Power laws for family sizes in a duplication model

by Rick Durrett* and Jason Schweinsberg†

Cornell University

May 28, 2004

Abstract

Qian, Luscombe, and Gerstein (2001) introduced a model of the diversification of protein folds in a genome that we may formulate as follows. Consider a multitype Yule process starting with one individual in which there are no deaths and each individual gives birth to a new individual at rate one. When a new individual is born, it has the same type as its parent with probability $1 - r$ and is a new type, different from all previously observed types, with probability r. We refer to individuals with the same type as families and provide an approximation to the joint distribution of family sizes when the population size reaches N. We also show that if $1 \ll S \ll N^{1-r}$, then the number of families of size at least S is approximately $CNS^{-1/(1-r)}$, while if $N^{1-r} \ll S$ the distribution decays more rapidly than any power.

Running head: Power laws for gene family sizes.

*Partially supported by NSF grants from the probability program (0202935) and from a joint DMS/NIGMS initiative to support research in mathematical biology (0201037).

†Supported by an NSF Postdoctoral Fellowship.

AMS 2000 subject classifications. Primary 60J80; Secondary 60J85, 92D15, 92D20.

Key words and phrases. Power law, Yule processes, multitype branching processes, genome sequencing.

1 Introduction

Genome sequencing of various species has shown that gene and protein-fold family sizes have a power-law distribution. Huynen and van Nimwegen (1998) studied six bacteria, two Archea, and yeast. Li, Gu, Wang, and Nekrutenko (2001) and later Gu, Cavalcanti, Chen, Bouman, and Li (2002) analyzed the genomes of yeast, the nematode *C. elegans*, fruit fly (*Drosophila melanogaster*), and human. There have been several models advanced to explain this phenomenon. Rzhetsky and Gomez (2001) and Karev et al (2002) (see also Koonin, Wolf, and Karev (2002)) introduced a birth and death model in which, when there are i individuals in a family, a birth occurs at rate λ_i and a death occurs at rate δ_i. They proved, as most readers of this journal can easily verify, that if the birth rates are second ordered balanced, i.e.,

$$\lambda_{i-1}/\delta_i = 1 - a/i + O(1/i^2),$$

then the stationary distribution is asymptotically Ci^{-a}. See the appendix of Karev et al (2002) or Example 3.6 on page 297 of Durrett (1996) for more details.

Qian, Luscombe and Gerstein (2001) introduced an alternative model that we will study in detail here. Consider a continuous-time Yule process with infinitely many types. At time zero, a single individual of type 1 is born. No individuals die, and each individual independently gives birth to a new individual at rate 1. When a new individual is born, it has the same type as its parent with probability $1 - r$, where $0 < r < 1$. With probability r, the new individual has a type which is different from all previously observed types. When the kth individual born has a different type from its parent, we say that it has type k. In this model, one can think of the new types as resulting from mutations, and r is the probability of mutation. Alternatively, one could think of a Yule process with immigration in which each individual gives birth at rate $1 - r$ and new immigrants arrive at rate r times the current population size. We refer to individuals with the same type as families. The goal of this paper is to study the distribution of the family sizes at the time when the population size reaches N.

1.1 Approximation to the family size distribution

Let T_N be the first time that the population size reaches N. Let $R_{k,N}$ be the number of individuals of type k at time T_N. Let $X_{k,N}$ be the fraction of individuals at time T_N whose type is in $\{1, \ldots, k\}$. Let $V_{k,N}$ be the fraction of individuals at time T_N, among those whose type is in $\{1, \ldots, k\}$, that are of type k. This means that the fraction of individuals at time T_N that are of type k is $V_{k,N}X_{k,N}$ and the number of individuals of type k at time T_N is $R_{k,N} = NV_{k,N}X_{k,N}$. Note that for $1 \le k \le N$, we have

$$X_{k,N} = \prod_{j=k+1}^{N} (1 - V_{j,N}). \tag{1.1}$$

The following proposition follows from well-known connections between Yule processes and Polya urns. We review these connections and prove this proposition in section 2.

Proposition 1.1. *For each integer $k \ge 2$, the limit*

$$W_k = \lim_{N\to\infty} V_{k,N}$$

exists a.s. The random variables $W_2, W_3, \ldots$ are independent. Furthermore, $P(W_k > 0) = r$ for all $k \geq 2$ and conditional on the event that $W_k > 0$, the distribution of W_k is Beta$(1, k-1)$.

For $1 \leq k \leq N$, let

$$Y_{k,N} = \prod_{j=k+1}^{N} (1 - W_j). \tag{1.2}$$

Let $\Delta = \{(x_i)_{i=1}^{\infty} : 0 \leq x_i \leq 1 \text{ for all } i \text{ and } \sum_{i=1}^{\infty} x_i = 1\}$. Note that the sequence of family sizes $(N^{-1}R_{k,N})_{k=1}^{\infty} = (V_{k,N}X_{k,N})_{k=1}^{\infty}$, whose kth term is the fraction of the population having type k at time T_N, is a sequence in Δ. Proposition 1.1 and equations (1.1) and (1.2) suggest that, for large N, the distribution of this sequence can be approximated by a sequence defined in terms of the W_k's. More specifically, let $Q_{r,N}$ be the distribution of the sequence in Δ whose first term is Y_1, whose kth term is W_kY_k for $2 \leq k \leq N$, and whose kth term is zero for $k > N$. Theorem 1.2 below exploits the coupling of the $X_{k,N}$ and $Y_{k,N}$ given above to show that the distribution of $(N^{-1}R_{k,N})_{k=1}^{\infty}$ can be approximated by $Q_{r,N}$ to within an error of $O(N^{-1/2})$. We prove this result in section 3.

Theorem 1.2. *We have* $E\left[\max_{1\leq k\leq N} |X_{k,N} - Y_{k,N}|\right] \leq \dfrac{5}{\sqrt{N}}$.

The distributions $Q_{r,N}$ first arose in the work of Durrett and Schweinsberg (2004) and Schweinsberg and Durrett (2003), who studied the effect of beneficial mutations on the genealogy of a population. The distributions $Q_{r,N}$ arise in that context because, shortly after a beneficial mutation, the number of individuals with the beneficial gene behaves like a supercritical branching process, which means that the number with descendants surviving a long time into the future behaves like a Yule process. In this setting, r is the rate of recombination, and individuals descended from a lineage with a recombination get traced back to a different ancestor than other individuals, just as individuals descended from an individual with a mutation in the present model are of a different type than the others. Schweinsberg and Durrett's (2003) approximation had an error of $O((\log N)^{-2})$ because of deaths and other complexities in the model, but Theorem 1.2 shows that the distributions $Q_{r,N}$ give a much more accurate approximation to the family-size distribution in the simpler model studied here.

1.2 A power law for the number of families of moderate size

Let $F_{S,N}$ denote the number of families at time T_N whose size is at least S. Define

$$g(S) = r\Gamma\left(\frac{2-r}{1-r}\right) N S^{-1/(1-r)}. \tag{1.3}$$

The theorem below, which is proved in section 4, shows that if $1 \ll S \ll N^{1-r}$, then $g(S)$ provides a good approximation to the number of families of size at least S, in the sense that $F_{S,N} - g(S)$ is $o(g(S))$.

Theorem 1.3. *There are constants $0 < C_1, C_2 < \infty$ so that*

$$E[|F_{S,N} - g(S)|] \leq C_1 g(S)\left[S^{-1/5} + (NS^{-1/(1-r)})^{-1/5}\right] + C_2.$$

Note that $S^{-1/5}$ and $(NS^{-1/(1-r)})^{-1/5}$ are both small and $g(S)$ is large when $1 \ll S \ll N^{1-r}$.

Theorem 1.3 confirms Qian, Luscombe, and Gerstein's (2001) power law but it also conflicts with their results. Since they considered the number of folds that occur exactly V times rather than at least V times, it follows from differentiating the right-hand side of (1.3) that for large N we would expect a decay with the power $b = 1 + 1/(1-r)$. This quantity is always larger than 2, while they observed powers b between 0.9 and 1.2 for eukaryotes and between 1.2 and 1.8 for prokaryotes. Despite this discrepancy, they were able to fit their model by starting the process at time zero with $N_0 > 1$ families. For example, for *Haemophilus influenzae* they took $r = 0.3$, $N_0 = 90$, and ran the process for 1249 generations. For *C. elegans* they took $r = 0.018$, $N_0 = 280$ and ran for 18,482 generations.

Figure 1 shows one simulation of the system with $r = 0.018$, $N_0 = 1$, and $N = 20,000$. In contrast to biologists who do a log-log plot of the number of gene families of size k, see e.g., Figure 1 in Harrison and Gerstein (2002), or Figure 8 in Karev et al (2002), we look at the tail of the distribution and plot the log of the family size on the x-axis and the number of families of at least that size on the y-axis. The curve fit by Karev et al (2002) has asymptotic power 1.9 in contrast to the 2.018 that comes from our formula, but note that the straight line fitted to our simulation of the distribution function has slope 0.91. Figure 2 shows the average of 10,000 simulations of the process with the *C. elegans* parameters. The straight line shows that Theorem 1.3 very accurately predicts the expected number of families until the log of the family size is 4. This simulation also shows that the power law breaks down when $S \gg N^{1-r}$ and brings us to our next topic.

1.3 Sizes of the largest families

The next result proves what was seen in the simulation: the expected number of families of size at least AN^{1-r} decays faster than any power of A.

Proposition 1.4. *For all $m > 0$, we have* $\lim_{A\to\infty} \left(\limsup_{N\to\infty} A^m E[F_{AN^{1-r},N}] \right) = 0.$

Proposition 1.5 below gives more specific information about the size of the largest families. Recall that $R_{k,N}$ is the number of individuals of type k at time T_N. Propositions 1.4 and 1.5 are proved in section 5.

Proposition 1.5. *Fix a positive integer k. Then, the limit*

$$Z = \lim_{N\to\infty} N^{r-1} R_{k,N}$$

exists almost surely. Furthermore, $P(Z > 0) = 1$ if $k = 1$ and $P(Z > 0) = r$ otherwise. Conditional on $Z > 0$, the distribution of Z is the same as the distribution of $Z_1 Z_2^{r-1}$, where the joint moment generating function of (Z_1, Z_2) is

$$\phi(s_1, s_2) = E[e^{-s_1 Z_1 - s_2 Z_2}] = \frac{1}{(1+s_2)^{k-1}(1+s_2+s_1(1+s_2)^r)} \tag{1.4}$$

for all $s_1 > 0$, $s_2 > 0$.

It is clear from Proposition 1.5 that for fixed k, the number of individuals of type k, if it is nonzero, is of order N^{1-r}. Note that when $k = 1$, the right-hand side of (1.4) reduces to $[1 + s_2 + s_1(1 + s_2)^r]^{-1}$. When $s_1 = 0$ this becomes $1/(1 + s_2)$, and when $s_2 = 0$ this becomes $1/(1 + s_1)$. Thus, both Z_1 and Z_2 have exponential distributions with mean one. By evaluating the mixed partial derivative of the right-hand side of (1.4) at $(0, 0)$, it is easy to calculate $E[Z_1Z_2] = 2 - r$ when $k = 1$. Since $E[Z_i] = \mathrm{Var}(Z_i) = 1$ for $i = 1, 2$, it follows that the correlation between Z_1 and Z_2 is simply $1 - r$. Kotz, Balakrishnan, and Johnson (2000) discuss several bivariate exponential distributions that have appeared in the literature, but we are not aware that the distribution of (Z_1, Z_2) has been studied previously and we are not aware of any expression for the joint density that would help us understand the distribution of $Z_1Z_2^{r-1}$. Figure 3 gives the observed sizes of the first family in 100,000 replications of the model with $r = 0.1$ and $N = 10,000$.

The reason we obtain a bivariate exponential distribution when $k = 1$ is that the individuals of type 1 form a Yule process in which particles split at rate $1-r$, while total population is a Yule process in which particles split at rate 1. Therefore, if $R_1(t)$ denotes the number of individuals at time t having type 1 and $R(t)$ denotes the total number of individuals at time t, then as $t \to \infty$ we have $e^{-(1-r)t}R_1(t) \to Z_1$ a.s. and $e^{-t}R(t) \to Z_2$ a.s., where Z_1 and Z_2 each have an exponential distribution with mean 1 (see Athreya and Ney (1972)). It follows that $R_1(t)R(t)^{r-1} \to Z_1Z_2^{r-1}$ a.s. as $t \to \infty$. More details are given in section 5.

