_g(a)\, \nu_1(da) \Big\} = \widehat{J}(\nu). \end{aligned} \tag{3.21}$$

If $\nu, \nu' \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ are both shift-invariant then $\nu_\lambda = \lambda\nu + (1-\lambda)\nu'$ is also shift-invariant for any $0 < \lambda < 1$. Moreover, $\nu \mapsto \int m(a,c)\nu(b,dc)$ is continuous for each $a, b \in \mathcal{X}$, implying that the set $\mathcal{S} = \{\nu : \nu \text{ is shift-invariant }\}$ is convex and closed in the topology we use on $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$. Note that if $g \in \mathcal{C}$, then so is U_g and the mapping $\nu \mapsto \int g d\nu - \log \int e^{U_g(a)}\nu_1(da)$ is continuous and convex. Consequently, the identity (3.21) implies that $\nu \mapsto H(\nu \,\|\, \nu_1 \otimes \mathbb{Q})$ is lower semicontinuous and convex. For any $\alpha < \infty$, the level set $\{\nu : J(\nu) \le \alpha\}$ is the intersection of the convex, closed sets $\mathcal{S}$ and $\{\nu : H(\nu \,\|\, \nu_1 \otimes \mathbb{Q}) \le \alpha\}$. Consequently, $J(\cdot)$ is a convex rate function. □

3.3 Proof of the lower bound in Theorem 2.2

Recall the definition of the multiplicity $m(a,c)$ of the symbol a in c and of the matrix $A_{\tilde g}$ with index set $\mathcal{X} \times \mathcal{X}$ associated with the transformed multitype Galton-Watson tree,

$$A_{\tilde g}(a,b) = \sum_{c\in\mathcal{X}^*} \tilde{\mathbb{Q}}\{c\,|\,b\}m(a,c), \text{ for } a,b \in \mathcal{X}. \tag{3.22}$$

By our assumptions the matrix $A_{\tilde g}$ which has the same set of non-zero entries as A, is weakly irreducible. Recall that, by the Perron-Frobenius theorem, see e.g. [DZ98, Theorem 3.1.1], the largest eigenvalue $\varrho_{\tilde g}$ of the irreducible restriction of $A_{\tilde g}$ to $\mathcal{X}_r$ is real and positive, with strictly positive right and left eigenvectors. Since $A_{\tilde g}$ is weakly irreducible, the largest eigenvalue of $A_{\tilde g}$ is also $\varrho_{\tilde g}$. Further, recall that $A_{\tilde g}(a,b) = 0$ whenever $b \in \mathcal{X}_t$ and $a \in \mathcal{X}_r$ or $b \le a \in \mathcal{X}_t$, while $\sum_{b\in\mathcal{X}_r} A_{\tilde g}(a,b) > 0$ for any $a \in \mathcal{X}_t$. Consequently, there exists a unique right eigenvector $u_{\tilde g} \in \mathbb{R}^{\mathcal{X}}$ for the eigenvalue $\varrho_{\tilde g}$ of $A_{\tilde g}$ having strictly positive entries, which add up to one. The next lemma guides the choice of $\tilde g$ associated with a large deviations lower bound at $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ for which $J(\nu) < \infty$.

Lemma 3.5. *Suppose $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ with ν_1 strictly positive. The following statements are equivalent.*

(i) *ν is shift-invariant and $\nu \ll \nu_1 \otimes \mathbb{Q}$.*
(ii) *There exists a function $\tilde g : \mathcal{X} \times \mathcal{X}^* \to \mathbb{R}$ with $U_{\tilde g} = 0$, such that $\varrho_{\tilde g} = 1$ and the corresponding Perron-Frobenius eigenvector $u_{\tilde g}$ satisfies $\nu(a,c) = \tilde{\mathbb{Q}}\{c\,|\,a\}u_{\tilde g}(a)$, for every $(a,c) \in \mathcal{X} \times \mathcal{X}^*$.*

Moreover, if (ii) holds, then $H(\nu \,\|\, \nu_1 \otimes \mathbb{Q}) = \int \tilde g(b,c)\, \nu(db\,,dc)$.

Proof. Suppose first that ν is shift-invariant and $\nu \ll \nu_1 \otimes \mathbb{Q}$. Define $\tilde g$ by

$$\tilde g(a,c) = \log\Big(\frac{\nu(a,c)}{\nu_1(a)\mathbb{Q}\{c\,|\,a\}}\Big) \text{ when } \mathbb{Q}\{c\,|\,a\} > 0\,, \tag{3.23}$$

and otherwise $\tilde g(a,c) = 0$. Then, for all $a \in \mathcal{X}$,

$$\sum_{c\in\mathcal{X}^*} \mathbb{Q}\{c\,|\,a\}e^{\tilde g(a,c)} = 1,$$

and hence $U_{\tilde g}(a) = 0$. We infer that

$$\tilde{\mathbb{Q}}\{c\,|\,a\} = e^{\tilde g(a,c)}\mathbb{Q}\{c\,|\,a\}. \tag{3.24}$$

Using this and the definition (3.23) of $\tilde g$ we see that

$$\nu(a,c) = e^{\tilde g(a,c)}\mathbb{Q}\{c\,|\,a\}\nu_1(a) = \tilde{\mathbb{Q}}\{c\,|\,a\}\nu_1(a). \tag{3.25}$$

To identify $\varrho_{\tilde g}$, by Perron-Frobenius theorem, we only have to find the eigenvalue corresponding to a strictly positive (right) eigenvector, which turns out to be ν_1. Indeed, for all $a \in \mathcal{X}$,

$$\sum_{b\in\mathcal{X}} A_{\tilde g}(a,b)\nu_1(b) = \sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*} \tilde{\mathbb{Q}}\{c\,|\,b\}m(a,c)\nu_1(b) = \sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*} \nu(b,c)m(a,c) = \nu_1(a),$$

using the shift-invariance of ν in the final step. This shows that $\varrho_{\tilde g} = 1$ and, by uniqueness of the eigenvector, $\nu_1 = u_{\tilde g}$. Hence (ii) follows from (3.25).

Conversely, fix $\tilde g$ for which $\varrho_{\tilde g} = 1$ and (ii) holds. Summing over $c \in \mathcal{X}^*$ in (ii) we have that $\nu_1 = u_{\tilde g}$ and hence $\nu \ll \nu_1 \otimes \mathbb{Q}$. Moreover, for all $a \in \mathcal{X}$,

$$\nu_1(a) = \sum_{b\in\mathcal{X}} A_{\tilde g}(a,b)\nu_1(b) = \sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*} m(a,c)\tilde{\mathbb{Q}}\{c\,|\,b\}\nu_1(b) = \sum_{(b,c)\in\mathcal{X}\times\mathcal{X}^*} m(a,c)\nu(b,c),$$

hence ν is shift-invariant. Moreover, using $\nu(a,c) = \tilde{\mathbb{Q}}\{c \,|\, a\}\nu_1(a)$ and the definition of $\tilde{\mathbb{Q}}$, we get

$$H(\nu \,\|\, \nu_1 \otimes \mathbb{Q}) = \sum_{(a,c)\in\mathcal{X}\times\mathcal{X}^*} \nu(a,c) \log \frac{\tilde{\mathbb{Q}}\{c \,|\, a\}}{\mathbb{Q}\{c \,|\, a\}} = \int \tilde{g}(a,c)\, \nu(da, dc),$$

which completes the proof. □

The next lemma is key to the proof of the lower bound in Theorem 2.2. It allows us to focus on those shift-invariant $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ with strictly positive first marginal, for which $\tilde{g}$ of Lemma 3.5 is bounded above. If $\nu \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ and $a \in \mathcal{X}$ we write $\nu(\,\cdot \,|\, a) = \nu(\,\cdot\,, a)/\nu_1(a)$.

