BRANCHING PROCESSES

Paul Balister
Department of Mathematical Sciences
University of Memphis
Memphis, TN 38152, USA
E-mail: pbalistr@memphis.edu

We introduce the basic theory of Galton-Watson branching processes, and the probabilistic tools needed to analyse them. The aim is to give a basic treatment of branching processes, including results on the limiting behaviour for subcritical, critical, and supercritical processes. We introduce just enough probabilistic theory to make the results rigourous, but avoid unnecessary technicalities as far as possible. We also include some results on multi-type processes, and an elegant connection with branching numbers of trees.

1. Notation and Preliminaries

In this article we shall be dealing with random processes, and so will be relying heavily on notation and results from probability theory. We shall start by summarising some probabilistic notation and tools in this section, while in Section 2 we shall define and begin the study of branching processes. Readers familiar with probability theory may therefore safely skip this section.

1.1. *Probability models*

In general, we describe random processes by use of a *probability model*. This consists of a (typically very large) set Ω of *outcomes*, one element $\omega \in \Omega$ of which we assume has been picked by "fate". Every detail of our random process is determined by the choice of ω. When we ask questions about our process we are asking for some information about ω. The simplest type of question is a Yes/No type question, which we shall call an *event*. Mathe-

matically we describe an event as a subset E of the set of all outcomes Ω. Informally E is the set of outcomes for which the answer to our question is "Yes". Thus if $\omega \in E$ then we say that the event E *occurred*, while if $\omega \notin E$ then we say that E did not occur. For example, if we toss a coin, we can obtain either heads (H) or tails (T). If we toss two coins, then there are four possible outcomes: $\Omega = \{\text{HH,HT,TH,TT}\}$. One possible event is $E = \{\text{we have at least one head}\} = \{\text{HH,HT,TH}\}$, another event is $E = \{\text{the coins showed the same face}\} = \{\text{HH,TT}\}$.

More general questions may result in a value, say a real number. These we can describe as a function $X\colon \Omega \to \mathbb{R}$, which for each outcome ω gives a value $X(\omega)$. For example, in the case of the two coin tosses above, $X(\omega)$ might be the number of heads. The function X is called a *random variable*. We usually denote random variables by capital letters X, Y, Z, and omit the dependence on ω when writing, for example, $X + Y = X(\omega) + Y(\omega)$. Given a random variable X, one can construct events such as $\{X < 5\} = \{\omega \in \Omega \mid X(\omega) < 5\}$ etc.. Once again, we usually omit the reference to ω. For example, in the above example of two coin tosses, and with X denoting the number of heads, we have $\{X < 2\} = \{\text{HT,TH,TT}\}$. Conversely, given an event E we can define a random variable, called the *indicator function* of E by

$$\mathbb{1}_E = \begin{cases} 1 & \text{if } E \text{ occurs;} \\ 0 & \text{if } E \text{ does not occur.} \end{cases}$$

Or, more formally, $\mathbb{1}_E(\omega) = 1$ if $\omega \in E$ and $\mathbb{1}_E(\omega) = 0$ if $\omega \notin E$.

Finally, we specify the *probability* of each event E, denoted by $\mathbb{P}(E)$, as a real number between 0 and 1. Heuristically, this represents the proportion of times the event should occur in many occurrences of the model. These probabilities satisfy the following laws:

P1.
$$\mathbb{P}(\emptyset) = 0, \qquad \text{and} \qquad \mathbb{P}(\Omega) = 1.$$

P2. If E_1 and E_2 are events with $E_1 \subseteq E_2$ (i.e., if E_1 occurs then so does E_2), then
$$\mathbb{P}(E_1) \leq \mathbb{P}(E_2).$$

P3. If E_1 and E_2 are *disjoint*, $E_1 \cap E_2 = \emptyset$ (so E_1 and E_2 can't both occur simultaneously), then
$$\mathbb{P}(\text{either } E_1 \text{ or } E_2 \text{ occurs}) = \mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2).$$

P4. If $E_1 \subseteq E_2 \subseteq E_3 \subseteq \dots$ is a increasing nested sequence of events then

$$\mathbb{P}(\text{at least one } E_n \text{ occurs}) = \mathbb{P}(\bigcup_{n=1}^{\infty} E_n) = \lim_{n\to\infty} \mathbb{P}(E_n).$$