1.4 A new Chinese restaurant

Our model has a close relation to a construction called the "Chinese restaurant process", which was first proposed by Dubins and Pitman and is described in Pitman (2002). Consider a restaurant with infinitely many circular tables. The first customer sits at the first table. For $n \geq 1$, the $(n + 1)$st customer sits at a new table with probability $\theta/(n + \theta)$ and otherwise picks one of the previous n customers at random and sits at that person's table. Note that our model is a variation of this in which the $(n + 1)$st customer sits at a new table with constant probability r, rather than with probability $\theta/(n + \theta)$.

The Chinese restaurant process describes the family size distribution in a Yule process with immigration when immigration occurs at constant rate θ. When there are n individuals in the Yule process, they are each splitting at rate 1 and immigration occurs at rate θ, so the probability that the $(n+1)$st individual starts a new family is $\theta/(n+\theta)$. Furthermore, the Chinese restaurant process describes the distribution of gene frequencies in a population in which each lineage experiences mutation at rate $\theta/2$ and whose ancestral structure is given by Kingman's coalescent (see Kingman, 1982), meaning that each pair of lineages merges at rate 1. Working backwards in time, when there are $n + 1$ lineages, coalescence occurs at rate $n(n + 1)/2$ while mutations occur at rate $\theta(n + 1)/2$. Consequently, the probability of having mutation before coalescence is $\theta/(n + \theta)$. For more details, see section 1.3 of Durrett (2002).

The Chinese restaurant process gives rise to a random partition Π_N of $\{1, \ldots, N\}$, where i and j are in the same block of Π_N if and only if the ith and jth customers are seated at the same table. If π is a partition of $\{1, \ldots, N\}$ with k blocks of sizes $n_1, \ldots, n_k$, one can check that

$$P(\Pi_N = \pi) = \frac{\theta^{k-1}}{(1 + \theta)(2 + \theta) \ldots (N - 1 + \theta)} \prod_{i=1}^{k} (n_i - 1)!, \tag{1.5}$$

which leads to the famous Ewens sampling formula (Ewens, 1972). Since the right-hand side of (1.5) depends only on $n_1, \ldots, n_k$, the random partition Π_N is exchangeable. Generalizations of the Chinese restaurant process can be used to construct arbitrary exchangeable and partially exchangeable random partitions (see Pitman, 1995).

If $M_{k,N}$ denotes the number, of the first N customers, who are seated at the kth table, then (see Pitman, 1995) the distribution of the Δ-valued sequence $(N^{-1}M_{1,N}, N^{-1}M_{2,N}, \ldots)$ converges to the Poisson-Dirichlet distribution with parameters $(0, \theta)$. This distribution is defined as follows. Let $(D_i)_{i=1}^{\infty}$ be an i.i.d. sequence of random variables having a Beta$(1, \theta)$ distribution. Then the sequence whose kth term is $D_k \prod_{j=1}^{k-1}(1 - D_j)$ has the Poisson-Dirichlet distribution with parameters $(0, \theta)$. The Poisson-Dirichlet distributions were studied extensively by Pitman and Yor (1997). In addition to their relevance in genetics, the Poisson-Dirichlet distributions have been shown to arise as limits for the component counts in various combinatorial structures, as observed by Arratia, Barbour, and Tavaré (2000); see also Pitman (2002). Pfaff (2003) studied the number of blocks of the partition Π_N of size between a and b for small θ. For this model, the number of large components does not follow a power law.

In the modified Chinese restaurant process studied here in which each customer sits at a new table with probability r, one can again define the associated random partition Θ_N of $\{1, \ldots, N\}$ such that i and j are in the same block if and only if the ith and jth customers are seated at the same table. In our branching processes interpretation, this means that the ith and jth individuals born have the same type. It is straightforward to derive an analog of the Ewens sampling formula in this case. If π is a partition of $\{1, \ldots, N\}$ into k blocks of sizes $n_1, \ldots, n_k$, and if $a_1 < a_2 < \cdots < a_k$ are the first integers in these blocks, then

$$P(\Theta_N = \pi) = \frac{\theta^{k-1}(1-\theta)^{N-k}}{(n-1)!} \left[\prod_{i=1}^{k} (n_i - 1)! \right] \prod_{j=2}^{k} (a_j - 1).$$

This formula depends on $a_2, \ldots, a_k$ as well as the block sizes $n_1, \ldots, n_k$, so the random partition Θ_N is not exchangeable. Nevertheless, one can still look for approximations to the distribution of the block sizes. We see from Theorem 1.2 that the distributions $Q_{r,N}$ play the role of the Poisson-Dirichlet distributions in this model.

2 Branching processes and Polya urns

In this section, we review some well-known connections between Polya urns and continuous-time branching processes, which will be useful later in the paper. Athreya and Karlin (1968) showed how to embed the urn process in a continuous-time branching process. This technique was reviewed in Athreya and Ney (1972). See Janson (2004) for a thorough survey of recent developments and generalizations.

Recall the following version of Polya's urn model. Suppose we start with a white balls and b black balls in the urn. We then draw a ball at random from the urn. If the ball we draw is white, we return it to the urn and add an additional white ball to the urn. If the ball we draw is black, we return it to the urn and add another black ball. This process can be repeated indefinitely. To see the connection with branching processes, consider a two-type branching process in which there are no deaths and each individual gives birth at rate 1. If at some time there are a individuals of type 1 and b individuals of type 2, then the probability that the next individual born will have

type 1 is $a/(a+b)$, which is the same as the probability that the next ball added to an urn with a white balls and b black balls will be white. It follows that the distribution of the number of type 1 individuals when the population size reaches N is the same as the distribution of the number of white balls in the urn when the number of balls in the urn is N.

Let $\zeta_i = 1$ if the ith ball added to the urn is white, and let $\zeta_i = 0$ if the ith ball added to the urn is black. Fix a positive integer N. Let $S \subset \{1, \ldots, N\}$, and let $S^c = \{1, \ldots, N\} \setminus S$. Let $|S|$ denote the cardinality of S. It is easy to check that

$$P(\zeta_i = 1 \text{ for } i \in S \text{ and } \zeta_i = 0 \text{ for } i \in S^c) = \frac{(a + |S| - 1)!(b + N - |S| - 1)!(a + b - 1)!}{(a-1)!(b-1)!(a+b+N-1)!}. \quad (2.1)$$

Since the right-hand side of (2.1) depends only on $|S|$ and not on the particular elements of S, the sequence $(\zeta_i)_{i=1}^{\infty}$ is exchangeable. By de Finetti's Theorem, there exists a probability measure μ on $[0, 1]$ such that for all N and all $S \subset \{1, \ldots, N\}$, we have

$$P(\zeta_i = 1 \text{ for } i \in S \text{ and } \zeta_i = 0 \text{ for } i \in S^c) = \int_0^1 x^{|S|}(1-x)^{N-|S|}\, \mu(dx), \quad (2.2)$$

where μ is the distribution of $\lim_{N\to\infty} N^{-1}|\{i \le N : \zeta_i = 1\}|$, the limiting fraction of white balls in the urn. It follows from Theorem 1 in section 9.1 of Chapter V of Athreya and Ney (1972) that μ is the Beta(a, b) distribution. One can also see this by checking that the right-hand sides of (2.1) and (2.2) agree in this case.

Proof of Proposition 1.1. Let S_k be the set of all i such that the type of the ith individual born is in $\{1, \ldots, k\}$. Let $\mathcal{H}_k$ be the σ-field generated by the sets $S_k, S_{k+1}, \ldots$. Note that if $j > k$ then $V_{j,N}$ is $\mathcal{H}_k$-measurable for all N. Therefore, to prove the proposition, it suffices to show that, for all $k \ge 2$, the limit W_k exists a.s. and satisfies the following conditions:

1. $P(W_k > 0) = r$.
2. The conditional distribution of W_k given $W_k > 0$ is Beta$(1, k-1)$.
3. W_k is independent of $\mathcal{H}_k$.

Note that the third condition implies that W_k is independent of $(W_j)_{j=k+1}^{\infty}$.

Enumerate the elements of S_k as $i_1 < i_2 < i_3 < \ldots$. Define a sequence $(\zeta_j^k)_{j=1}^{\infty}$ such that $\zeta_j^k = 1$ if the i_jth individual has type k and $\zeta_i^k = 0$ otherwise. Note that $i_j = j$ for $j \le k$. Also, $\zeta_j^k = 0$ for $j = 1, \ldots, k-1$. Furthermore, $\zeta_k^k = 1$ if and only if the kth individual has a new type, and whether or not this individual has a new type does not affect the births of individuals of types greater than k. Thus, $P(\zeta_k^k = 1|\mathcal{H}_k) = r$. If $\zeta_k^k = 0$ then clearly $W_k = 0$. Because of the connection between branching processes and Polya urns, if $\zeta_k^k = 1$ then the sequence $(\zeta_{j+k}^k)_{j=1}^{\infty}$ has the same distribution as the Polya urn sequence $(\zeta_i)_{i=1}^{\infty}$ defined above. Furthermore, the values of ζ_j^k do not affect the births of individuals of types greater than k, so this relationship holds even after conditioning on $\mathcal{H}_k$. It follows that, conditional on $\zeta_k^k = 1$, the random variable W_k has a Beta$(1, k-1)$ distribution and W_k is independent of $\mathcal{H}_k$. □

Now, fix N and to simplify notation, write X_k, Y_k, V_k, and R_k for $X_{k,N}$, $Y_{k,N}$, $V_{k,N}$, and $R_{k,N}$ respectively. We will use this notation throughout the rest of the paper when the value of

N is clear from the context. Let $\mathcal{F}_k$ be the σ-field generated by the random variables V_j and W_j for $j \geq k+1$. It follows from (1.1) and (1.2) that X_k and Y_k are $\mathcal{F}_k$-measurable. Let $\mathcal{G}_k$ be the σ-field generated by the random variables V_j for $j \geq k+1$ and W_j for $j \geq k$.

We can write $W_k = \xi_k \tilde{W}_k$, where ξ_k has a Bernoulli(r) distribution and is independent of $\mathcal{F}_k$, and $\tilde{W}_k$ has a Beta$(1, k-1)$ distribution and is independent of ξ_k and $\mathcal{F}_k$. Since $E[\tilde{W}_k] = 1/k$ and $E[\tilde{W}_k^2] = 2/[k(k+1)]$, we have $E[W_k|\mathcal{F}_k] = r/k$ and $E[W_k^2|\mathcal{F}_k] = 2r/[k(k+1)]$. Note that $V_k = 0$ whenever $W_k = 0$. On $\{W_k > 0\}$, define $\tilde{V}_k = R_k - 1$. Define $\tilde{V}_k = 0$ on $\{W_k = 0\}$. Then

$$V_k = \left[\frac{1+\tilde{V}_k}{NX_k}\right] 1_{\{W_k>0\}} = \left[\left(\frac{1}{k}\right)\left(\frac{k}{NX_k}\right) + \left(\frac{\tilde{V}_k}{NX_k-k}\right)\left(\frac{NX_k-k}{NX_k}\right)\right] 1_{\{W_k>0\}}. \tag{2.3}$$

It follows from (2.2) that

$$\textit{the conditional distribution of } \tilde{V}_k \textit{ given } \mathcal{G}_k \textit{ is Binomial}(NX_k - k, W_k) \tag{2.4}$$

because there are $NX_k - k$ individuals, after the first k, with types in $\{1, \ldots, k\}$, and, conditional on $\mathcal{G}_k$, each has type k with probability W_k. Therefore,

$$E[V_k|\mathcal{G}_k] = \left[\frac{1}{NX_k} + W_k\left(\frac{NX_k-k}{NX_k}\right)\right] 1_{\{W_k>0\}} \tag{2.5}$$

and

$$E[V_k|\mathcal{F}_k] = E[E[V_k|\mathcal{G}_k]|\mathcal{F}_k] = r/k.$$

3 Approximating the family size distribution

In this section, we prove Theorem 1.2, which implies that the distribution $Q_{r,N}$ is a good approximation to the family size distribution in the Yule process with infinitely many types. To prove this result, we need to show that the X_k, which are related to the V_j by (1.1), are close to the Y_k, which are likewise related to the W_j by (1.2). We begin by showing that $E[X_k]$ and $E[Y_k]$ are the same.

Lemma 3.1. *We have* $E[X_k] = E[Y_k] = \prod_{j=k+1}^{N} \left(1 - \frac{r}{j}\right)$ *for* $1 \leq k \leq N$.