Lemma 3.6. *Suppose O is an open subset of $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ and $\nu \in O$ with $J(\nu) < \infty$. Then, for any $\delta > 0$, there exists $\tilde{\nu} \in O$ with $J(\tilde{\nu}) \le J(\nu) + \delta$, such that $\tilde{\nu}_1$ is strictly positive and $\tilde{\nu}(c \,|\, a) \le \mathbb{Q}\{c \,|\, a\}/y$ for some $y > 0$ and all $(a,c) \in \mathcal{X} \times \mathcal{X}^*$.*

Proof. Recall our assumption that X is weakly irreducible and critical. This implies the existence of a strictly positive probability vector u_0 on $\mathcal{X}$ such that

$$\nu^*(a,c) = \mathbb{Q}\{c \,|\, a\} u_0(a) \in \mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$$

is shift-invariant with $\nu_1^*(a) = u_0(a)$ and $J(\nu^*) = 0$. Fixing $\nu \in O$ with $J(\nu) < \infty$, we have for each $0 < \varepsilon < 1$ that $\nu_\varepsilon = (1-\varepsilon)\nu + \varepsilon\nu^*$ is shift-invariant in $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$ with $(\nu_\varepsilon)_1$ strictly positive and

$$\nu_\varepsilon(c \,|\, a) = 0 \qquad \Longleftrightarrow \qquad (a,c) \in \mathcal{X} \times \mathcal{X}^* \text{ satisfies } \mathbb{Q}\{c \,|\, a\} = 0.$$

By convexity of $J(\cdot)$ we know that $J(\nu_\varepsilon) \le (1-\varepsilon)J(\nu)$. Further, $\int f\, d\nu_\varepsilon \to \int f\, d\nu$ as $\varepsilon \downarrow 0$, for any $f : \mathcal{X} \times \mathcal{X}^* \to \mathbb{R}$ which is either bounded or satisfies $f(b,c) = m(a,c)1_{b_0}(b)$ for some $a, b_0 \in \mathcal{X}$. As O is open in $\mathcal{M}(\mathcal{X} \times \mathcal{X}^*)$, it follows that $\nu_\varepsilon \in O$ for all $\varepsilon > 0$ small enough.

In view of the above, we may and shall assume hereafter that ν_1 is strictly positive and $\nu(c \,|\, a) = 0$ exactly for those values $(a,c) \in \mathcal{X} \times \mathcal{X}^*$ where $\mathbb{Q}\{c \,|\, a\} = 0$. In particular, the matrix $A_{0,0}$ given by

$$A_{0,0}(a,b) = \sum_{c\in\mathcal{X}^*} m(a,c)\nu(c \,|\, b), \text{ for } a, b \in \mathcal{X}, \tag{3.26}$$

has nonnegative entries and is weakly irreducible. Its Perron-Frobenius eigenvalue, denoted $\varrho(A_{0,0})$, equals 1, and the corresponding right eigenvector $u_{0,0}$ equals ν_1 and hence is a strictly positive probability vector on $\mathcal{X}$. The corresponding left eigenvector $v_{0,0}$ is a probability vector which is strictly positive on $\mathcal{X}_r$. Clearly, for each $b \in \mathcal{X}_r$ there exists $c_1 = c_1(b)$ such that $\mathbb{Q}\{c_1 \,|\, b\} > 0$, hence also $\nu(c_1 \,|\, b) > 0$. Recall that for $b \in \mathcal{X}_t$ we have $\mathbb{Q}\{c \,|\, b\} > 0$ (and hence $\nu(c \,|\, b) > 0$) for only finitely many $c \in \mathcal{X}^*$. Consequently, $\nu(c \,|\, b) \le \mathbb{Q}\{c \,|\, b\}/y$ for some $y > 0$ and all $c \in \mathcal{X}^*$, $b \in \mathcal{X}_t$. The proof of the lemma is complete if the same applies for all $b \in \mathcal{X}_r$.

Assuming hereafter that this is not the case, with $\sum_{a\in\mathcal{X}_t} m(a,c)$ uniformly bounded under $\mathbb{Q}$, there must exist $b_0 \in \mathcal{X}_r$ and $c_2 = c_2(b_0) \in \mathcal{X}^*$ such that $\mathbb{Q}\{c_2 \,|\, b_0\} > 0$ (and hence also $\nu(c_2 \,|\, b_0) > 0$), with $\sum_{a\in\mathcal{X}_r} m(a,c_2)$ large enough to guarantee that $\sum_{a\in\mathcal{X}_r} v_{0,0}(a)(m(a,c_2) - m(a,c_1(b_0))) > 0$. Let $c_1(b)$ be arbitrary for $b \in \mathcal{X}_t$, and $c_2 = c_1(b)$ for all $b \ne b_0$.

Using these c_1 and c_2 we next construct probability measures $\nu_{x,y}(\,\cdot \,|\, b)$ on $\mathcal{X}^*$ for $0 < y < y_0$ and $|x| < 1/2$, such that for each $b \in \mathcal{X}$ and $c \in \mathcal{X}^*$ we have

- $\nu_{x,y}(c \,|\, b) \le \mathbb{Q}\{c \,|\, b\}/y$,
- $\nu_{x,y}(c \,|\, b) \to \nu_{0,0}(c \,|\, b) = \nu(c \,|\, b)$ as $x \to 0$ and $y \downarrow 0$,
- $\nu_{x,y}(c \,|\, b) = 0$ if and only if $\nu(c \,|\, b) = 0$.

Further,

$$\limsup_{\substack{x\to 0\\ y\downarrow 0}} H\big(\nu_{x,y}(\cdot\,|\,b)\,\big\|\,\mathbb{Q}\{\cdot\,|\,b\}\big) \le H\big(\nu_{0,0}(\,\cdot\,|\,b)\,\big\|\,\mathbb{Q}\{\cdot\,|\,b\}\big)\,, \tag{3.27}$$

and

$$A_{x,y}(a,b)=\sum_{c\in\mathcal{X}^*} m(a,c)\nu_{x,y}(c\,|\,b)\to A_{0,0}(a,b), \qquad \text{for any } a,b\in\mathcal{X}. \tag{3.28}$$

Note that $A_{x,y}(a,b)=0$ if and only if $A_{0,0}(a,b)=0$, so with $A_{0,0}$ weakly irreducible, the same applies to $A_{x,y}$. The function $f(x,y)=\varrho(A_{x,y})$ is thus continuous in this range of (x,y), as is also the strictly positive Perron-Frobenius right eigenvector $u_{x,y}$ of $A_{x,y}$, normalized to be a probability vector on $\mathcal{X}$. Our construction is such that $A_{x,0}=A_{0,0}+xB$ where $B(a,b)=\nu(c_2\,|\,b)\nu(c_1\,|\,b)(m(a,c_2)-m(a,c_1))$. Therefore, $f(x,0)$ is continuously differentiable at $x=0$ with

$$\frac{\partial f}{\partial x}(0,0)=\frac{\sum_{a,b} v_{0,0}(a)B(a,b)u_{0,0}(b)}{\sum_a v_{0,0}(a)u_{0,0}(a)}>0\,.$$

By the implicit function theorem, there exist $x(y)\to 0$ as $y\downarrow 0$ such that $f(x(y),y)=f(0,0)=1$ for all $y>0$ small enough. It follows that $\nu_{x,y}(b,c)=\nu_{x,y}(c\,|\,b)u_{x,y}(b)$ defines a shift-invariant probability measure $\nu_{x,y}\in\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$ for $x=x(y)$ and all $y>0$ small enough. Moreover,

$$\int m(a,c)\nu_{x(y),y}(b,dc)=A_{x(y),y}(a,b)u_{x(y),y}(b)\to A_{0,0}(a,b)u_{0,0}(b)=\int m(a,c)\nu(b,dc)\,,$$

for each $a,b\in\mathcal{X}$ and $y\downarrow 0$, implying the convergence of $\nu_{x(y),y}$ to ν in the topology of $\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$, and by (3.27) and shift-invariance, also

$$\begin{aligned}\limsup_{y\downarrow 0} J(\nu_{x(y),y}) &= \limsup_{y\downarrow 0}\sum_{b\in\mathcal{X}} u_{x(y),y}(b)H\big(\nu_{x(y),y}(\,\cdot\,|\,b)\,\big\|\,\mathbb{Q}\{\,\cdot\,|\,b\}\big)\\ &\le \sum_{b\in\mathcal{X}} u_{0,0}(b)H\big(\nu_{0,0}(\,\cdot\,|\,b)\,\big\|\,\mathbb{Q}\{\,\cdot\,|\,b\}\big)=J(\nu)\,,\end{aligned}$$

which completes the proof of the lemma subject to the construction of $\nu_{x,y}(\,\cdot\,|\,b)$.