P5. If $E_1 \supseteq E_2 \supseteq E_3 \supseteq \dots$ is a decreasing nested sequence of events then

$$\mathbb{P}(\text{all } E_n \text{ occur}) = \mathbb{P}(\bigcap_{n=1}^{\infty} E_n) = \lim_{n\to\infty} \mathbb{P}(E_n).$$

As a consequence of P3 and P4 one can show

P6. If $E_1, E_2, \dots$ are *pairwise disjoint* events, $E_i \cap E_j = \emptyset$ for all $i \neq j$, then

$$\mathbb{P}(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

Properties P4, P5 and P6 are sometimes referred to as *continuity of probability*. Note that in P4 and P5, $\mathbb{P}(E_n)$ is a bounded monotonic sequence of real numbers, so the limits on the right do indeed exist.

Technical Note: If Ω is finite (or countably infinite), then one can just define $\mathbb{P}(E)$ as the sum of the probabilities $\mathbb{P}(\{\omega\})$ of each outcome $\omega \in E$. For uncountably infinite Ω this is in general not possible. Moreover, it is generally not possible to satisfactorily define the probability for *every* subset E of Ω. Instead we restrict our events to only *some* subsets of Ω. The collection $\mathcal{F}$ of these valid events will form a σ-field: it will be closed under any finite number of set operations, as well as countable unions and intersections. In all our examples, essentially any property one can write down about ω will be in $\mathcal{F}$, so we usually ignore this technicality. To make everything precise in full generality needs knowledge of *measure theory*, which we will try to avoid as much as possible in these notes. See [6] for more details.

We say an event E occurs *almost surely* if $\mathbb{P}(E) = 1$. This is not quite the same as saying E *always* happens ($E = \Omega$), but as the probability of it not happening is zero, $\mathbb{P}(\Omega \setminus E) = 0$, it is the next best thing.

1.2. *Expectation*

For a random variable X that is *discrete*, i.e., takes only finitely or countably many real values $\{t_1, t_2, \dots\}$, we can define the *expectation* or average value of X by

$$\mathbb{E}(X) = \sum_i t_i \, \mathbb{P}(X = t_i), \tag{1.1}$$

in other words, $\mathbb{E}(X)$ is the average of the values taken weighted by the probability that they occur. If infinitely many values are possible then we must be careful about convergence of this sum. In the case when $X \geq 0$ (always, i.e., all $t_i \geq 0$), then the sum either converges to some $c \in \mathbb{R}$ or tends to infinity. In these cases we write $\mathbb{E}(X) = c$ or $\mathbb{E}(X) = \infty$ respectively. If X can be negative then we need the sum to be absolutely convergent ($\mathbb{E}(|X|) < \infty$) to make the sum independent of the order of summation. If $\mathbb{E}(|X|) = \infty$ then we say $\mathbb{E}(X)$ *diverges*.

Sometimes we allow $\pm\infty$ as a value for X, in which case we use the convention that $\infty \cdot 0 = 0$ in (1.1). In this case $\mathbb{E}(X)$ will diverge unless $\mathbb{P}(X = \pm\infty) = 0$. It is worth remembering though that even if X is almost surely finite, $\mathbb{E}(X)$ can still diverge.

For random variables that take a continuous range of values the definition of the expectation is a bit more technical. One can often represent the expected value as some type of integral, but perhaps the simplest definition is to define it as the limit of approximate expectations

$$\mathbb{E}(X) = \lim_{n\to\infty} \sum_{i=-\infty}^{\infty} \tfrac{i}{n}\, \mathbb{P}(\tfrac{i}{n} \leq X < \tfrac{i+1}{n}).$$

We have the following properties of expectation.

E1. For any event E, $\mathbb{P}(E) = \mathbb{E}(\mathbb{1}_E)$.
E2. $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
E3. $\mathbb{E}(\lambda X) = \lambda\, \mathbb{E}(X)$ for any (non-random) constant λ.
E4. If $X \leq Y$ then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.
E5. $\mathbb{E}(|X|) = 0$ if and only if $X = 0$ almost surely.