Proof. We prove the formula for $E[Y_k]$ by backwards induction on k. Clearly $E[Y_N] = 1$. Suppose the formula holds for some $k \geq 2$. Then

$$E[Y_{k-1}] = E[(1-W_k)Y_k] = E[(1-W_k)]E[Y_k] = \left(1 - \frac{r}{k}\right) \prod_{j=k+1}^{N} \left(1 - \frac{r}{j}\right) = \prod_{j=k}^{N} \left(1 - \frac{r}{j}\right).$$

To get the same formula for $E[X_k]$, first note that $E[X_{k,j}] = 1$ for $1 \leq j \leq k$. If $n \geq k$ then conditional on $X_{k,n}$, the probability that the $(n+1)$st individual has a type in $\{1, \ldots, k\}$ is $(1-r)X_{k,n}$. Therefore,

$$E[X_{k,n+1}] = \frac{nE[X_{k,n}] + (1-r)E[X_{k,n}]}{n+1} = \left(1 - \frac{r}{n+1}\right) E[X_{k,n}],$$

so the formula for $E[X_k]$ follows by induction on n. □

Lemma 3.2. *We have* $\left(\frac{k}{N}\right)^r e^{-r^2/k} \le E[X_k] \le \left(\frac{k}{N}\right)^r e^{r/k}$ *for* $1 \le k \le N$.

Proof. By Lemma 3.1, we have $\log E[X_k] = \sum_{j=k+1}^{N} \log(1 - r/j)$. Note that if $0 \le x < 1$ then $\log(1-x) = -\sum_{k=1}^{\infty}(x^k/k)$. Summing, we see that if $0 \le x \le 1/2$, $-(x+x^2) \le \log(1-x) \le -x$. Therefore,

$$\log E[X_k] \le -\sum_{j=k+1}^{N} \frac{r}{j} = \frac{r}{k} - \sum_{j=k}^{N} \frac{r}{j} \le \frac{r}{k} - \int_k^N \frac{r}{x}\,dx = \frac{r}{k} + \log\left(\frac{k}{N}\right)^r, \tag{3.1}$$

$$\log E[X_k] \ge -\sum_{j=k+1}^{N} \left(\frac{r}{j} + \frac{r^2}{j^2}\right) \ge -\int_k^N \frac{r}{x}\,dx - \int_k^N \frac{r^2}{x^2}\,dx \ge \log\left(\frac{k}{N}\right)^r - \frac{r^2}{k}. \tag{3.2}$$

The result now follows by exponentiating both sides in (3.1) and (3.2). □

Lemma 3.3. *We have* $E[X_k^2(W_k - V_k)^2] \le r\left(\frac{1}{N^2} + \frac{2}{N^{1+r}k^{1-r}}\right)$ *for* $2 \le k \le N$.

Proof. By (2.3), we have

$$V_k - W_k = \left[\left(\frac{1}{k} - W_k\right)\left(\frac{k}{NX_k}\right) + \left(\frac{\tilde{V}_k}{NX_k - k} - W_k\right)\left(\frac{NX_k - k}{NX_k}\right)\right] 1_{\{W_k > 0\}}. \tag{3.3}$$

When we take the conditional expectation given $\mathcal{G}_k$ of the square of the right-hand side of (3.3), the cross-term vanishes because (2.4) implies

$$E\left[\frac{\tilde{V}_k}{NX_k - k} - W_k \middle| \mathcal{G}_k\right] = 0.$$

Since X_k and W_k are $\mathcal{G}_k$ measurable, using (2.4) again gives

$$\begin{aligned} &E[(V_k - W_k)^2 | \mathcal{G}_k] \\ &\quad = E\left[\left(\frac{1}{k} - W_k\right)^2\left(\frac{k}{NX_k}\right)^2 1_{\{W_k>0\}} + \left(\frac{\tilde{V}_k}{NX_k - k} - W_k\right)^2\left(\frac{NX_k - k}{NX_k}\right)^2 1_{\{W_k>0\}} \middle| \mathcal{G}_k\right] \\ &\quad = \left(\frac{1}{k} - W_k\right)^2\left(\frac{k}{NX_k}\right)^2 1_{\{W_k>0\}} + \frac{W_k(1-W_k)}{NX_k - k}\left(\frac{NX_k - k}{NX_k}\right)^2 1_{\{W_k>0\}}. \end{aligned}$$

Since W_k is independent of $\mathcal{F}_k$ and the conditional distribution of W_k given $W_k > 0$ is Beta$(1, k-1)$,

$$E[(V_k - W_k)^2 | \mathcal{F}_k] = \frac{k-1}{k^2(k+1)}\left(\frac{k}{NX_k}\right)^2 r + \frac{NX_k - k}{(NX_k)^2}\left(\frac{r}{k} - \frac{2r}{k(k+1)}\right) \le \frac{r}{N^2X_k^2} + \frac{r}{NkX_k}.$$

Thus, using Lemma 3.2, we get

$$\begin{aligned} E[X_k^2(W_k - V_k)^2] = E[X_k^2 E[(W_k - V_k)^2 | \mathcal{F}_k]] &\le E\left[\frac{r}{N^2} + \frac{rX_k}{Nk}\right] \\ &\le \frac{r}{N^2} + \frac{r}{N^{1+r}k^{1-r}}e^{r/k} \le r\left(\frac{1}{N^2} + \frac{2}{N^{1+r}k^{1-r}}\right), \end{aligned}$$

since for $k \ge 2$, we have $e^{1/2} \le 2$. □

Lemma 3.4. *For every real number a, we have $E[X_k(X_k - Y_k)(W_k - V_k)(W_k - a)] = 0$.*

Proof. Using the fact that $E[W_k - V_k|\mathcal{F}_k] = 0$ and that W_k is $\mathcal{G}_k$-measurable, we have

$$\begin{aligned} E[(W_k - V_k)(W_k - a)|\mathcal{F}_k] &= E[W_k(W_k - V_k)|\mathcal{F}_k] \\ &= E[E[W_k(W_k - V_k)|\mathcal{G}_k]|\mathcal{F}_k] = E[W_k^2 - W_k E[V_k|\mathcal{G}_k]|\mathcal{F}_k]. \end{aligned}$$

Using (2.5) now, the above equals

$$\begin{aligned} &E\left[W_k^2 - W_k\left(\frac{1}{NX_k} + W_k\left(\frac{NX_k - k}{NX_k}\right)\right)1_{\{W_k>0\}}\middle|\mathcal{F}_k\right] \\ &\quad = \frac{1}{NX_k}\big(kE[W_k^2|\mathcal{F}_k] - E[W_k|\mathcal{F}_k]\big) = \frac{r}{NX_k}\left(\frac{2}{k+1} - \frac{1}{k}\right). \end{aligned}$$

It follows that

$$\begin{aligned} &E[X_k(X_k - Y_k)(W_k - V_k)(W_k - a)] = E[E[X_k(X_k - Y_k)(W_k - V_k)(W_k - a)|\mathcal{F}_k]] \\ &= E[X_k(X_k - Y_k)E[(W_k - V_k)(W_k - a)|\mathcal{F}_k] = E\left[(X_k - Y_k)\left(\frac{r}{N}\right)\left(\frac{2}{k+1} - \frac{1}{k}\right)\right] = 0, \end{aligned}$$

where the last equality follows from Lemma 3.1. □

Lemma 3.5. *We have $E[(X_k - Y_k)^2] \le 3/N$ for $1 \le k \le N$.*

Proof. Suppose $2 \le k \le N$. We will bound $E[(X_{k-1} - Y_{k-1})^2]$ in terms of $E[(X_k - Y_k)^2]$. First, note that it follows from (1.1) and (1.2) that

$$X_{k-1} - Y_{k-1} = (1 - V_k)X_k - (1 - W_k)Y_k = X_k(W_k - V_k) + (X_k - Y_k)(1 - W_k). \tag{3.4}$$

Thus,

$$\begin{aligned} E[(X_{k-1} - Y_{k-1})^2] &= E[X_k^2(W_k - V_k)^2] + E[(X_k - Y_k)^2(1 - W_k)^2] \\ &\quad + 2E[X_k(X_k - Y_k)(W_k - V_k)(1 - W_k)]. \end{aligned} \tag{3.5}$$

By Lemma 3.4 with $a = 1$, the third term on the right-hand side of (3.5) vanishes. Using Lemma 3.3 and the fact that $E[(X_k - Y_k)^2(1 - W_k)^2] \le E[(X_k - Y_k)^2]$, we get

$$E[(X_{k-1} - Y_{k-1})^2] \le E[(X_k - Y_k)^2] + r\left(\frac{1}{N^2} + \frac{2}{N^{1+r}k^{1-r}}\right).$$

Since $X_N = Y_N = 1$, it follows that for $1 \le k \le N$, we have

$$\begin{aligned} E[(X_k - Y_k)^2] &\le \sum_{j=2}^{N} r\left(\frac{1}{N^2} + \frac{2}{N^{1+r}j^{1-r}}\right) \le \frac{r}{N} + \frac{2r}{N^{1+r}}\sum_{j=2}^{N}\frac{1}{j^{1-r}} \\ &\le \frac{1}{N} + \frac{2r}{N^{1+r}}\int_1^N \frac{1}{x^{1-r}}\,dx \le \frac{1}{N} + \frac{2r}{N^{1+r}}\left(\frac{N^r}{r}\right) = \frac{3}{N}, \end{aligned} \tag{3.6}$$

which completes the proof. □

Proof of Theorem 1.2. Let $M = \max_{1\le k\le N} |X_k - Y_k|$. Fix $x > 0$. Let $T = \max\{k : |X_k - Y_k| \ge x\}$ if $M \ge x$, and let $T = 0$ otherwise. For $2 \le k \le N$, define

$$\rho_k = X_k(W_k - V_k) - (X_k - Y_k)(W_k - r/k) \tag{3.7}$$

so that by (3.4), $X_{k-1} - Y_{k-1} = \rho_k + (X_k - Y_k)(1 - r/k)$. Let

$$H_k = \begin{cases} X_k - Y_k & \text{for } k \ge T \\ X_T - Y_T + \sum_{j=k+1}^{T} \rho_j & \text{for } k < T. \end{cases}$$

This definition is chosen so that

$$H_{k-1} - H_k = \rho_k - (r/k)H_k 1_{\{k>T\}} \tag{3.8}$$

Our first step is to show

$$P(M \ge x) \le x^{-2} E[H_1^2]. \tag{3.9}$$

To establish (3.9), we mimic the proof of Kolmogorov's Maximal Inequality in Durrett (1996). Let $A_k = \{T = k\}$, so the event that $M \ge x$ is the event $\cup_{k=1}^N A_k$. Then

$$\begin{aligned} E[H_1^2] &\ge \sum_{k=1}^N E[H_1^2 1_{A_k}] = \sum_{k=1}^N E[(H_k^2 + 2H_k(H_1 - H_k) + (H_1 - H_k)^2)1_{A_k}] \\ &\ge \sum_{k=1}^N E[H_k^2 1_{A_k}] + 2\sum_{k=1}^N E[H_k(H_1 - H_k)1_{A_k}]. \end{aligned}$$

If $j \le k$ then

$$E[\rho_j|\mathcal{F}_k] = E[E[\rho_j|\mathcal{F}_j]|\mathcal{F}_k] = E[X_j E[W_j - V_j|\mathcal{F}_j] - (X_j - Y_j)E[W_j - r/j|\mathcal{F}_j]|\mathcal{F}_k] = 0. \tag{3.10}$$

Therefore,

$$\begin{aligned} \sum_{k=1}^N E[H_k(H_1 - H_k)1_{A_k}] &= \sum_{k=1}^N E[E[H_k(\rho_2 + \cdots + \rho_k)1_{A_k}|\mathcal{F}_k]] \\ &= \sum_{k=1}^N E[H_k 1_{A_k} E[\rho_2 + \cdots + \rho_k|\mathcal{F}_k]] = 0. \end{aligned}$$

It follows that

$$E[H_1^2] \ge \sum_{k=1}^N E[H_k^2 1_{A_k}] \ge \sum_{k=1}^N x^2 P(A_k) = x^2 P(M \ge x),$$

which implies (3.9).