We now turn to this construction. For any $|x|<1/2$ we define the probability measure

$$\nu_{x,0}(c\,|\,b)=\nu(c\,|\,b)+x\nu(c_2\,|\,b)\nu(c_1\,|\,b)(1_{\{c=c_2\}}-1_{\{c=c_1\}})\,. \tag{3.29}$$

In particular, $\nu_{x,0}(c\,|\,b)=0$ exactly where $\nu(c\,|\,b)=0$ and $A_{x,0}=A_{0,0}+xB$ as stated. Let $y_0=\mathbb{Q}\{c_2\,|\,b_0\}\min_{b\in\mathcal{X}_r}\mathbb{Q}\{c_1\,|\,b\}>0$ further reducing y_0 as needed to ensure that $\nu(c\,|\,b)\le\mathbb{Q}\{c\,|\,b\}/y_0$ for any $c\in\mathcal{X}^*$ and $b\in\mathcal{X}_t$. For any $0<y<y_0$ define the probability measures $\nu_{x,y}(\,\cdot\,|\,b\,)$ by

$$\begin{aligned}\nu_{x,y}(c\,|\,b)&=\min(\nu_{x,0}(c\,|\,b),\mathbb{Q}\{c\,|\,b\}/y) \text{ for } c\ne c_1\,,\\ \nu_{x,y}(c_1\,|\,b)&=\nu_{x,0}(c_1\,|\,b)+\sum_{c\ne c_1}(\nu(c\,|\,b)-\mathbb{Q}\{c\,|\,b\}/y)_+\,,\end{aligned} \tag{3.30}$$

with $_+$ indicating the positive part. Our choice of y_0 results in $\nu_{x,y}(\,\cdot\,|\,b)=\nu(\,\cdot\,|\,b)$ whenever $b\in\mathcal{X}_t$ and further guarantees that

$$\nu_{x,y}(c_2\,|\,b_0)=\nu_{x,0}(c_2\,|\,b_0)\le\mathbb{Q}\{c_2\,|\,b_0\}/y$$

and $\nu_{x,y}(c_1\,|\,b)\le 1\le\mathbb{Q}\{c_1\,|\,b\}/y$ for all $b\in\mathcal{X}_r$, $|x|<1/2$ and $0<y<y_0$. Hence we have as stated that $\nu_{x,y}(c\,|\,b)\le\mathbb{Q}\{c\,|\,b\}/y$ for all $c\in\mathcal{X}^*$, and $\nu_{x,y}(c\,|\,b)=0$ if and only if $\nu(c\,|\,b)=0$. Moreover, $A_{x,y}=A_{x,0}+E_y$, for

$$E_y(a,b)=\sum_{c\in\mathcal{X}^*}(m(a,c_1)-m(a,c))\big(\nu(c\,|\,b)-\mathbb{Q}\{c\,|\,b\}/y\big)_+,$$

in particular, $E_y(a,b) = 0$ for $b \in \mathcal{X}_t$. Writing $n(c) = n$ if $c \in \mathcal{X}^n$. Recall that $\sum_c n(c)\nu(c\,|\,b) = \sum_a A_{0,0}(a,b) < \infty$ for all $b \in \mathcal{X}$, so by dominated convergence

$$|E_y(a,b)| \le \sum_{c\in\mathcal{X}^*} (n(c_1)+n(c))\nu(c\,|\,b)1_{\{\nu(c\,|\,b)>\mathbb{Q}\{c\,|\,b\}/y\}} \quad \underset{y\downarrow 0}{\longrightarrow} \quad 0\,,$$

and consequently, as stated, each entry of $A_{x,y}$ is continuous in $(x,y) \in (-1/2,1/2)\times[0,y_0)$. By the same argument, $\sum_{c\ne c_1}(\nu(c\,|\,b) - \mathbb{Q}\{c\,|\,b\}/y)_+ \to 0$ as $y \downarrow 0$, implying the pointwise convergence $\nu_{x,y}(c\,|\,b) \to \nu(c\,|\,b)$ for each $(b,c) \in \mathcal{X}\times\mathcal{X}^*$. Turning to (3.27), note that it suffices to consider only $b \in \mathcal{X}_r$. Recall that for any $q > 0$ the function $z\log(z/q)$ increases in $z \in [q,1]$, and if $\nu_{x,y}(c\,|\,b) \ne \nu_{0,0}(c\,|\,b)$ and $c \ne c_1$, $c \ne c_2$, then necessarily $0 < \mathbb{Q}\{c\,|\,b\} \le \nu_{x,y}(c\,|\,b) < \nu_{0,0}(c\,|\,b) < 1$. Consequently,

$$\sum_{\substack{c\ne c_1\\ c\ne c_2}} \nu_{x,y}(c\,|\,b)\log\frac{\nu_{x,y}(c\,|\,b)}{\mathbb{Q}\{c\,|\,b\}} \le \sum_{\substack{c\ne c_1\\ c\ne c_2}} \nu_{0,0}(c\,|\,b)\log\frac{\nu_{0,0}(c\,|\,b)}{\mathbb{Q}\{c\,|\,b\}}\,,$$

yielding (3.27) since $\nu_{x,y}(c_i\,|\,b) \to \nu_{0,0}(c_i\,|\,b)$ and $\mathbb{Q}\{c_i\,|\,b\} > 0$ for $i = 1,2$ and $b \in \mathcal{X}_r$. □

Using Lemma 3.6 we now establish the lower bound in Theorem 2.2.

Lemma 3.7. *For each open set $O \subset \mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$,*

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big\{M_X \in O\,\big|\,|T| = n\big\} \ge -\inf_{\nu\in O} J(\nu).$$

Proof. Suppose that ν is an approximate minimizer on the right hand side. We can assume without loss of generality that $J(\nu) < \infty$, hence ν is shift-invariant with $\nu \ll \nu_1 \otimes \mathbb{Q}$. By Lemma 3.6 we may and shall assume in addition that ν_1 is strictly positive and the function $\tilde g$ associated to ν via (3.23) is bounded from above. Recall from Lemma 3.5 that $\varrho_{\tilde g} = 1$, and the corresponding Perron-Frobenius eigenvector $u_{\tilde g}$ satisfies

$$\nu(a,c) = \tilde{\mathbb{Q}}\{c\,|\,a\}\,u_{\tilde g}(a),\ \text{for every } (a,c) \in \mathcal{X}\times\mathcal{X}^*, \tag{3.31}$$

and further that $H(\nu\,\|\,\nu_1\otimes\mathbb{Q}) = \int \tilde g(b,c)\,\nu(db,dc)$. It thus suffices to show that

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big\{M_X \in O\,\big|\,|T| = n\big\} \ge -\int \tilde g(b,c)\,\nu(db,dc). \tag{3.32}$$