Using E2 and induction one can swap any *finite* sum with expectations:

$$\mathbb{E}(\textstyle\sum_{i=1}^{n} X_n) = \sum_{i=1}^{n} \mathbb{E}(X_n),$$

however, for infinite sums we have $\sum_{n=1}^{\infty} \mathbb{E}(X_n) \neq \mathbb{E}(\sum_{n=1}^{\infty} X_n)$ in general. Indeed, in general $\lim_{n\to\infty} \mathbb{E}(X_n) \neq \mathbb{E}(\lim_{n\to\infty} X_n)$. There are however some important cases when we can swap limits or infinite sums with expectations (see [6] for more details).

Monotone Convergence Theorem (MCT): If X_n, $n \in \mathbb{N}$, are random variables such that $0 \leq X_1 \leq X_2 \leq \ldots$, then

$$\mathbb{E}(\lim_{n\to\infty} X_n) = \lim_{n\to\infty} \mathbb{E}(X_n). \tag{1.2}$$

Tonelli's Theorem[a]**:** If $X_n \geq 0$ then

$$\mathbb{E}(\textstyle\sum_{n=1}^{\infty} X_n) = \sum_{n=1}^{\infty} \mathbb{E}(X_n). \tag{1.3}$$

Proof that MCT⇒Tonelli: Apply MCT to $Y_n = \sum_{r=1}^{n} X_r$. □

Note that the limits and sums in MCT and Tonelli may be non-negative real numbers or $+\infty$.

Dominated Convergence Theorem (DCT): If X_n, $n \in \mathbb{N}$, and Y are random variables such that $|X_n| \leq Y$, $\mathbb{E}(Y) < \infty$, and $\lim_{n\to\infty} X_n$ exists almost surely, then

$$\mathbb{E}(\lim_{n\to\infty} X_n) = \lim_{n\to\infty} \mathbb{E}(X_n). \tag{1.4}$$

Fubini's Theorem[a]**:** If $\sum_{n=1}^{\infty} \mathbb{E}(|X_n|) < \infty$ then

$$\mathbb{E}(\textstyle\sum_{n=1}^{\infty} X_n) = \sum_{n=1}^{\infty} \mathbb{E}(X_n). \tag{1.5}$$

Proof that DCT⇒Fubini: Apply DCT to $Y_n = \sum_{r=1}^{n} X_r$ and $Y = \sum_{n=1}^{\infty} |X_n|$, using Tonelli to show that $\mathbb{E}(Y) < \infty$, and noting that $\mathbb{P}(\lim_{n\to\infty} Y_n \text{ exists}) \geq \mathbb{P}(\sum_{n=1}^{\infty} |X_n| < \infty) = 1$ since $\mathbb{E}(Y) < \infty$. □

Note that the limits and sums in DCT and Fubini will be necessarily finite.

One can also easily generalise both the MCT and DCT to the case of a parameterised collection of random variables indexed by real numbers $\{X_t : t \in \mathbb{R},\ t \geq 0\}$.

The following observation is often very useful.

Markov's Inequality: If $X \geq 0$ and $c > 0$ then

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}. \tag{1.6}$$

Proof: If $X \geq 0$ then $c\mathbb{1}_{\{X\geq c\}} \leq X$, so $\mathbb{E}(c\mathbb{1}_{\{X\geq c\}}) \leq \mathbb{E}(X)$. But $\mathbb{E}(c\mathbb{1}_{\{X\geq c\}}) = c\,\mathbb{P}(X \geq c)$, so $\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$. □

1.3. *Conditional Probability*

Often we talk about *conditioning* on some event E (which we shall assume has non-zero probability). Intuitively, this means we are assuming that E

[a]Actually, just a very special case of Tonelli's/Fubini's Theorem.

occurs and rescaling our probability to this part of Ω. We define the *conditional probability* of A conditioned on E by

$$\mathbb{P}(A \mid E) = \frac{\mathbb{P}(A \cap E)}{\mathbb{P}(E)}. \tag{1.7}$$

It can be easily checked that $\mathbb{P}(A \mid E)$ gives a probability model on the set of outcomes $\Omega' = E$, where we ignore any outcomes that don't lie in E. If we have a partition of $\Omega = E_1 \cup E_2 \cup \dots$ into (finitely or countably infinitely many) disjoint events E_n then one can recover the probability of any event A by taking a weighted average of the conditional probabilities