We now obtain a bound on $E[H_1^2]$. Using (3.8) and the fact that the random variable H_k and the event $\{k > T\}$ are $\mathcal{F}_k$-measurable, we have

$$\begin{aligned} E[H_{k-1}^2|\mathcal{F}_k] &= E[(\rho_k + H_k(1 - (r/k)1_{\{k>T\}}))^2|\mathcal{F}_k] \\ &= E[\rho_k^2|\mathcal{F}_k] + 2H_k(1 - (r/k)1_{\{k>T\}})E[\rho_k|\mathcal{F}_k] + H_k^2((1 - (r/k)1_{\{k>T\}})^2. \end{aligned}$$

Since $E[\rho_k|\mathcal{F}_k] = 0$ by (3.10), it follows that $E[H_{k-1}^2|\mathcal{F}_k] \le E[\rho_k^2|\mathcal{F}_k] + H_k^2$, and thus $E[H_{k-1}^2] \le E[\rho_k^2] + E[H_k^2]$. Since $H_N = 0$, we can combine this result with (3.9) to get

$$P(M \ge x) \le x^{-2} E[H_1^2] \le x^{-2} \sum_{k=2}^{N} E[\rho_k^2].$$

To bound $E[\rho_k^2]$ we recall the definition in (3.7) and use Lemma 3.5 and the fact that W_k is independent of X_k and Y_k to get

$$\begin{aligned} E[(X_k - Y_k)^2 (W_k - r/k)^2] &\le E[(X_k - Y_k)^2] E\left[W_k^2 - \frac{2r}{k} W_k + \frac{r^2}{k^2}\right] \\ &\le \frac{3}{N}\left(\frac{2r}{k(k+1)} - \frac{2r^2}{k^2} + \frac{r^2}{k^2}\right) \le \frac{6r}{Nk(k+1)}. \end{aligned}$$

Combining this result with Lemma 3.3 and Lemma 3.4 with $a = r/k$, we get

$$E[\rho_k^2] \le r\left(\frac{1}{N^2} + \frac{2}{N^{1+r} k^{1-r}} + \frac{6}{Nk(k+1)}\right). \tag{3.11}$$

The telescoping sum $\sum_{k=2}^{N} 6r/[Nk(k+1)] \le 3r/N$, so it follows from (3.6) and (3.11) that

$$P(M \ge x) \le x^{-2} \sum_{k=2}^{N} E[\rho_k^2] \le x^{-2}\left(\frac{3}{N} + \frac{3r}{N}\right) \le \frac{6}{Nx^2}.$$

Thus,

$$E[M] = \int_0^\infty P(M \ge x)\, dx \le \frac{2}{\sqrt{N}} + \int_{2/\sqrt{N}}^\infty \frac{6}{Nx^2}\, dx = \frac{2}{\sqrt{N}} + \frac{3}{\sqrt{N}} = \frac{5}{\sqrt{N}},$$

which proves the theorem. □

4 The power law

In this section, we prove Theorem 1.3, which gives the power law for the family size distribution. Our first lemma gives a bound on the moments of the binomial distribution. Throughout this section, we allow the value of the constant C to change from line to line.

Lemma 4.1. *Fix $m \ge 1$. There exists a constant C such that for all n and p such that $np \ge 1$, if X has a Binomial(n,p) distribution then*

$$E\left[\left|\frac{X}{n} - p\right|^m\right] \le C\left(\frac{p}{n}\right)^{m/2}.$$

Proof. For now, we assume that $p \le 1/2$. The proof is based on two bounds for binomial tail probabilities. If $z > 0$ then

$$P\left(\frac{X}{n} - p \le -z\right) \le e^{-nz^2/2p}, \tag{4.1}$$

and if $0 < z < 1 - p$, then

$$P\left(\frac{X}{n} - p \geq z\right) \leq e^{-nz^2/2(p+z)}. \tag{4.2}$$

Equation (4.1) follows from (3.52) on p. 121 of Johnson, Kotz, and Kemp (1992). To prove (4.2), we use the fact that if $p < a < 1$, then $P(X/n \geq a) \leq e^{-nH(a)}$, where

$$H(a) = a\log(a/p) + (1-a)\log((1-a)/(1-p)).$$

This is proved, for example, in Arratia and Gordon (1989). We have $H'(a) = \log(a/p) - \log((1-a)/(1-p))$ and $H''(a) = 1/[a(1-a)]$. Since $H(p) = H'(p) = 0$, by Taylor's Theorem there exists $z \in [p, a]$ such that $H(a) = \frac{1}{2}H''(z)(a-p)^2$. Note that the function $a \mapsto H''(a)$ is decreasing on $(0, 1/2)$ and increasing on $(1/2, 1)$. Therefore, if $a \leq 1/2$ then $H(a) \geq \frac{1}{2}H''(a)(a-p)^2 \geq \frac{1}{2a}(1-p)^2$ and if $a \geq 1/2$ then $H(a) \geq \frac{1}{2}H''(1/2)(a-p)^2 \geq 2(a-p)^2 \geq \frac{1}{2a}(a-p)^2$. Equation (4.2) follows by substituting $z = a - p$.

Now, using Lemma 5.7 in chapter 1 of Durrett (1996), we get

$$E\left[\left|\frac{X}{n} - p\right|^m\right] = \int_0^p mz^{m-1}P\left(\left|\frac{X}{n} - p\right| > z\right) dz + \int_p^{1-p} mz^{m-1}P\left(\left|\frac{X}{n} - p\right| > z\right) dz. \tag{4.3}$$

Using (4.1) and (4.2), then $z \leq p$, and making the substitution $z = y\sqrt{4p/n}$, the first term on the right is

$$\begin{aligned} &\leq \int_0^p mz^{m-1}\left(e^{-nz^2/2p} + e^{-nz^2/2(p+z)}\right) dz \leq 2m\int_0^p z^{m-1}e^{-nz^2/4p}\, dz \\ &\leq 2m\int_0^\infty \left(\frac{4p}{n}\right)^{m/2} y^{m-1}e^{-y^2}\, dy \leq C\left(\frac{p}{n}\right)^{m/2}. \end{aligned} \tag{4.4}$$

Likewise, using $z/(p+z) \geq 1/2$ for $z \geq p$ and substituting $z = 4y/n$, the second term on the right in (4.3) is

$$\begin{aligned} &\leq \int_p^{1-p} mz^{m-1}e^{-nz^2/2(p+z)}\, dz \leq m\int_p^{1-p} z^{m-1}e^{-nz/4}\, dz \\ &\leq m\int_0^\infty \left(\frac{4}{n}\right)^m y^{m-1}e^{-y}\, dy \leq \frac{C}{n^m}. \end{aligned} \tag{4.5}$$

It follows from (4.3), (4.4), and (4.5) that if $p \leq 1/2$ and $np \geq 1$, then

$$E\left[\left|\frac{X}{n} - p\right|^m\right] \leq C\left(\frac{p}{n}\right)^{m/2} + C\left(\frac{1}{n}\right)^m \leq C\left(\frac{p}{n}\right)^{m/2}. \tag{4.6}$$

The fact that $np \geq 1$ was used only for the second inequality in (4.6). Therefore, if $p \geq 1/2$ and $np \geq 1$, we can use the first inequality in (4.6) to get

$$E\left[\left|\frac{X}{n} - p\right|^m\right] = E\left[\left|\frac{n-X}{n} - (1-p)\right|^m\right] \leq C\left(\frac{1-p}{n}\right)^{m/2} + C\left(\frac{1}{n}\right)^m \leq C\left(\frac{p}{n}\right)^{m/2},$$

which completes the proof of the lemma. □

We will also need, both in this section and section 5, bounds on tail probabilities for the Gamma distribution.

Lemma 4.2. *Suppose X has a Gamma$(m,1)$ distribution and $a > 0$. Then*

$$P(X \le a) \le a^m/\Gamma(m+1) \qquad \text{and} \qquad P(X \ge a) \le 2^m e^{-a/2}.$$

Proof. The density of X is $f(x) = e^{-x}x^{m-1}/\Gamma(m)$ for $x > 0$. Therefore,

$$P(X \le a) = \int_0^a \frac{e^{-x}x^{m-1}}{\Gamma(m)}\,dx \le \frac{1}{\Gamma(m)}\int_0^a x^{m-1}\,dx = \frac{a^m}{m\Gamma(m)} = \frac{a^m}{\Gamma(m+1)}.$$

Also, recall that $E[e^{\lambda X}] = (1-\lambda)^{-m}$ for all $\lambda < 1$, so taking $\lambda = 1/2$ and using Markov's Inequality, we get $P(X \ge a) \le e^{-\lambda a}E[e^{\lambda X}] = (1-\lambda)^{-m}e^{-\lambda a} = 2^m e^{-a/2}$. □

The next lemma bounds the moments of X_k.

Lemma 4.3. *Fix a real number $m \ge 1$. Then there is a positive constant C such that for all $k \ge 1$,*

$$E[X_k^m] \le \left(\frac{k}{N}\right)^{mr}\left(1 + \frac{C}{k}\right).$$

Proof. Let $M_{k,l} = \sum_{j=1}^k R_{j,l}$ be the number of individuals at time T_l with types in $\{1,\dots,k\}$. Note that $M_{k,N} = NX_k$. Conditional on $M_{k,l}$, the probability that the $(l+1)$st individual born has a type in $\{1,\dots,k\}$ is $(1-r)M_{k,l}/l$. Therefore,

$$E[M_{k,l+1}^m | M_{k,l}] = M_{k,l}^m + (1-r)\left(\frac{M_{k,l}}{l}\right)\left[(M_{k,l}+1)^m - M_{k,l}^m\right].$$

Since $b^m - a^m = \int_a^b mx^{m-1}\,dx \le mb^{m-1}(b-a)$ for $0 \le a \le b$, the above is

$$\begin{aligned}
&\le M_{k,l}^m + (1-r)\left(\frac{M_{k,l}}{l}\right)m(M_{k,l}+1)^{m-1} \\
&= M_{k,l}^m + \frac{(1-r)m}{l}M_{k,l}^m + \frac{(1-r)m}{l}\left[(M_{k,l}+1)^{m-1} - M_{k,l}^{m-1}\right]M_{k,l}.
\end{aligned}$$

Using the integration inequality again this is

$$\begin{aligned}
&\le M_{k,l}^m\left(1 + \frac{(1-r)m}{l}\right) + \frac{(1-r)m}{l}\left[(m-1)(M_{k,l}+1)^{m-2}\right]M_{k,l} \\
&\le M_{k,l}^m\left(1 + \frac{(1-r)m}{l}\right) + \frac{m(m-1)}{l}(M_{k,l}+1)^{m-1}.
\end{aligned}$$

Since $M_{k,l} \ge 1$, we have

$$E[M_{k,l+1}^m | M_{k,l}] \le M_{k,l}^m\left(1 + \frac{(1-r)m}{l}\right) + \frac{C}{l}M_{k,l}^{m-1}. \tag{4.7}$$

We now establish the lemma for integer values of m by induction. When $m = 1$, the result is an immediate consequence of Lemma 3.2 and the inequality $e^{r/k} \le 1 + C/k$. Suppose the result holds for $m - 1$. Then, since $M_{k,l} = lX_{k,l}$, we have

$$\begin{aligned} E[M_{k,l}^{m-1}] &= l^{m-1} E\left[\left(\frac{M_{k,l}}{l}\right)^{m-1}\right] \\ &\le l^{m-1}\left(\frac{k}{l}\right)^{(m-1)r}\left(1 + \frac{C}{k}\right) \le Ck^{(m-1)r} l^{(m-1)(1-r)}. \end{aligned}$$

Therefore, taking expectations of both sides in (4.7), we get

$$E[M_{k,l+1}^m] \le \left(1 + \frac{(1-r)m}{l}\right) E[M_{k,l}^m] + Ck^{(m-1)r} l^{(m-1)(1-r)-1}.$$

Since $M_{k,k} = k$, iterating the last result shows that $E[X_k^m] = E[M_{k,N}]/N^m$ is at most

$$\frac{1}{N^m}\left[k^m \prod_{j=k}^{N-1}\left(1 + \frac{(1-r)m}{j}\right) + \sum_{l=k}^{N-1} Ck^{(m-1)r} l^{(m-1)(1-r)-1}\left(\prod_{j=l+1}^{N-1}\left(1 + \frac{(1-r)m}{j}\right)\right)\right].$$

Since $1 + x \le e^x$ for $x > 0$, we have

$$\begin{aligned} \prod_{j=k}^{N-1}\left(1 + \frac{(1-r)m}{j}\right) &\le \exp\left(\sum_{j=k}^{N-1} \frac{(1-r)m}{j}\right) \le \exp\left((1-r)m\left(\frac{1}{k} + \int_k^N x^{-1}\, dx\right)\right) \\ &= \exp\left(\frac{(1-r)m}{k} + (1-r)m \log\left(\frac{N}{k}\right)\right) \le \left(\frac{N}{k}\right)^{(1-r)m}\left(1 + \frac{C}{k}\right). \end{aligned}$$

Thus,

$$\begin{aligned} E[X_k^m] &\le \frac{1}{N^m}\left[k^m\left(\frac{N}{k}\right)^{(1-r)m}\left(1 + \frac{C}{k}\right) + C\sum_{l=k}^{N-1} k^{(m-1)r} l^{(m-1)(1-r)-1}\left(\frac{N}{l}\right)^{(1-r)m}\right] \\ &= \left(\frac{k}{N}\right)^{mr}\left[1 + \frac{C}{k} + Ck^{-r}\sum_{l=k}^{N-1} l^{-2+r}\right] \le \left(\frac{k}{N}\right)^{mr}\left[1 + \frac{C}{k}\right]. \end{aligned}$$

The result for integer values of m follows by induction.