Since $\tilde g$ is bounded above, fixing $\varepsilon > 0$ we can choose an open set $\widetilde O \subset O$ such that $\nu \in \widetilde O$ and $\langle \tilde g,\mu\rangle \le \langle \tilde g,\nu\rangle + \varepsilon$ for all $\mu \in \widetilde O$. We use the transformed probability measures $\tilde{\mathbb{P}}$ and the formula (3.10) for their density, to get

$$\begin{aligned}
\mathbb{P}\big\{M_X \in O,\, |T| = n\big\} &\ge \tilde{\mathbb{E}}\Big\{\frac{d\mathbb{P}}{d\tilde{\mathbb{P}}}(T)1_{\{M_X\in\widetilde O\}}1_{\{|T|=n\}}\Big\}\\
&= \tilde{\mathbb{E}}\Big\{\prod_{v\in V}\exp\Big(-\tilde g(X(v),C(v))\Big)1_{\{M_X\in\widetilde O\}}1_{\{|T|=n\}}\Big\}\\
&\ge \exp\big(-n\langle\tilde g,\nu\rangle - n\varepsilon\big)\times\tilde{\mathbb{P}}\Big\{M_X \in \widetilde O,\, |T| = n\Big\}.
\end{aligned}$$

Dividing by $\mathbb{P}\{|T| = n\}$ and recalling Lemma 3.1 gives

$$\begin{aligned}
&\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\big\{M_X \in O\,\big|\,|T| = n\big\}\\
&\qquad \ge -n\langle\tilde g,\nu\rangle - n\varepsilon + \liminf_{n\to\infty}\frac{1}{n}\log\tilde{\mathbb{P}}\Big\{M_X \in \widetilde O\Big|\,|T| = n\Big\}.
\end{aligned}$$

The result follows once we show that

$$\limsup_{n\to\infty}\frac{1}{n}\log\tilde{\mathbb{P}}\Big\{M_X\notin\widetilde{O}\,\Big|\,|T|=n\Big\}<0. \tag{3.33}$$

We use the upper bound (but now with the law $\mathbb{P}$ replaced by $\tilde{\mathbb{P}}$) to establish (3.33). Indeed, since $\tilde{g}$ is bounded from above, we have $\tilde{\mathbb{Q}}\{e^{\eta N}\,|\,a\}<\infty$ for all $a\in\mathcal{X}$ and $\eta>0$. So, denoting

$$\tilde{J}(\nu)=\begin{cases} H(\nu\,\|\,\nu_1\otimes\tilde{\mathbb{Q}}) & \text{if } \nu \text{ is shift-invariant},\\ \infty & \text{otherwise},\end{cases} \tag{3.34}$$

the upper bound gives

$$\limsup_{n\to\infty}\frac{1}{n}\log\tilde{\mathbb{P}}\Big\{M_X\notin\widetilde{O}\,\Big|\,|T|=n\Big\}\le-\inf_{\tilde{\nu}\in\Gamma}\tilde{J}(\tilde{\nu}),$$

where $\Gamma\subset\widetilde{O}^c$ is a compact subset of $\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)$. It suffices to show that the infimum is positive. Suppose, for contradiction, that there exists a sequence $\tilde{\nu}_n$ with $\tilde{J}(\tilde{\nu}_n)\downarrow 0$. By compactness of Γ and lower semicontinuity of $\nu\mapsto\tilde{J}(\nu)$, we can extract a limit point $\tilde{\nu}\in\Gamma$ with $\tilde{J}(\tilde{\nu})=0$, and hence $\tilde{\nu}$ is shift-invariant and $H(\tilde{\nu}\,\|\,\tilde{\nu}_1\otimes\tilde{\mathbb{Q}})=0$. This implies that $\tilde{\nu}(a,c)=\tilde{\mathbb{Q}}\{c\,|\,a\}\tilde{\nu}_1(a)$, for every $(a,c)\in\mathcal{X}\times\mathcal{X}^*$. Then, using shift-invariance of $\tilde{\nu}$, for any $b\in\mathcal{X}$,

$$\sum_{(a,c)\in\mathcal{X}\times\mathcal{X}^*}\tilde{\mathbb{Q}}\{c\,|\,a\}m(b,c)\tilde{\nu}_1(a)=\sum_{(a,c)\in\mathcal{X}\times\mathcal{X}^*}\tilde{\nu}(a,c)m(b,c)=\tilde{\nu}_1(b).$$

By the uniqueness of the Perron-Frobenius eigenvector we infer that $\tilde{\nu}_1=u_{\tilde{g}}=\nu_1$ and this implies $\tilde{\nu}=\nu$, which contradicts $\tilde{\nu}\in\Gamma$. □

We complete the proof of Theorem 2.2 by noting that the rate function J has compact level sets, i.e. is a *good* rate function. This follows from abstract considerations as stated, e.g., in [DZ98, Theorem 1.2.18].

3.4 Proof of Theorem 2.1

Note that X is an irreducible, critical multitype Galton-Watson tree with offspring law

$$\mathbb{Q}\{c\,|\,b\}=p(n)\prod_{i=1}^{n}Q\{a_i\,|\,b\},\ \text{for } c=(n,a_1,\dots,a_n), \tag{3.35}$$

such that all exponential moments are finite. We derive Theorem 2.1 from Theorem 2.2 by applying the contraction principle to the continuous linear mapping $F:\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)\to\mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, defined by

$$F(\nu)(a,b)=\sum_{c\in\mathcal{X}^*}m(b,c)\nu(a,c)\ \text{for all } \nu\in\mathcal{M}(\mathcal{X}\times\mathcal{X}^*)\text{ and } a,b\in\mathcal{X}. \tag{3.36}$$

Indeed, Theorem 2.2 implies the large deviation principle for $F(M_X)$ conditioned on $\{|T|=n\}$ with the good rate function $I(\mu)=\inf\{J(\nu):F(\nu)=\mu\}$, see for example [DZ98, Theorem 4.2.1]. Convexity of I follows easily from the linearity of F and convexity of J. It is easy to see that on $\{|T|=n\}$ we have $L_X=\frac{n}{n-1}F(M_X)$. It follows that conditioned on $\{|T|=n\}$ the random variables L_X are exponentially equivalent to $F(M_X)$, hence L_X satisfy the same large deviation principle as $F(M_X)$, see [DZ98, Theorem 4.2.13]. Without loss of generality we restrict the space for the large deviation principle of L_X to the set of all probability vectors on $\mathcal{X}\times\mathcal{X}$, see [DZ98, Lemma 4.1.5(b)].

Turning to the proof of (2.3), recall that ν is shift-invariant if and only if $\sum_a F(\nu)(a,b)=\nu_1(b)$ for all $b\in\mathcal{X}$. Hence, if also $F(\nu)=\mu$, then necessarily $\nu_1=\mu_2$ and consequently,

$$I(\mu)=\inf\big\{H(\nu\,\|\nu_1\otimes\mathbb{Q})\ :\ F(\nu)=\mu,\nu_1=\mu_2\big\}\ . \tag{3.37}$$

Note that $\nu_1(a) = 0$ yields $\sum_b F(\nu)(a,b) = 0$. Hence if $\mu_1(a) > 0 = \mu_2(a)$ for some $a \in \mathcal{X}$ then $\{\nu : F(\nu) = \mu, \nu_1 = \mu_2\}$ is an empty set, and therefore $I(\mu) = \infty$. Assuming hereafter that $\mu_1 \ll \mu_2$, it is not hard to check that

$$I(\mu) = \sum_{a\in\mathcal{X}} \mu_2(a)\,\widetilde{I}\Big(\frac{\mu(a,\cdot)}{\mu_2(a)}, \mathbb{Q}\{\,\cdot\,|\,a\}\Big)\ , \tag{3.38}$$

where for $\phi : \mathcal{X} \to \mathbb{R}_+$ and $q \in \mathcal{M}(\mathcal{X}^*)$,

$$\widetilde{I}(\phi, q) = \inf\Big\{H(\widetilde{\nu}\,\|\,q)\ :\ \widetilde{\nu} \in \mathcal{M}(\mathcal{X}^*),\ \ \phi(b) = \sum_{c\in\mathcal{X}^*} m(b,c)\,\widetilde{\nu}(c) \text{ for all } b \in \mathcal{X}\Big\}\ . \tag{3.39}$$

Suppose now that $q(c) = p(n)\prod_{i=1}^n \widehat{q}(a_i)$ for all $c = (n, a_1, \ldots, a_n)$, where $\widehat{q}(\cdot)$ is a probability vector on $\mathcal{X}$ and $p(\,\cdot\,)$ a probability measure with mean one on the nonnegative integers, whose exponential moments are all finite. With $z = \sum_b \phi(b)$ we show next that,

$$\widetilde{I}(\phi, q) = zH\big(\phi/z\,\|\,\widehat{q}\big) + I_p(z)\ . \tag{3.40}$$

Once this is done, we combine (3.40) for $\widehat{q}(\cdot) = Q\{\,\cdot\,|\,a\}$ and $z = \mu_1(a)/\mu_2(a)$ with the representation (3.38) of $I(\mu)$, which directly yields the formula (2.3), thus completing the proof of the theorem.

To prove (3.40), suppose first that $z = 0$, i.e. $\phi(b) = 0$ for all $b \in \mathcal{X}$. In this case, $\widetilde{\nu}((0,\emptyset)) = 1$ is the only possible measure in (3.39), leading to $\widetilde{I}(\phi, q) = -\log q((0,\emptyset)) = -\log p(0)$, whereas it follows from (2.2) that $I_p(0) = -\log p(0)$ establishing (3.40) for such $\phi(\cdot)$. Assume hereafter that $z > 0$. Now the possible measures $\widetilde{\nu}(\cdot)$ in (3.39) are of the form $\widetilde{\nu}(c) = s(n)v_n(a_1, \ldots, a_n)$ for $c = (n, a_1, \ldots, a_n)$, with $v_0 = 1$, where $s(\cdot)$ is a probability measure on the nonnegative integers whose mean is z, and $v_n(\,\cdot\,)$, $n \geq 1$, are probability measures on $\mathcal{X}^n$ with marginals $v_{n,i}(\,\cdot\,)$ such that

$$\phi(b) = \sum_{n=1}^{\infty} s(n) \sum_{i=1}^{n} v_{n,i}(b) \qquad \text{for all } b \in \mathcal{X}\ . \tag{3.41}$$

By the assumed structure of $q(\,\cdot\,)$ we have for such $\widetilde{\nu}(\,\cdot\,)$ that

$$H(\widetilde{\nu}\,\|\,q) = \sum_{n=1}^{\infty} s(n)H(v_n\,\|\,\widehat{q}^n) + H(s\,\|\,p)\ , \tag{3.42}$$

where $\widehat{q}^n$ denotes the product measure on $\mathcal{X}^n$ with equal marginals $\widehat{q}$. Recall that

$$\sum_{n=1}^{\infty} s(n)H\big(v_n\,\|\,\widehat{q}^n\big) \geq \sum_{n=1}^{\infty} s(n)\sum_{i=1}^{n} H\big(v_{n,i}\,\|\,\widehat{q}\big) \geq zH\Big(z^{-1}\sum_{n=1}^{\infty} s(n)\sum_{i=1}^{n} v_{n,i}\,\Big\|\,\widehat{q}\Big)\ ,$$

with equality whenever $v_n = \prod_{i=1}^n v_{n,i}$ and $v_{n,i}$ are independent of n and i (see [DZ98, Lemma 7.3.25] for the first inequality, with the second inequality following by convexity of $H(\cdot\,\|\,\widehat{q})$ and the fact that $\sum_n s(n)n = z$). So, in view of (3.41),

$$H(\widetilde{\nu}\,\|\,q) \geq zH(\phi/z\,\|\,\widehat{q}) + H(s\,\|\,p)\,, \tag{3.43}$$

with equality when $v_n = (z^{-1}\phi)^n$ for all $n \geq 1$. Recall that with all exponential moments of $p(\cdot)$ finite, $I_p(z) = \inf\{H(s\,\|\,p) : s(\cdot) \text{ a probability measure on } \{0, 1, \ldots\} \text{ and } \sum_n s(n)n = z\}$ (see [DZ98, (2.1.27)] for a similar identity). Combining this with (3.43) leads to (3.40) and completes our proof.

3.5 Proof of Theorem 2.3

In the first step we extend the result of Theorem 2.2 to k-generation empirical offspring measures, for each $k \geq 2$, in case $\mathbb{Q}$ is irreducible and the offspring size is bounded by some non-random $N_0 < \infty$.

For each $k \geq 0$, let $\mathcal{X}(k)$ be the *finite* set of typed trees with height at most k and maximal degree N_0+1, equipped with the discrete topology (in particular, $\mathcal{X}(0) = \mathcal{X}$). Let $\pi_k : \bar{\mathcal{X}} \to \mathcal{X}(k)$ be the canonical projection obtained by removing all vertices in generations exceeding k and $\pi_{k,l} : \mathcal{X}(k) \to \mathcal{X}(l)$, $k \geq l$, the projections obtained by removing all vertices in generations exceeding l.

If X is a finite typed tree and v is a vertex in this tree, we denote by X^v the subtree rooted in v and let the k-generation empirical offspring measures M_X^k associated to X be defined as

$$M_X^k(b) = \frac{1}{|T|} \sum_{v \in V} \delta_{\pi_k(X^v)}(b), \text{ for all } b \in \mathcal{X}(k) \tag{3.44}$$

(for example $M_X^1(b) = M_X(a,c)$ where $b \in \mathcal{X}(1)$ has root of type a with n children of types $a_1, \ldots, a_n$ and $c = (n, a_1, \ldots, a_n)$). Given $a \in \mathcal{X}(k-1)$ and $b \in \mathcal{X}(k)$ we write $m_k(a,b)$ for the number of children v of the root in b such that $b^v = a$. A measure μ on $\mathcal{X}(k)$ is called *shift-invariant* if

$$\mu \circ \pi_{k,k-1}^{-1}(a) = \sum_{b \in \mathcal{X}(k)} m_k(a,b)\mu(b), \text{ for all } a \in \mathcal{X}(k-1). \tag{3.45}$$

We equip the space $\mathcal{M}(\mathcal{X}(k))$ of probability measures on $\mathcal{X}(k)$ with the smallest topology which makes the functionals $\mu \mapsto \int f d\mu$ continuous for each bounded $f : \mathcal{X}(k) \to \mathbb{R}$ (since the maximal degree is bounded in $\mathcal{X}(k)$, it follows that $\mu \mapsto \int m_k(a,x) d\mu(x)$ is also continuous for each $a \in \mathcal{X}(k-1)$).

Define $\mu \circ \pi_{k,k-1}^{-1} \otimes_1 \mathbb{Q}$ as the measure on $\mathcal{X}(k)$ obtained by providing children for each vertex of the $k-1$ generation, independently according to the transition mechanism $\mathbb{Q}$, and define the function

$$J_k(\mu) = \begin{cases} H\big(\mu \,\big\|\, \mu \circ \pi_{k,k-1}^{-1} \otimes_1 \mathbb{Q}\big) & \text{if } \mu \text{ is shift-invariant,} \\ \infty & \text{otherwise,} \end{cases} \tag{3.46}$$

on $\mathcal{M}(\mathcal{X}(k))$. Note that $J_1(\cdot)$ coincides with the good rate function $J(\cdot)$ of Theorem 2.2.