Now suppose $n < m < n + 1$, where n is a positive integer. Let $p = (n - m + 1)^{-1}$ and let $q = (m - n)^{-1}$. Note that $p^{-1} + q^{-1} = 1$ and $n/p + (n+1)/q = m$. By Hölder's Inequality,

$$\begin{aligned} E[X_k^m] = E[X_k^{n/p} X_k^{(n+1)/q}] &\le E[X_k^n]^{n-m+1} E[X_k^{n+1}]^{m-n} \\ &\le \left(\frac{k}{N}\right)^{mr}\left(1 + \frac{C}{k}\right), \end{aligned}$$

so the lemma is true for all real numbers $m \ge 1$. □

To prove Theorem 1.3, we will approximate the family sizes NV_kX_k by $NW_k(k/N)^r$. To use this approximation, we will need a bound on the probability that the difference between these two quantities is large. Note that

$$V_kX_k - W_k\left(\frac{k}{N}\right)^r = X_k(V_k - W_k) + W_k\left(X_k - \left(\frac{k}{N}\right)^r\right). \tag{4.8}$$

The next two lemmas deal separately with the two terms on the right-hand side of (4.8).

Lemma 4.4. *There is a positive constant C so that for all $\delta > 0$*

$$\sum_{k=1}^{N} P\left(\left|W_k\left(X_k - \left(\frac{k}{N}\right)^r\right)\right| > \frac{\delta S}{2N}\right) \le C\left(\frac{N^{1-r}}{\delta S}\right)^{2/(3-2r)}.$$

Proof. Conditioning on $\mathcal{F}_k$ and noting that W_k is independent of $\mathcal{F}_k$ gives

$$E\left[W_k^2\left(X_k - \left(\frac{k}{N}\right)^r\right)^2\right] = E[W_k^2]E\left[\left(X_k - \left(\frac{k}{N}\right)^r\right)^2\right].$$

If $k \ge 2$, Lemmas 3.2 and 4.3 give that the above is equal to

$$\begin{aligned}
&\frac{2r}{k(k+1)}\left(E[X_k^2] - 2E[X_k]\left(\frac{k}{N}\right)^r + \left(\frac{k}{N}\right)^{2r}\right) \\
&\quad\le \frac{2r}{k^2}\left[\left(\frac{k}{N}\right)^{2r}\left(1 + \frac{C}{k}\right) - 2\left(\frac{k}{N}\right)^{2r} e^{-r^2/k} + \left(\frac{k}{N}\right)^{2r}\right] \\
&\quad\le \frac{2r}{k^2}\left(\frac{k}{N}\right)^{2r}\left[1 + \frac{C}{k} - 2\left(1 - \frac{r^2}{k}\right) + 1\right] \le \frac{C}{N^{2r}k^{3-2r}}.
\end{aligned}$$

Fix a positive integer L. Using a trivial inequality for $k \le L$ and Chebyshev's Inequality,

$$\begin{aligned}
\sum_{k=1}^{N} P\left(\left|W_k\left(X_k - \left(\frac{k}{N}\right)^r\right)\right| > \frac{\delta S}{2N}\right) &\le L + \sum_{k=L+1}^{N} \frac{C}{N^{2r}k^{3-2r}}\left(\frac{\delta S}{2N}\right)^{-2} \\
&\le L + \frac{CN^{2-2r}}{(\delta S)^2}\int_L^\infty \frac{1}{x^{3-2r}}\,dx = L + \frac{CN^{2-2r}L^{-(2-2r)}}{(2-2r)(\delta S)^2}.
\end{aligned} \tag{4.9}$$

If $L = (N^{1-r}/\delta S)^{2/(3-2r)}$, then the right-hand side of (4.9) is bounded by

$$\left(\frac{N^{1-r}}{\delta S}\right)^{2/(3-2r)} + C\left(\frac{N^{1-r}}{\delta S}\right)^{2-(2-2r)2/(3-2r)} \le C\left(\frac{N^{1-r}}{\delta S}\right)^{2/(3-2r)},$$

as claimed. □

Lemma 4.5. *There is a constant C so that for all $\delta > 0$, we have*

$$\sum_{k=1}^{N} P\left(|X_k(V_k - W_k)| > \frac{\delta S}{2N}\right) \le 1 + \frac{CN}{(\delta S)^{3/2(1-r)}}.$$

Proof. Recall from Section 2 that $W_k = \xi_k \tilde{W}_k$, where $\xi_k = 1_{\{W_k>0\}}$ has a Bernoulli(r) distribution and $\tilde{W}_k$ has a Beta$(1, k-1)$ distribution and is independent of ξ_k. Also recall that $\tilde{V}_k = (NV_kX_k - 1)1_{\{W_k>0\}}$ is a random variable such that the conditional distribution of $\tilde{V}_k$ given $\mathcal{G}_k$ is Binomial$(NX_k - k, \tilde{W}_k)$. Using (2.3), we see that for all $k \geq 2$ we have

$$\begin{aligned} P\left(|X_k(V_k - W_k)| > \frac{\delta S}{2N}\right) &= P\left(|NX_k(V_k - W_k)|1_{\{W_k>0\}} > \frac{\delta S}{2}\right) \\ &= P\left(|(1 - k\tilde{W}_k) + (\tilde{V}_k - \tilde{W}_k(NX_k - k))|1_{\{W_k>0\}} > \frac{\delta S}{2}\right) \\ &\leq P\left(|1 - k\tilde{W}_k| > \frac{\delta S}{4}\right) + P\left(|\tilde{V}_k - \tilde{W}_k(NX_k - k)| > \frac{\delta S}{4}\right). \quad (4.10) \end{aligned}$$

Let $m = 3/2(1-r)$. The reason for this choice will become clear in (4.14). Until then the reader should keep in mind that m is a fixed real number. Since $\Gamma(x+1) = x\Gamma(x)$ for all real x, we have $\Gamma(k)/\Gamma(m+k) \leq Ck^{-m}$ for some constant C. Therefore,

$$E[\tilde{W}_k^m] = \frac{\Gamma(k)\Gamma(m+1)}{\Gamma(m+k)} \leq Ck^{-m}, \quad (4.11)$$

so using $(a+b)^m \leq 2^m(a^m + b^m)$ for $a, b \geq 0$, we have

$$E[(1 + k\tilde{W}_k)^m] \leq 2^m(1 + E[(k\tilde{W}_k)^m]) \leq C.$$

Therefore, by Markov's Inequality, if $k \geq 2$ then

$$P\left(|1 - k\tilde{W}_k| > \frac{\delta S}{4}\right) \leq P\left(|1 + k\tilde{W}_k| > \frac{\delta S}{4}\right) \leq \left(\frac{\delta S}{4}\right)^{-m} E[(1 + k\tilde{W}_k)^m] \leq \frac{C}{(\delta S)^m}, \quad (4.12)$$

which bounds the first term on the right-hand side of (4.10).

Because of the restriction $np \geq 1$ in Lemma 4.1, we must split the second term in (4.10) into two pieces, depending on the value of $\tilde{W}_k(NX_k - k)$. Let V_k' be a random variable such that, conditional on $\mathcal{G}_k$, the distribution of V_k' is Binomial$(NX_k - k, 1/(NX_k - k))$. We set $V_k' = 0$ if $NX_k - k = 0$. Note that when $\tilde{W}_k(NX_k - k) < 1$, the conditional distribution of V_k' given $\mathcal{G}_k$ stochastically dominates the conditional distribution of $\tilde{V}_k$ given $\mathcal{G}_k$. By Lemma 4.1, $E[|V_k' - 1|^m|\mathcal{G}_k] \leq C$. Note also that $|\tilde{V}_k - \tilde{W}_k(NX_k - k)| = 0$ on the event $\{NX_k - k = 0\}$. Therefore, if $k \geq 2$ then

$$\begin{aligned} &P\left(|\tilde{V}_k - \tilde{W}_k(NX_k - k)|1_{\{\tilde{W}_k(NX_k-k)<1\}} > \frac{\delta S}{4}\right) \\ &\quad = E\left[E\left[P\left(|\tilde{V}_k - \tilde{W}_k(NX_k - k)|1_{\{\tilde{W}_k(NX_k-k)<1\}} > \frac{\delta S}{4}\right)\middle|\mathcal{G}_k\right]\right] \\ &\quad \leq E\left[E\left[P\left(|V_k' - 1| + 1 > \frac{\delta S}{4}\right)\middle|\mathcal{G}_k\right]\right] \\ &\quad \leq \left(\frac{\delta S}{4}\right)^{-m} E[E[(|V_k' - 1| + 1)^m|\mathcal{G}_k]] \\ &\quad \leq \left(\frac{\delta S}{4}\right)^{-m} 2^m(C+1) \leq \frac{C}{(\delta S)^m}. \quad (4.13) \end{aligned}$$

By Lemma 4.1, we get, for $k \geq 2$,

$$
\begin{aligned}
&P\bigg(|\tilde{V}_k - \tilde{W}_k(NX_k - k)|1_{\{\tilde{W}_k(NX_k-k)\geq 1\}} > \frac{\delta S}{4}\bigg) \\
&\qquad = E\bigg[E\bigg[P\bigg(\bigg|\frac{\tilde{V}_k}{NX_k - k} - \tilde{W}_k\bigg|1_{\{\tilde{W}_k(NX_k-k)\geq 1\}} > \frac{\delta S}{4(NX_k - k)}\bigg)\bigg|\mathcal{G}_k\bigg]\bigg] \\
&\qquad \leq E\bigg[\bigg(\frac{\delta S}{4(NX_k - k)}\bigg)^{-m} E\bigg[\bigg|\frac{\tilde{V}_k}{NX_k - k} - \tilde{W}_k\bigg|^m 1_{\{\tilde{W}_k(NX_k-k)\geq 1\}}\bigg|\mathcal{G}_k\bigg]\bigg] \\
&\qquad \leq E\bigg[\bigg(\frac{4(NX_k - k)}{\delta S}\bigg)^m C\bigg(\frac{\tilde{W}_k}{NX_k - k}\bigg)^{m/2}\bigg] \leq \frac{C}{(\delta S)^m} E[\tilde{W}_k^{m/2}(NX_k - k)^{m/2}].
\end{aligned}
$$

By conditioning on $\mathcal{F}_k$ and noting that W_k is independent of this σ-field, we see that this is at most

$$
\frac{C}{(\delta S)^m} E[\tilde{W}_k^{m/2}] E[(NX_k)^{m/2}] \leq \frac{C}{(\delta S)^m} \frac{C}{k^{m/2}} N^{m/2} \bigg(\frac{k}{N}\bigg)^{mr/2}
$$

by (4.11) and Lemma 4.3. Recalling that $m = 3/2(1-r)$, the above is at most

$$
\frac{C}{(\delta S)^m} \bigg(\frac{N}{k}\bigg)^{m(1-r)/2} = \frac{C}{(\delta S)^m} \bigg(\frac{N}{k}\bigg)^{3/4}. \tag{4.14}
$$

Note that

$$
\sum_{k=2}^{N} \bigg(\frac{N}{k}\bigg)^{3/4} = N^{3/4} \sum_{k=2}^{N} k^{-3/4} \leq C N^{3/4} N^{1/4} = CN.
$$

Combining this fact with (4.10), (4.12), (4.13), and (4.14), which hold for $k \geq 2$, we get

$$
\sum_{k=1}^{N} P\bigg(|X_k(V_k - W_k)| > \frac{\delta S}{2N}\bigg) \leq 1 + \frac{CN}{(\delta S)^m},
$$

which completes the proof. □

Now that we have shown that $V_k X_k$ and $W_k(k/N)^r$ are close with high probability, the next step is to calculate the probability that $W_k(k/N)^r$ is large. The next two lemmas provide upper and lower bounds on the probability that $W_k(k/N)^r$ is large. Recall from (1.3) that

$$
g(S) = r\Gamma\bigg(\frac{2-r}{1-r}\bigg) N S^{-1/(1-r)}.
$$

Lemma 4.6. *There is a constant C so that for $0 < \delta < 1/2$ and $S \leq \frac{1}{2}N^{1-r}$,*

$$
\sum_{k=1}^{N} P\bigg(W_k\bigg(\frac{k}{N}\bigg)^r \geq \frac{(1-\delta)S}{N}\bigg) \leq C + g(S)(1 + C\delta).
$$

Proof. The C on the left takes care of the term $k = 1$. Since the conditional distribution of W_k given $W_k > 0$ is Beta$(1, k-1)$, we have, for all $k \geq 2$ and $a \in (0,1)$,

$$P(W_k \geq a) = r \int_a^1 (k-1)(1-x)^{k-2}\, dx = r(1-a)^{k-1}.$$

Using the facts that $(1 - a/x)^x \leq e^{-a}$ if $0 \leq a \leq x$ and $1/(1-x) \leq 1 + 2x$ if $0 \leq x \leq 1/2$, we have, for $k \geq 2$,

$$\begin{aligned} P\biggl(W_k \geq \frac{S(1-\delta)}{k^r N^{1-r}}\biggr) &= r\biggl(1 - \frac{S(1-\delta)k^{1-r}}{kN^{1-r}}\biggr)^k \biggl(1 - \frac{S(1-\delta)}{k^r N^{1-r}}\biggr)^{-1} \\ &\leq re^{-S(1-\delta)(k/N)^{1-r}} \biggl(1 + \frac{2S}{k^r N^{1-r}}\biggr). \end{aligned} \tag{4.15}$$

Note that

$$\sum_{k=2}^{N} e^{-S(1-\delta)(k/N)^{1-r}} \biggl(1 + \frac{2S}{k^r N^{1-r}}\biggr) \leq \int_0^\infty e^{-S(1-\delta)(x/N)^{1-r}} \biggl(1 + \frac{2S}{x^r N^{1-r}}\biggr)\, dx. \tag{4.16}$$

Letting $y = S(1-\delta)(x/N)^{1-r}$, which means that $x = y^{1/(1-r)}M$ where $M = N(S(1-\delta))^{-1/(1-r)}$ and $dx = y^{1/(1-r)-1}M/(1-r)\, dy$ we have

$$\begin{aligned} \int_0^\infty e^{-S(1-\delta)(x/N)^{1-r}}\, dx &= \int_0^\infty e^{-y} y^{1/(1-r)-1} \biggl(\frac{M}{1-r}\biggr)\, dy \\ &= \Gamma\biggl(\frac{1}{1-r}\biggr)\frac{M}{1-r} = \Gamma\biggl(\frac{2-r}{1-r}\biggr) N(S(1-\delta))^{-1/(1-r)}. \end{aligned} \tag{4.17}$$

The same change of variables gives

$$\begin{aligned} \int_0^\infty e^{-S(1-\delta)(x/N)^{1-r}} \biggl(\frac{2S}{x^r N^{1-r}}\biggr)\, dx &= \frac{2S}{N^{1-r}} \int_0^\infty e^{-y} \bigl(y^{1/(1-r)}M\bigr)^{-r} y^{1/(1-r)-1} \biggl(\frac{M}{1-r}\biggr)\, dy \\ &= \frac{2SM^{1-r}}{(1-r)N^{1-r}} \int_0^\infty e^{-y}\, dy = \frac{2}{(1-r)(1-\delta)} \leq C. \end{aligned} \tag{4.18}$$

Because $(1-\delta)^{-1/(1-r)} \leq 1 + C\delta$, the claim follows from (4.15), (4.16), (4.17), and (4.18). □

Lemma 4.7. *There is a constant C so that for $0 < \delta < 1/2$ and $S \leq \frac{1}{3}N^{1-r}$,*

$$\sum_{k=1}^{N} P\biggl(W_k \biggl(\frac{k}{N}\biggr)^r \geq \frac{(1+\delta)S}{N}\biggr) \geq -C + g(S)\bigl(1 - C\delta - Ce^{-S/2}\bigr).$$

Proof. Recall from the beginning of the proof of Lemma 3.2 that if $0 \leq x \leq 1/2$, then $\log(1-x) \geq -(x + x^2)$. It follows that if $0 \leq a/y \leq 1/2$ and $a \geq 0$, then $y\log(1 - a/y) \geq -a - a^2/y$, and so

$$\biggl(1 - \frac{a}{y}\biggr)^y \geq e^{-a}e^{-a^2/y} \geq e^{-a}\biggl(1 - \frac{a^2}{y}\biggr).$$

Therefore, if $k \geq 2$ then

$$P\left(W_k\left(\frac{k}{N}\right)^r \geq \frac{(1+\delta)S}{N}\right) = r\left(1 - \frac{S(1+\delta)}{k^r N^{1-r}}\right)^{k-1} \geq r\left(1 - \frac{S(1+\delta)k^{1-r}}{kN^{1-r}}\right)^k$$
$$\geq re^{-S(1+\delta)(k/N)^{1-r}}\left(1 - \frac{S^2(1+\delta)^2 k^{1-2r}}{N^{2-2r}}\right). \tag{4.19}$$

We have

$$\sum_{k=2}^{N} e^{-S(1+\delta)(k/N)^{1-r}} \geq \left(\int_0^\infty e^{-S(1+\delta)(x/N)^{1-r}}\,dx\right) - 2 - \int_N^\infty e^{-S(1+\delta)(x/N)^{1-r}}\,dx. \tag{4.20}$$

It follows from (4.17) with δ replaced by $-\delta$ and $M = N(S(1+\delta))^{-1/(1-r)}$ that

$$\int_0^\infty e^{-S(1+\delta)(x/N)^{1-r}}\,dx = \Gamma\left(\frac{2-r}{1-r}\right)M \geq \Gamma\left(\frac{2-r}{1-r}\right)NS^{-1/(1-r)}(1 - C\delta). \tag{4.21}$$

To estimate the second term in (4.19) we note that

$$\sum_{k=2}^{N} e^{-S(1+\delta)(k/N)^{1-r}}\left(\frac{S^2(1+\delta)^2 k^{1-2r}}{N^{2-2r}}\right) \leq \frac{CS^2}{N^{2-2r}}\int_0^\infty e^{-S(1+\delta)(x/N)^{1-r}} x^{1-2r}\,dx.$$

Making the change of variables $x = y^{1/(1-r)}M$ and reasoning as in (4.18), we see that this equals

$$\frac{CS^2}{N^{2-2r}}\int_0^\infty e^{-y}(y^{1/(1-r)}M)^{1-2r} y^{1/(1-r)-1}\left(\frac{M}{1-r}\right)dy$$
$$= \frac{C}{(1+\delta)^2}\int_0^\infty e^{-y} y\,dy \leq C. \tag{4.22}$$

Finally, by using our favorite change of variables and using Lemma 4.2, we get

$$\int_N^\infty e^{-S(1+\delta)(x/N)^{1-r}}\,dx = \int_{S(1+\delta)}^\infty e^{-y} y^{1/(1-r)-1}\left(\frac{M}{1-r}\right)dy$$
$$\leq \frac{M}{1-r}\int_S^\infty e^{-y} y^{1/(1-r)-1}\,dy \leq CNS^{-1/(1-r)}e^{-S/2}. \tag{4.23}$$

The Lemma now follows by combining (4.19), (4.20), (4.21), (4.22), and (4.23). □

Proof of Theorem 1.3. Let A_k be the event that $\{NX_kV_k \geq S\}$, so $F_{S,N} = \sum_{k=1}^N 1_{A_k}$. First, note that if the theorem is true for $S = \frac{1}{3}N^{1-r}$, then we know that $E[F_{S,N}] \leq C$ for all $S > \frac{1}{3}N^{1-r}$, which implies the assertion in the theorem. Therefore, it suffices to prove the result for $S \leq \frac{1}{3}N^{1-r}$, in which case the conclusions of Lemmas 4.6 and 4.7 will hold as long as we choose $\delta < 1/2$. Let $A_k^- = \{NW_k(k/N)^r \geq (1-\delta)S\}$ and let $A_k^+ = \{NW_k(k/N)^r \geq (1+\delta)S\}$. Let $F_S^- = \sum_{k=1}^N 1_{A_k^-}$ and $F_S^+ = \sum_{k=1}^N 1_{A_k^+}$. Writing F_S for $F_{S,N}$, we have

$$|F_S - g(S)| \leq |F_S - F_S^-| + |F_S^- - E[F_S^-]| + |E[F_S^-] - g(S)|. \tag{4.24}$$

To prove the theorem, we will bound the expectations of the three terms on the right-hand side of (4.24).

Note that $A_k^+ \subset A_k^-$ for all k and $A_k \triangle A_k^- \subset (A_k^- \setminus A_k^+) \cup \{|V_k X_k - W_k(k/N)^r| > \delta S/N\}$. By Lemmas 4.6 and 4.7, we have

$$\sum_{k=1}^{N} P(A_k^- \setminus A_k^+) \le C + Cg(S)\Big\{\delta + e^{-S/2}\Big\} \tag{4.25}$$

By (4.8) and Lemmas 4.4 and 4.5, we have

$$\begin{aligned}\sum_{k=1}^{N} P\bigg(\bigg|V_k X_k - W_k\Big(\frac{k}{N}\Big)^r\bigg| > \frac{\delta S}{N}\bigg) &\le 1 + C\bigg(\frac{N^{1-r}}{\delta S}\bigg)^{2/(3-2r)} + C\bigg(\frac{N}{(\delta S)^{3/2(1-r)}}\bigg) \\ &\le 1 + Cg(S)\bigg\{\frac{(NS^{-1/(1-r)})^{-1/(3-2r)}}{\delta^{2/(3-2r)}} + \frac{S^{-1/2(1-r)}}{\delta^{3/2(1-r)}}\bigg\}. \end{aligned} \tag{4.26}$$

Combining the last two results, we get

$$E[|F_S - F_S^-|] \le C + Cg(S)(D_1 + D_2), \tag{4.27}$$

where D_1 and D_2 are the terms in braces in (4.25) and (4.26). To bound the second term of (4.24), we use Jensen's Inequality and the fact that the A_k^- are independent to get

$$\begin{aligned} E[|F_S^- - E[F_S^-]|] &\le E[(F_S^- - E[F_S^-])^2]^{1/2} = \mathrm{Var}(F_S^-)^{1/2} = \bigg[\sum_{k=1}^{N} \mathrm{Var}(1_{A_k^-})\bigg]^{1/2} \\ &\le \bigg[\sum_{k=1}^{N} P(A_k^-)\bigg]^{1/2} \le Cg(S)^{1/2} \le Cg(S)\big(NS^{-1/(1-r)}\big)^{-1/2}. \end{aligned} \tag{4.28}$$

Furthermore, note that since Lemma 4.6 gives an upper bound for $E[F_S^-]$ that is greater than $g(S)$ and Lemma 4.7 gives a lower bound for $E[F_S^+]$ that is smaller than $g(S)$, the difference $|E[F_S^-] - g(S)|$ is less than or equal to the difference between these two bounds, which itself was bounded in (4.25). Combining this observation with (4.24), (4.27), and (4.28), we see that

$$E[|F_S - g(S)|] \le C + Cg(S)(D_1 + D_2 + (NS^{-1/(1-r)})^{-1/2}).$$

To prove the theorem, we need to show that each part of $D_1 + D_2$ is bounded by S^{-a} or $(NS^{-1/(1-r)})^{-b}$ for some positive constants a and b. Letting $R = NS^{-1/(1-r)}$ to simplify notation, it is enough to bound

$$\delta + \frac{R^{-1/(3-2r)}}{\delta^{2/(3-2r)}} + \frac{S^{-1/2(1-r)}}{\delta^{3/2(1-r)}}.$$

To do this, we let $\delta = A(S^{-c} + R^{-d})$, where $3c < 1$ and $2d < 1$, and choose A to ensure that $\delta < 1/2$. To optimize the bound we set $(1-3c)/2(1-r) = c$ and $(1-2d)/(3-2r) = d$. Solving gives $c = 1/(5-2r)$ and $d = 1/(5-2r)$. □

5 Sizes of the largest families

In this section, we study the largest families, whose sizes are $O(N^{1-r})$, and we prove Propositions 1.4 and 1.5. The key to our arguments is the following well-known result about Yule processes, which is discussed in Chapter III of Athreya and Ney (1972). Suppose $(X(t), t \geq 0)$ is a Yule process started with one individual at time zero in which each individual splits into two at rate λ. Then, there exists a random variable W such that

$$\lim_{t\to\infty} e^{-\lambda t} X(t) = W \quad a.s. \tag{5.1}$$

and W has an exponential distribution with mean 1. A corollary of this fact is that if $X(0) = m$ then $e^{-\lambda t}X(t)$ converges a.s. to a limiting random variable whose distribution is Gamma$(m, 1)$.