Lemma 3.8. *Suppose that X is an irreducible, critical multitype Galton-Watson tree with uniformly bounded offspring sizes, conditioned to have exactly n vertices. Then, for $n \to \infty$, the k-generation empirical offspring measure M_X^k satisfies a large deviation principle in $\mathcal{M}(\mathcal{X}(k))$ with speed n and convex, good rate function $J_k(\cdot)$.*

Proof. For $l \geq 0$ let $\mathcal{X}\{l\} \subset \mathcal{X}(l)$ be the support of $\pi_l(X)$ for a multitype Galton-Watson tree X corresponding to the transition mechanism $\mathbb{Q}$ starting at any strictly positive measure for $X(\rho)$. Let $\mathcal{X}_m\{l\}$ be the partition of $\mathcal{X}\{l\}$ according to the height $m = 0, 1, \ldots, l$ of the tree. Let

$$\mathcal{I} : \mathcal{X}\{k\} \to \mathcal{X}\{k-1\} \times \mathcal{X}\{k-1\}^* \text{ given by } \begin{cases} \mathcal{I}_1(b) &=& \pi_{k,k-1}(b) \in \mathcal{X}\{k-1\}, \\ \mathcal{I}_2(b) &=& (n, b^{v_1}, \ldots, b^{v_n}) \in \mathcal{X}\{k-1\}^*, \end{cases}$$

where $v_1, \ldots, v_n$ are the vertices in the first generation of $b \in \mathcal{X}\{k\}$ ordered from left to right.

To prove Lemma 3.8 we intend to apply Theorem 2.2 to a multitype Galton-Watson tree $\widetilde{X}$ on the enlarged finite type space $\mathcal{X}\{k-1\}$. We mark the objects related to this new tree by $\widetilde{\cdot}$.

The process $\widetilde{X}$ is constructed by choosing $\widetilde{X}(\rho)$ using the law of $\pi_{k-1}(X)$, and the offspring number and types of a vertex v as $\widetilde{C}(v) = \mathcal{I}_2(b)$ for the typed tree $b \in \mathcal{X}\{k\}$ obtained by providing children for each vertex in generation $k-1$ of $\widetilde{X}(v)$ independently according to the transition mechanism $\mathbb{Q}$.

With $\mathbb{Q}$ irreducible, it is easy to check that any $a \in \mathcal{X}\{k-1\}$ can be reached by finitely many steps of the transition mechanism $\widetilde{\mathbb{Q}}$ for $\widetilde{X}$ starting at any $b \in \mathcal{X}_{k-1}\{k-1\}$. Further,

$$b \in \mathcal{X}_l\{k-1\} \text{ for some } l < k-1 \qquad \Longrightarrow \qquad \operatorname{supp} \widetilde{\mathbb{Q}}\{\cdot \,|\, b\} \subset \bigcup_{n=0}^{N_0} \{n\} \times \mathcal{X}\{l-1\}^n,$$

implying that $\widetilde{A}(a,b) = 0$ whenever $a \in \mathcal{X}_m\{k-1\}$ for some $m \geq l$. Consequently, $\widetilde{\mathbb{Q}}$ is weakly irreducible on $\mathcal{X}\{k-1\}$. Let μ_0 denote the Perron-Frobenius eigenvector of the irreducible matrix

A, normalized to be a strictly positive probability vector on $\mathcal{X}$. Then, $\mu_l = \mu_{l-1} \otimes_1 \mathbb{Q}$ for $l \geq 1$ are strictly positive probability vectors on $\mathcal{X}\{l\}$, such that $\mu_l \circ \pi_{l,l-1}^{-1} = \mu_{l-1}$ for all $l \geq 1$. Moreover, with μ_0 the right eigenvector corresponding to the eigenvalue 1 of the matrix A, it follows by induction on $l \geq 1$ that μ_l are shift-invariant on $\mathcal{X}(l)$. In particular, for any $a \in \mathcal{X}\{k-1\}$,

$$\sum_{b \in \mathcal{X}\{k-1\}} \widetilde{A}(a,b)\mu_{k-1}(b) := \sum_{b \in \mathcal{X}\{k-1\}} \sum_{c \in \mathcal{X}\{k-1\}^*} \widetilde{m}(a,c)\widetilde{\mathbb{Q}}(c \,|\, b)\mu_{k-1}(b)$$
$$= \sum_{\bar{b} \in \mathcal{X}\{k\}} m_k(a,\bar{b})\, \mu_{k-1} \otimes_1 \mathbb{Q}(\bar{b}) = \mu_{k-1}(a)\,.$$

With μ_{k-1} a strictly positive right eigenvector for the eigenvalue 1 and the matrix $\widetilde{A}$, we see that $\widetilde{\mathbb{Q}}$ is also critical. Consequently, we have from Theorem 2.2 that $M_{\widetilde{X}}$ satisfy the large deviation principle in $\mathcal{M}(\mathcal{X}\{k-1\} \times \mathcal{X}\{k-1\}^*)$ with the good rate function $\widetilde{J}(\cdot)$ corresponding to $\widetilde{\mathbb{Q}}$.

For each $\nu_1 \in \mathcal{M}(\mathcal{X}\{k-1\})$ the measure $\nu_1 \circ \widetilde{\mathbb{Q}}$ is supported on the closed (finite) set $\mathcal{I}(\mathcal{X}\{k\})$. Consequently, $M_{\widetilde{X}}$ is supported on $\mathcal{I}(\mathcal{X}\{k\})$ as is any ν for which $\widetilde{J}(\nu) < \infty$, allowing us to restrict this large deviation principle to $\mathcal{M}(\mathcal{I}(\mathcal{X}\{k\}))$. Identifying $\mathcal{M}(\mathcal{I}(\mathcal{X}\{k\}))$ with $\mathcal{M}(\mathcal{X}\{k\})$ via the mapping $\mu = \nu \circ \mathcal{I}$, the law of $M_{\widetilde{X}}$ is exactly mapped to that of M_X^k. Moreover,

$$\nu \in \mathcal{M}(\mathcal{I}(\mathcal{X}\{k\})) \text{ is shift-invariant} \qquad \Longleftrightarrow \qquad \mu \text{ is shift-invariant on } \mathcal{X}(k),$$

as defined in (3.45), with $\nu_1 = \mu \circ \pi_{k,k-1}^{-1}$ and $(\nu_1 \otimes \widetilde{\mathbb{Q}}) \circ \mathcal{I} = (\mu \circ \pi_{k,k-1}^{-1}) \otimes_1 \mathbb{Q}$. This leads to the large deviation principle for M_X^k with the good rate function $J_k(\cdot)$, restricted to $\mathcal{M}(\mathcal{X}\{k\})$.

To complete the proof it suffices to check that any shift-invariant measure $\mu \in \mathcal{M}(\mathcal{X}(k))$ with

$$\mu \ll \mu \circ \pi_{k,k-1}^{-1} \otimes_1 \mathbb{Q} \text{ in } \mathcal{M}(\mathcal{X}(k))$$

is supported by $\mathcal{X}\{k\}$. To this end, fix a shift-invariant μ in $\mathcal{M}(\mathcal{X}(k))$ and note that $\int N[m] d\mu = 1$ for $m = 1, \ldots, k$. Hence we can associate shifted probability measures $S^m(\mu) \in \mathcal{M}(\mathcal{X}(k-m))$ with μ such that

$$S^0(\mu) = \mu, \quad S^m(\mu) = S(S^{m-1}(\mu)), \qquad \text{for } m = 1, \ldots, k,$$

where $S(\mu)$ is defined as in (2.14). The shift-invariance of μ implies that $S^m(\mu) \circ \pi_{k-m,1}^{-1}$ is independent of $m = 0, \ldots, k-1$.