Lemma 5.1. *Fix a positive integer m, and suppose $k \leq 2(m+1)$. There exists a positive constant C such that*

$$\limsup_{N\to\infty} P(R_{k,N} \geq AN^{1-r}) \leq \frac{C}{A^{m+1}}.$$

Proof. Let $J_k(t)$ be the number of individuals of type k in the population at time t, so that $R_{k,N} = J_k(T_N)$. Because each individual of type k gives birth to another individual of type k at rate $1 - r$, the process $(J_k(t + T_{2(m+1)}), t \geq 0)$ is a Yule process. By (5.1), there is a random variable Z_1 such that, as $t \to \infty$, we have $e^{-(1-r)t} J_k(t + T_{2(m+1)}) \to Z_1$ almost surely. By the strong Markov property of $(J_k(t), t \geq 0)$, the conditional distribution of Z_1 given $J_k(T_{2(m+1)})$ is Gamma$(J_k(T_{2(m+1)}), 1)$. Let $K(t)$ denote the total population size at time t. Because each individual gives birth at rate 1, the process $(K(t + T_{2(m+1)}), t \geq 0)$ is also a Yule process. As $t \to \infty$, we have $e^{-t}K(t + T_{2(m+1)}) \to Z_2$ almost surely, where Z_2 has a Gamma$(2(m + 1), 1)$ distribution. Note that

$$\lim_{N\to\infty} \frac{R_{k,N}}{N^{1-r}} = \lim_{t\to\infty} \frac{J_k(t + T_{2(m+1)})}{[K(t + T_{2(m+1)})]^{1-r}} = \lim_{t\to\infty} \frac{e^{-(1-r)t} J_k(t + T_{2(m+1)})}{[e^{-t}K(t + T_{2(m+1)})]^{1-r}} = \frac{Z_1}{Z_2^{1-r}} \quad \text{a.s.}$$

Therefore, using the Portmanteau Theorem for the first inequality,

$$\limsup_{N\to\infty} P(R_{k,N} \geq AN^{1-r}) \leq P\left(\frac{Z_1}{Z_2^{1-r}} \geq A\right) \leq P(Z_1 \geq A^{1/2}) + P(Z_2^{1-r} \leq A^{-1/2}). \tag{5.2}$$

Note that since $J_k(T_{2(m+1)}) \leq 2(m + 1)$, the distribution of Z_1 is stochastically dominated by the Gamma$(2(m + 1), 1)$ distribution. Therefore, by Lemma 4.2, we have $P(Z_1 \geq A^{1/2}) \leq 2^{2(m+1)} e^{-A^{1/2}/2}$. Furthermore, if $A \geq 1$ then

$$P(Z_2^{1-r} \leq A^{-1/2}) \leq P(Z_2 \leq A^{-1/2}) \leq \frac{A^{-(m+1)}}{\Gamma(2m + 3)} \leq \frac{1}{A^{m+1}}.$$

These bounds, combined with (5.2), prove the lemma for $A \geq 1$. The result for $A < 1$ is obvious. □

Lemma 5.2. *Fix a positive integer m, and suppose $k > 2(m+1)$. There exists a positive constant C, depending on m but not on k, such that*

$$\limsup_{N\to\infty} P(R_{k,N} \geq AN^{1-r}) \leq \frac{C}{A^{k/2}}.$$

Proof. The proof is similar to the proof of Lemma 5.1. Let $J_k(t)$ be the number of individuals of type k in the population at time t, and let $K(t)$ be the population size at time t. The process $(J_k(t+T_k), t \geq 0)$ is a Yule process started with either 0 or 1 individuals, and $(K(t+T_k), t \geq 0)$ is a Yule process started with k individuals. Therefore, as $t \to \infty$, we have $e^{-(1-r)t}J_k(t+T_k) \to Z_1$ a.s. and $e^{-t}K(t+T_k) \to Z_2$ a.s., where Z_1 has an exponential distribution conditional on $J_k(T_k) = 1$ and is zero otherwise, and Z_2 has a Gamma$(k, 1)$ distribution. We have

$$\lim_{N\to\infty} \frac{R_{k,N}}{N^{1-r}} = \lim_{t\to\infty} \frac{J_k(t+T_k)}{[K(t+T_k)]^{1-r}} = \lim_{t\to\infty} \frac{e^{-(1-r)t}J_k(t+T_k)}{[e^{-t}K(t+T_k)]^{1-r}} = \frac{Z_1}{Z_2^{1-r}} \quad \text{a.s.}$$

so (5.2) holds again. Since $k > 1$ we have $P(J_k(T_k) = 1) = r$ and so $P(Z_1 \geq A^{-1/2}) = re^{-A^{1/2}}$. By Lemma 4.2, if $A \geq 1$ then $P(Z_2^{1-r} \leq A^{-1/2}) \leq P(Z_2 \leq A^{-1/2}) \leq A^{-k/2}$. These bounds and (5.2) prove the lemma for $A \geq 1$, and the case of $A < 1$ is clear. □

Proof of Proposition 1.4. By Lemmas 5.1 and 5.2, there is a positive constant C such that if $A > 1$ then

$$\begin{aligned}
\limsup_{N\to\infty} A^m E[F_{AN^{1-r},N}] &= \limsup_{N\to\infty} \sum_{k=1}^{N} A^m P(R_{k,N} \geq AN^{1-r}) \\
&\leq \limsup_{N\to\infty} \left(2(m+1)A^m \frac{C}{A^{m+1}} + A^m \sum_{k=2m+3}^{N} \frac{C}{A^{k/2}} \right) \\
&\leq \frac{2(m+1)C}{A} + \frac{CA^{-3/2}}{1 - A^{-1/2}},
\end{aligned}$$

and the proposition follows by letting $A \to \infty$. □

Proof of Proposition 1.5. Define $J_k(t)$ and $K(t)$ as in the proof of Lemma 5.2. Let Z_1 and Z_2 be the almost sure limits as $t \to \infty$ of $e^{-(1-r)t}J_k(t+T_k)$ and $e^{-t}K(t+T_k)$ respectively, and let $Z = Z_1 Z_2^{r-1}$. Then, again following the proof of Lemma 5.2, the random variables $N^{r-1}R_{k,N}$ converge almost surely to Z as $N \to \infty$. We have $Z_1 = 0$, and thus $Z = 0$, if $J_k(T_k) = 0$. However, we have $P(Z > 0 | J_k(T_k) = 1) = 1$ because, conditional on $J_k(T_k) = 1$, the random variable Z_1 has an exponential distribution with mean 1. If $k = 1$ then $P(Z > 0) = P(J_k(T_k) = 1) = 1$ but for $k \geq 2$, we have $P(Z > 0) = P(J_k(T_k) = 1) = r$.

It remains only to establish the joint moment generating function for (Z_1, Z_2), conditional on $Z > 0$. To do this, we will replace our original branching process with infinitely many types by a two-type branching process. In the new branching process, individuals of type A will correspond to individuals of type k in the original branching process, and individuals of type B will correspond to individuals of all other types in the original branching process. We will denote the new branching process by $((L_A(t), L_B(t)), t \geq 0)$, and we will start the process at time T_k,

so that $L_A(t) = J_k(t + T_k)$ and $L_B(t) = K(t + T_k) - J_k(t + T_k)$. Conditional on $Z > 0$, the new branching process starts with 1 individual of type A and $k - 1$ individuals of type B.

We now follow the development in section 7 of chapter V of Athreya and Ney (1972). To fit our example into this framework, we can think of each individual as living for an exponential(1) amount of time. When a type A individual dies, it gives birth to two type A individuals with probability $1-r$ and one type A and one type B with probability r. When a type B individual dies, it gives birth to two type B individuals. Therefore, the offspring distribution can be written as $p_A(2, 0) = 1-r$, $p_A(1, 1) = r$, and $p_B(0, 2) = 1$. The corresponding generating functions are given by $f_A(s_1, s_2) = (1 - r)s_1^2 + rs_1s_2$ and $f_B(s_1, s_2) = s_2^2$. Define also $u_A(s_1, s_2) = f_A(s_1, s_2) - s_1 = (1 - r)s_1^2 + rs_1s_2 - s_1$ and $u_B(s_1, s_2) = f_B(s_1, s_2) - s_2 = s_2^2 - s_2$.

Let $E_{(a,b)}$ denote expectation when the two-type branching process starts with a individuals of type A and b individuals of type B. For $s_1, s_2 \in [0, 1]$, define $G_A(s_1, s_2, t) = E_{(1,0)}[s_1^{L_A(t)}s_2^{L_B(t)}]$ and $G_B(s_1, s_2, t) = E_{(0,1)}[s_1^{L_A(t)}s_2^{L_B(t)}]$. Kolmogorov's backward equations (see (2) on p. 201 of Athreya and Ney (1972)) give

$$\begin{aligned}\frac{\partial}{\partial t}G_A(s_1, s_2, t) &= u_A(G_A(s_1, s_2, t), G_B(s_1, s_2, t)) \\ &= (1 - r)G_A(s_1, s_2, t)^2 + rG_A(s_1, s_2, t)G_B(s_1, s_2, t) - G_A(s_1, s_2, t) \qquad (5.3)\end{aligned}$$

and

$$\frac{\partial}{\partial t}G_B(s_1, s_2, t) = u_B(G_A(s_1, s_2, t), G_B(s_1, s_2, t)) = G_B(s_1, s_2, t)^2 - G_B(s_1, s_2, t). \qquad (5.4)$$

For fixed s_1 and s_2, the equations (5.3) and (5.4) become ordinary differential equations. Let $x(t) = G_A(s_1, s_2, t)$ and $y(t) = G_B(s_1, s_2, t)$. To find $G_B(s_1, s_2, t)$, we need to solve the equation $y'(t) = y(t)^2 - y(t)$ with the initial condition $y(0) = s_2$. The unique solution is

$$y(t) = G_B(s_1, s_2, t) = \frac{s_2e^{-t}}{s_2e^{-t} - s_2 + 1}.$$

Plugging this into (5.3), we see that, to find $G_A(s_1, s_2, t)$, we need to solve

$$x'(t) = (1 - r)x(t)^2 + r\left(\frac{s_2e^{-t}}{s_2e^{-t} - s_2 + 1}\right)x(t) - x(t), \qquad x(0) = s_1,$$

and one can check that the unique solution to this equation is given by

$$x(t) = G_A(s_1, s_2, t) = \frac{e^{-t}}{(s_2e^{-t} - s_2 + 1)^r}\left(\frac{1}{s_1} - \frac{1}{s_2} + \frac{(s_2e^{-t} - s_2 + 1)^{1-r}}{s_2}\right)^{-1}.$$

Because individuals in this branching process evolve independently, we have

$$E_{(a,b)}[s_1^{L_A(t)}s_2^{L_B(t)}] = [G_A(s_1, s_2, t)]^a[G_B(s_1, s_2, t)]^b.$$

Let $\phi_t(s_1, s_2) = E_{(1,k-1)}[e^{-s_1e^{-(1-r)t}L_A(t)-s_2e^{-t}L_B(t)}]$ for all $s_1 \geq 0$ and $s_2 \geq 0$, so that ϕ_t is the joint moment generating function of $(e^{-(1-r)t}L_A(t), e^{-t}L_B(t))$ when we start with one individual of type A and $k - 1$ individuals of type B. Then, for $s_1, s_2 > 0$, we have

$$\phi_t(s_1, s_2) = G_A(e^{-s_1e^{-(1-r)t}}, e^{-s_2e^{-t}}, t)[G_B(e^{-s_1e^{-(1-r)t}}, e^{-s_2e^{-t}}, t)]^{k-1}.$$

We know that $e^{-(1-r)t}L_A(t) \to Z_1$ a.s. Also, since $e^{-t}L_A(t) \to 0$ a.s., we have $e^{-t}L_B(t) \to Z_2$ a.s. Thus, by the Continuity Theorem for moment generating functions, to show that ϕ_t is the joint moment generating function of (Z_1, Z_2), we only need to show that for all $s_1 \geq 0$ and $s_2 \geq 0$, we have $\lim_{t\to\infty} \phi_t(s_1, s_2) = \phi(s_1, s_2)$. However, this is just a tedious but straightforward calculus exercise. By l'Hospital's Rule,

$$\lim_{t\to\infty} \frac{e^{-t}}{e^{-t}e^{-s_2e^{-t}} - e^{-s_2e^{-t}} + 1} = \frac{1}{1+s_2}. \tag{5.5}$$

From this, we see that

$$\lim_{t\to\infty} [G_B(e^{-s_1e^{-(1-r)t}}, e^{-s_2e^{-t}}, t)]^{k-1} = \lim_{t\to\infty} \left(\frac{e^{-s_2e^{-t}}e^{-t}}{e^{-s_2e^{-t}}e^{-t} - e^{-s_2e^{-t}} + 1} \right)^{k-1} = \left(\frac{1}{1+s_2} \right)^{k-1}.$$

Note that $G_A(e^{-s_1e^{-(1-r)t}}, e^{-s_2e^{-t}}, t)$ equals

$$\left(\frac{e^{-rt}}{(e^{-s_2e^{-t}}e^{-t} - e^{-s_2e^{-t}} + 1)^r} \right) \left(\frac{e^{-(1-r)t}}{\left(e^{s_1e^{-(1-r)t}} - e^{s_2e^{-t}} + e^{s_2e^{-t}}(1 - e^{-s_2e^{-t}} + e^{-s_2e^{-t}}e^{-t})^{1-r}\right)} \right). \tag{5.6}$$

By (5.5), the first factor in (5.6) goes to $1/(1+s_2)^r$ as $t \to \infty$. For the second factor, it is easiest to take the limit of the reciprocal. From (5.5) and the fact that

$$\lim_{t\to\infty} \frac{e^{s_1e^{-(1-r)t}} - e^{s_2e^{-t}}}{e^{-(1-r)t}} = s_1,$$

which can also be shown using l'Hospital's Rule, we get that the limit of the reciprocal of the second factor in (5.6) is $s_1 + (1+s_2)^{1-r}$. Thus,

$$\begin{aligned} \lim_{t\to\infty} \phi_t(s_1, s_2) &= \left(\frac{1}{1+s_2} \right)^{k-1} \frac{1}{(1+s_2)^r(s_1 + (1+s_2)^{1-r})} \\ &= \frac{1}{(1+s_2)^{k-1}(1+s_2+s_1(1+s_2)^r)} = \phi(s_1, s_2), \end{aligned}$$

which completes the proof of the proposition. □

References

R. Arratia, A. D. Barbour, and S. Tavaré (2000). Limits of logarithmic combinatorial structures. *Ann. Probab.* **28**, 1620–1644.