Recall that the measure $S^{k-1}(\mu)$ of each $(a,c) \in \mathcal{X}(1)$ is the expectation under μ of the number of vertices of generation $k-1$ of the tree whose type is $a \in \mathcal{X}$ and which have offspring $c \in \mathcal{X}^*$. Our assumption that $\mu \ll \mu \circ \pi_{k,k-1}^{-1} \otimes_1 \mathbb{Q}$ thus implies that

$$\text{supp } S^{k-1}(\mu) \subset \text{supp } \mu_1 = \mathcal{X}\{1\}.$$

Consequently, $S^m(\mu) \circ \pi_{k-m,1}^{-1}$ are supported by $\mathcal{X}\{1\}$ for all $m = 0, \ldots, k-1$, which implies that μ is supported by $\mathcal{X}\{k\}$ as claimed. $\square$

To move from the empirical k-generation offspring measures M_X^k to the empirical subtree measure T_X we use the Dawson-Gärtner theorem, see e.g. [DZ98, Theorem 4.6.1]. Note that the spaces $\mathcal{X}(k)$ and the canonical projections $\pi_{k,l}$, $k \geq l$, form a projective system of Polish spaces and that the projective limit coincides with the Polish space $\bar{\mathcal{X}}$.

Similarly, the probability measures on $\mathcal{X}(k)$ with the projections $\pi_{k,l}^*$ defined by $\pi_{k,l}^*(\mu) = \mu \circ \pi_{k,l}^{-1}$ form a projective system and the projective limit is the Polish space $\mathcal{M}(\bar{\mathcal{X}})$ described before Theorem 2.3 and the canonical projections $\pi_k^* : \mathcal{M}(\bar{\mathcal{X}}) \to \mathcal{M}(\mathcal{X}(k))$ can be defined by $\pi_k^*(\mu) = \mu \circ \pi_k^{-1}$. Details follow from an argument similar to the one given in [DZ98, Lemma 6.5.14]. Recalling that $M_X^k = T_X \circ \pi_k^{-1}$, the Dawson-Gärtner theorem yields the following corollary of Lemma 3.8 (see for example [DZ98, Corollary 6.5.15] for a similar derivation).

Corollary 3.9. *Suppose that X is an irreducible, critical multitype Galton-Watson tree with uniformly bounded offspring sizes, conditioned to have exactly n vertices. Then, for $n \to \infty$, the empirical subtree measure T_X satisfies a large deviation principle in $\mathcal{M}(\bar{\mathcal{X}})$ with speed n and convex, good rate function*

$$\tilde{K}(\mu) = \sup_{k \geq 1} J_k\big(\mu \circ \pi_k^{-1}\big).$$

To complete the proof of Theorem 2.3 it just remains to show that $\tilde{K}(\cdot) = K(\cdot)$. For this purpose first assume that $\mu \in \mathcal{M}(\bar{\mathcal{X}})$ is shift-invariant. Then, for each $k \geq 1$ and $a \in \mathcal{X}(k-1)$,

$$\begin{aligned}\big(\mu \circ \pi_k^{-1}\big) \circ \pi_{k,k-1}^{-1}(a) = S(\mu) \circ \pi_{k-1}^{-1}(a) &= \int d\mu(X) \sum_{i=1}^{N} \delta_{X_{k-1}^{v_i}}(a) \\ &= \int d\mu(X)\, m_k(a, \pi_k X) = \sum_{b \in \mathcal{X}(k)} \mu \circ \pi_k^{-1}(b)\, m_k(a,b).\end{aligned}$$

In other words, for each $k \geq 1$, the measure $\mu \circ \pi_k^{-1}$ is shift-invariant in $\mathcal{M}(\mathcal{X}(k))$. Conversely, if $\mu \circ \pi_k^{-1}$ is shift-invariant in $\mathcal{M}(\mathcal{X}(k))$ for every $k \geq 1$, the same calculation shows that $\mu = S(\mu)$ on the collection of sets of the form $\pi_k^{-1}(A)$ for any $k \geq 1$ and $A \subset \mathcal{X}(k)$. As this collection of sets is closed under finite intersections and it generates the Borel σ-field on $\bar{\mathcal{X}}$, we infer that μ itself is shift-invariant.

Recall the definition of the projections $\mathfrak{p}_0, \mathfrak{p}_1$ for backward trees from Section 2.3. For the proof of Theorem 2.3 it only remains to verify the following lemma.

Lemma 3.10. *For every shift-invariant probability measure μ on $\bar{\mathcal{X}}$ we have*

$$H\big(\mu^* \circ \mathfrak{p}_1^{-1} \,\big\|\, \mu^* \circ \mathfrak{p}_0^{-1} \otimes \mathbb{Q}\big) = \sup_{k \geq 2} H\big(\mu \circ \pi_k^{-1} \,\big\|\, \mu \circ \pi_{k-1}^{-1} \otimes_1 \mathbb{Q}\big). \tag{3.47}$$

Proof. Define projections $\pi_k^j : \bar{\mathcal{X}} \to \mathcal{X}(k)$ as follows: Order the vertices $v_1, v_2, \dots$ in generation $k-1$ of $x \in \bar{\mathcal{X}}$ from left to right, with v_1 the leftmost. The tree $\pi_k^j(x)$ is obtained by removing all vertices in generations exceeding k and all vertices in generation k whose parent is some v_l, $l \geq j$. In particular, $\pi_k^1(x) = \pi_{k-1}(x)$ and $\pi_k^j(x) = \pi_k(x)$ for all $j > N[k-1](x)$. Let $\mu \circ (\pi_k^j)^{-1} \otimes_j \mathbb{Q}$ denote the measure obtained by sampling X according to μ and then independently adding offspring according to $\mathbb{Q}$ to each of the vertices v_l for $l \geq j$ in generation $k-1$ of $\pi_k^j(X)$. Observe that we define this measure for *all* j and that in many cases no vertices in generation k are removed or added. Assume first that $\mu \circ \pi_k^{-1} \ll \mu \circ \pi_{k-1}^{-1} \otimes_1 \mathbb{Q}$. Then, in case $\mu \circ \pi_k^{-1}(x) > 0$ and $N[k-1](x) = n \geq 1$ we find that

$$\frac{\mu \circ \pi_k^{-1}(x)}{\mu \circ \pi_{k-1}^{-1} \otimes_1 \mathbb{Q}(x)} = \prod_{j=1}^{n} \frac{\mu \circ (\pi_k^{j+1})^{-1} \otimes_{j+1} \mathbb{Q}(x)}{\mu \circ (\pi_k^j)^{-1} \otimes_j \mathbb{Q}(x)}, \tag{3.48}$$

with all the terms on the right hand side positive. Recall the definition of the measure $\mu_{k-1} = \mu^* \circ p_{k-1}^{-1}$ and the projections $\mathfrak{p}_{0,k-1}$, $\mathfrak{p}_{1,k-1}$ on $\mathcal{X}[k-1]$ from Section 2.3. Also recall that $(y, v) \in \mathcal{X}[k-1]$ denotes the tree $y \in \bar{\mathcal{X}}$ with centre v in generation $k-1$ of y. Hence, for $1 \leq j \leq n$,

$$\frac{\mu \circ (\pi_k^{j+1})^{-1} \otimes_{j+1} \mathbb{Q}(x)}{\mu \circ (\pi_k^j)^{-1} \otimes_j \mathbb{Q}(x)} = \frac{\mu_{k-1} \circ \mathfrak{p}_{1,k-1}^{-1}(\pi_k^j(x), v_j)}{\mu_{k-1} \circ \mathfrak{p}_{0,k-1}^{-1} \otimes \mathbb{Q}(\pi_k^j(x), v_j)} \tag{3.49}$$

with all terms positive. Note that

$$N[k-1](x)=0 \qquad \Longrightarrow \qquad \mu\circ\pi_k^{-1}(x)=\mu\circ\pi_{k-1}^{-1}\otimes_1\mathbb{Q}(x),$$

whereas if $y=\pi_k^j(x)$ with $N[k-1](x)=n>0$ then $N[k-1](y)=n$ and $\mu\circ(\pi_k^j)^{-1}(y)=\mu\circ\mathfrak{p}_{1,k-1}^{-1}(y,v_j)$ for any $1\le j\le n$. Hence, (3.48) and (3.49) imply that

$$\begin{aligned} H\big(\mu\circ\pi_k^{-1}\,\big\|\,\mu\circ\pi_{k-1}^{-1}\otimes_1\mathbb{Q}\big) &= \sum_{x\in\mathcal{X}(k)}\sum_{j=1}^{N[k-1](x)}\mu\circ\pi_k^{-1}(x)\log\Big(\frac{\mu_{k-1}\circ\mathfrak{p}_{1,k-1}^{-1}(\pi_k^j(x),v_j)}{\mu_{k-1}\circ\mathfrak{p}_{0,k-1}^{-1}\otimes\mathbb{Q}(\pi_k^j(x),v_j)}\Big)\\ &= H\big(\mu_{k-1}\circ\mathfrak{p}_{1,k-1}^{-1}\,\big\|\,\mu_{k-1}\circ\mathfrak{p}_{0,k-1}^{-1}\otimes\mathbb{Q}\big). \end{aligned} \tag{3.50}$$