R. Arratia and L. Gordon (1989). Tutorial on large deviations for the binomial distribution. *Bulletin of Mathematical Biology* **51**, 125–131.

K. B. Athreya and S. Karlin (1968). Embedding of urn schemes into continuous time Markov branching processes. *Ann. Math. Statist.* **39**, 1801-1817.

K. B. Athreya and P. E. Ney (1972). *Branching Processes*. Springer-Verlag, Berlin.

R. Durrett (1996). *Probability: Theory and Examples*. 2nd ed. Duxbury, Belmont, CA.

R. Durrett (2002). *Probability models for DNA sequence evolution*. Springer-Verlag, New York.

R. Durrett and J. Schweinsberg (2004). Approximating selective sweeps. To appear in *Theor. Popul. Biol.* Available at http://www.math.cornell.edu/~jasonsch/webpage.html.

W. J. Ewens (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112.

Z. Gu, A. Cavalcanti, F.-C. Chen, P. Bouman, and W.-H. Li (2002). Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol. Biol. Evol.* **19**, 256–262

P. M. Harrison and M. Gerstein (2002). Studying genomes through the aeons: Protein families, pseudogenes, and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174

M. A. Hunyen, E. van Nimwegen (1998). The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15**, 583–589 S. Janson (2004). Functional limit theorems

for multitype branching processes and generalized Polya urns. *Stochastic Process. Appl.* **110**, 177–245.

N. L. Johnson, S. Kotz, and A. W. Kemp (1992). *Univariate Discrete Distributions*. Wiley, New York.

G. P. Karev, Y. I. Wolf, A. Y. Rzhetsky, F. S. Berezov, and E. V. Koonin (2002). Birth and death of protein domains: A simple model explains power law behavior. *BMC Evolutionary Biology.* **2**: article 18.

J. F. C. Kingman (1982). The coalescent. *Stochastic Process. Appl.* **13**, 235-248.

E. V. Koonin, Y. I. Wolf, and G. P. Karev (2002). The structure of the protein universe and genome evolution. *Nature.* **420**, 218–223

S. Kotz, N. Balakrishnan, and N. L. Johnson (2000). *Continuous Multivariate Distributions. Volume 1: Models and Applications.* 2nd ed. Wiley, New York.

W.-H. Li, Z. Gu, H. Wang, and A. Nekrutenko (2001). Evolutionary analyses of the human genome. *Nature.* **409**, 847–849

T. J. Pfaff (2003). A mean field model for species abundance. *Stochastic Process. Appl.* **104**, 325–347.

J. Pitman (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Relat. Fields*, **102**, 145–158.

J. Pitman (2002). *Combinatorial Stochastic Processes*. Lecture Notes for St. Flour Summer School, available at http://stat-www.berkeley.edu/users/pitman/bibliog.html.

J. Pitman and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

J. Qian, N. M. Luscombe, and M. Gerstein (2001). Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. *J. Mol. Biol.* **313**, 673–681.

A. Rzhetsky and S. M. Gomez (2001). Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics.* **17**, 988–996

J. Schweinsberg and R. Durrett (2003). Random partitions approximating the coalescence of lineages during a selective sweep. Preprint. Available at http://www.math.cornell.edu/~jasonsch/webpage.html.

Department of Mathematics
Malott Hall
Cornell University
Ithaca, NY 14853-4201
E-mail: jasonsch@math.cornell.edu
rtd1@cornell.edu

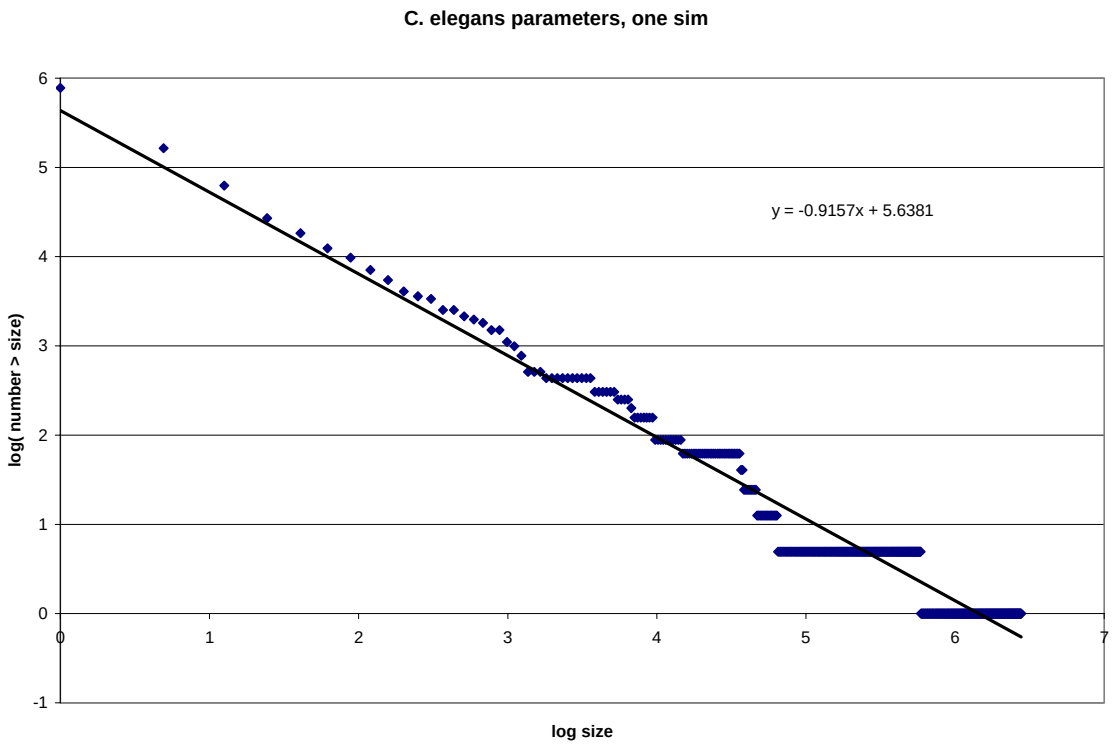

Figure 1. One simulation of the duplication model with C. elegans parameters. r = 0.018, and N = 20,000.

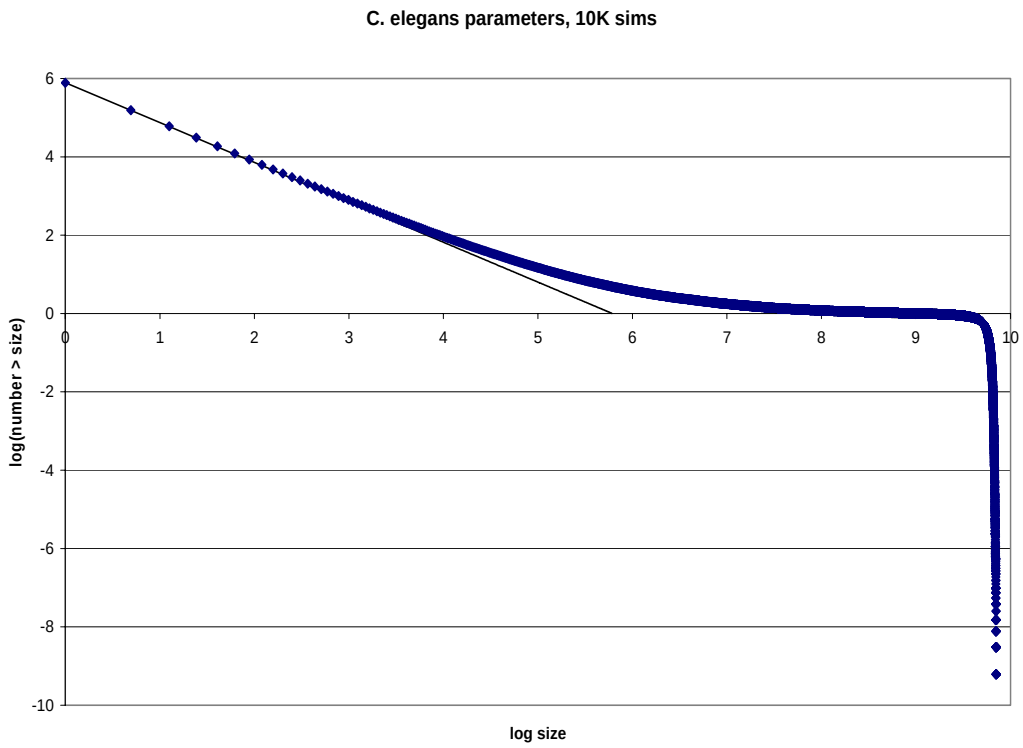

Figure 2. Average of 10,000 simulations of the duplication model with C. elegans parameters. r = 0.018, and N = 20,000. Straight line is the prediction of Theorem 1.3.

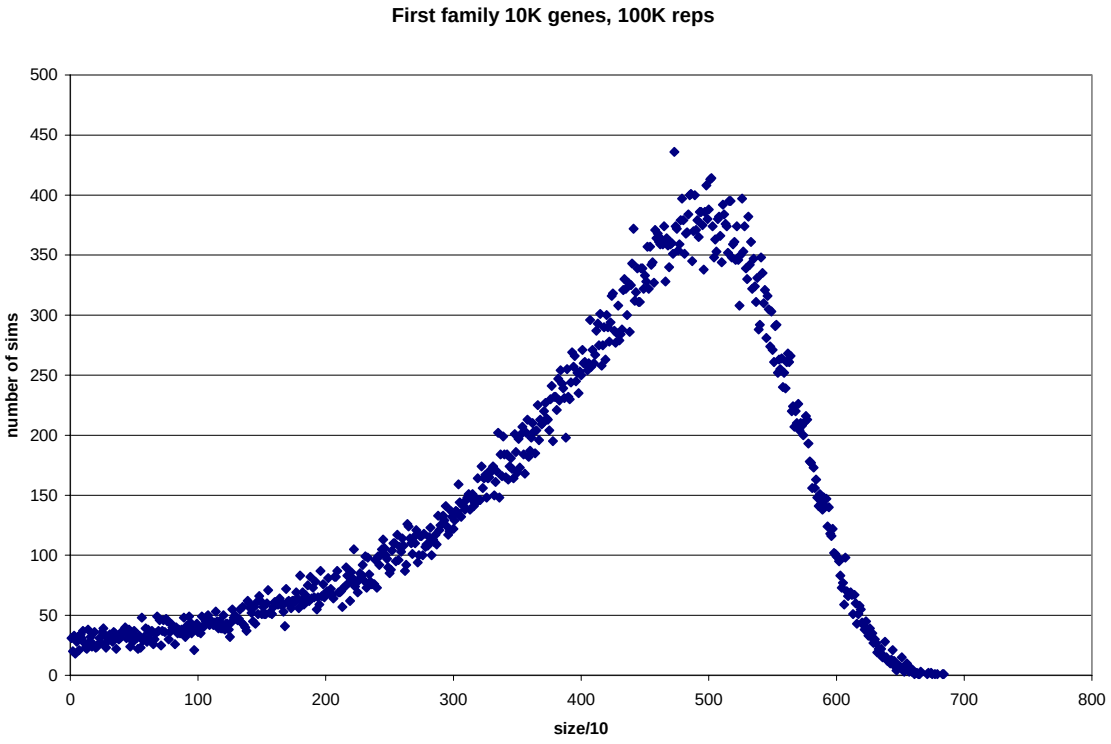

Figure 3. Distribution of the first family in 100,000 replications of the model with r = 0.1, and N = 10,000.