Finally, note that

$$\mu\circ\pi_k^{-1}(x)>0 \text{ and } \mu\circ\pi_{k-1}^{-1}\otimes_1\mathbb{Q}(x)=0 \text{ for some } x\in\mathcal{X}(k),$$

if and only if there exists $1\le j\le N[k-1](x)$ such that

$$\mu_{k-1}\circ\mathfrak{p}_{1,k-1}^{-1}(\pi_k^j(x),v_j)>0 \text{ and } \mu_{k-1}\circ\mathfrak{p}_{0,k-1}^{-1}\otimes\mathbb{Q}(\pi_k^j(x),v_j)=0.$$

Consequently, $\mu\circ\pi_k^{-1}\ll\mu\circ\pi_{k-1}^{-1}\otimes_1\mathbb{Q}$ if and only if $\mu_{k-1}\circ\mathfrak{p}_{1,k-1}^{-1}\ll\mu_{k-1}\circ\mathfrak{p}_{0,k-1}^{-1}\otimes\mathbb{Q}$, with (3.50) holding for any shift-invariant $\mu\in\mathcal{M}(\bar{\mathcal{X}})$ and $k\ge 2$. By the identities $p_k\circ\mathfrak{p}_0=\mathfrak{p}_{0,k}\circ p_k$ and $p_k\circ\mathfrak{p}_1=\mathfrak{p}_{1,k}\circ p_k$ this amounts to

$$H\big(\mu\circ\pi_k^{-1}\,\big\|\,\mu\circ\pi_{k-1}^{-1}\otimes_1\mathbb{Q}\big)=H\big(\mu^*\circ\mathfrak{p}_1^{-1}\circ p_{k-1}^{-1}\,\big\|\,(\mu^*\circ\mathfrak{p}_0^{-1}\otimes\mathbb{Q})\circ p_{k-1}^{-1}\big). \tag{3.51}$$

The variational characterization of the relative entropy states that, for two probability measures ν_1,ν_2 on the Polish space $\underline{\mathcal{X}}$,

$$H\big(\nu_1\circ p_k^{-1}\,\big\|\,\nu_2\circ p_k^{-1}\big)=\sup_{\phi\in C_b(\mathcal{X}[k])}\Big\{\int_{\underline{\mathcal{X}}}\phi\circ p_k\,d\nu_1-\log\int_{\underline{\mathcal{X}}}e^{\phi\circ p_k}\,d\nu_2\Big\},$$

where $C_b(\mathcal{X}[k])$ is the set of continuous, bounded functions on $\mathcal{X}[k]$ (see for example [DZ98, Lemma 6.2.13]). Obviously, this expression is increasing in k and by the same representation it is bounded by $H(\nu_1\,\|\,\nu_2)$, which together with (3.51) shows that the left hand side of (3.47) is at least as large as its right hand side.

Conversely, for any continuous bounded function $\phi:\underline{\mathcal{X}}\to\mathbb{R}$ and $\varepsilon>0$ there exists a uniformly continuous function $\psi:\underline{\mathcal{X}}\to\mathbb{R}$ such that

$$\Big|\log\int_{\underline{\mathcal{X}}}e^{\phi}\,d\nu_2-\log\int_{\underline{\mathcal{X}}}e^{\psi}\,d\nu_2\Big|<\varepsilon \quad\text{ and }\quad \Big|\int_{\underline{\mathcal{X}}}\phi\,d\nu_1-\int_{\underline{\mathcal{X}}}\psi\,d\nu_1\Big|<\varepsilon.$$

Moreover, with $\underline{\mathcal{X}}$ being the projective limit of $\mathcal{X}[k]$, we can find a $k\ge 1$ and a continuous, bounded function $\psi^k:\mathcal{X}[k]\to\mathbb{R}$ such that $|\psi^k\circ p_k(x)-\psi(x)|<\varepsilon$ for all $x\in\underline{\mathcal{X}}$. Hence

$$\sup_{k\ge 2}\sup_{\phi\in C_b(\mathcal{X}[k])}\Big\{\int_{\underline{\mathcal{X}}}\phi\circ p_k\,d\nu_1-\log\int_{\underline{\mathcal{X}}}e^{\phi\circ p_k}\,d\nu_2\Big\}\ge\sup_{\phi\in C_b(\underline{\mathcal{X}})}\Big\{\int_{\underline{\mathcal{X}}}\phi\,d\nu_1-\log\int_{\underline{\mathcal{X}}}e^{\phi}\,d\nu_2\Big\},$$

which together with (3.51) shows that the right hand side of (3.47) is at least as large as its left hand side. This completes the proof of the lemma. $\square$

Bibliography

[Al91] D. Aldous. The continuum random tree II: An overview. In: *Stochastic analysis*, Proc. Symp., Durham/UK 1990, Lond. Math. Soc. Lect. Note Ser. 167 (1991), 23–70.

[AN72] K.B. Athreya and P.E. Ney. Branching processes. Springer, New York, (1972).

[BP94] I. Benjamini and Y. Peres. Markov chains indexed by trees. *Ann. Probab.* 22 (1994), 219–243.

[DZ98] A. Dembo and O. Zeitouni. Large deviations techniques and applications. Springer, New York, (1998).

[DGPZ02] A. Dembo, N. Gantert, Y. Peres and O. Zeitouni. Large deviations for random walks on Galton-Watson trees: averaging and uncertainty. *Probab. Theory Relat. Fields*, 122 (2002), 241-288.

[Du96] R. Durrett. Probability: theory and examples. Duxbury Press, Belmont, CA, (1996).

[Ge88] H.O. Georgii. Gibbs Measures and Phase Transitions. de Gruyter, Berlin (1988).

[KM02] W. König and P. Mörters. Brownian intersection local times: upper tails and thick points. *Ann. Probab.* 30 (2002), 1605-1656.

[LG99] J.-F. Le Gall. The Hausdorff measure of the range of super-Brownian motion. In: *Perplexing problems in probability*, Eds. M. Bramson, R. Durrett, pp. 285–314. Birkhäuser, Basel, (1999).

[LPP95] R. Lyons, R. Pemantle and Y. Peres. Ergodic theory on Galton–Watson trees: speed of random walk and dimension of harmonic measure. *Ergodic Theory Dyn. Systems* 15 (1995), 593–619.

[MM78] A. Meir and J.W. Moon. On the altitude of nodes in random trees. *Canad. J. Math.* 30 (1978), 997–1015.

[Pe95] R. Pemantle. Tree-indexed processes. *Statist. Sci.* 10 (1995), no. 2, 200–231.

AMIR DEMBO, Department of Mathematics, Stanford University
Stanford, CA 94305, USA.

PETER MÖRTERS, Department of Mathematical Sciences, University of Bath
Bath BA2 7AY, United Kingdom.

SCOTT SHEFFIELD, Microsoft Research
One Microsoft Way, Redmond WA 98052, USA.