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \mid E_n)\, \mathbb{P}(E_n).$$

(To see this, use P6 to write $\mathbb{P}(A) = \sum_n \mathbb{P}(A \cap E_n)$ and then use (1.7).) Similarly we can define *conditional expectation* as expectation in the new model $\mathbb{P}(\cdot \mid E)$:

$$\mathbb{E}(X \mid E) = \frac{\mathbb{E}(X \mathbb{1}_E)}{\mathbb{P}(E)}.$$

Once again, if $\Omega = E_1 \cup E_2 \cup \dots$ is a partition of Ω into disjoint events then

$$\mathbb{E}(X) = \sum_n \mathbb{E}(X \mid E_n)\, \mathbb{P}(E_n). \tag{1.8}$$

We can also define the idea of conditioning on a random variable Y, by defining $\mathbb{E}(X \mid Y)$ as a random variable that for each $y \in \mathbb{R}$ takes the (non-random) value $\mathbb{E}(X \mid Y = y)$ whenever $Y = y$. There are some technical issues here, since this is undefined whenever $\mathbb{P}(Y = y) = 0$. However, this does not cause problems as long as Y is discrete, i.e., when Y takes only finitely or countably many distinct values. Using this notation, we have the following terse (but somewhat cryptic) reformulation of (1.8):

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y)). \tag{1.9}$$

Even when Y is not discrete, it is possible to define a random variable $\mathbb{E}(X \mid Y)$ so that it is a function of Y and (1.9) holds. However, the proof of this belongs to a course on measure theory, so we omit the details here (see for example [6]).

1.4. *Independence*

We say two events E_1 and E_2 are *independent* if

$$\mathbb{P}(E_1 \text{ and } E_2 \text{ both occur}) = \mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2).$$

More intuitively, this is equivalent to $\mathbb{P}(E_1 \mid E_2) = \mathbb{P}(E_1)$ and $\mathbb{P}(E_2 \mid E_1) = \mathbb{P}(E_2)$, i.e., knowing one event has occurred does not make it any more or less likely that the other has occurred. We say a random variable X is independent of an event E if *every* event constructed from X (such as $\{X < 1\}$, $\{X \in (2, 5)\}$, etc.) is independent of E. Similarly, two random variables X and Y are independent if every event constructed from X is independent of every event constructed from Y. More generally, a collection of events and/or random variables is independent of another collection of events and/or random variables, if every event constructed using information from the first collection is independent of any event constructed using information from the second collection.

For *independent* random variables X and Y we have

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$$

provided the right hand side is well defined. Note that this does not hold in general for non-independent random variables. As a particular example $\mathbb{E}(X^2) \neq (\mathbb{E}(X))^2$ unless X is almost surely a constant. Indeed, if one defines the *variance* of X by

$$\mathrm{Var}(X) = \mathbb{E}((X - \mu)^2)$$

where $\mu = \mathbb{E}(X)$ then $\mathrm{Var}(X)$ is clearly non-negative, and (by property E5) it is strictly positive unless $X = \mu$ almost surely. However,

$$\begin{aligned}\mathbb{E}((X - \mu)^2) &= \mathbb{E}(X^2 - 2\mu X + \mu^2)\\ &= \mathbb{E}(X^2) - 2\mu\,\mathbb{E}(X) + \mu^2\\ &= \mathbb{E}(X^2) - \mu^2,\end{aligned}$$

so

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$$

is the difference between $\mathbb{E}(X^2)$ and $(\mathbb{E}(X))^2$.

In general for any (not necessarily independent) X and Y, we have at least the following useful bound on $\mathbb{E}(XY)$.

Cauchy-Schwarz Inequality:

$$(\mathbb{E}(XY))^2 \leq \mathbb{E}(X^2)\,\mathbb{E}(Y^2). \tag{1.10}$$

Proof: For any constants α, β, we have $(\alpha X - \beta Y)^2 \geq 0$. Thus

$$\begin{aligned}0 \leq \mathbb{E}((\alpha X - \beta Y)^2) &= \mathbb{E}(\alpha^2 X^2 + \beta^2 Y^2 - 2\alpha\beta XY)\\ &= \alpha^2\mathbb{E}(X^2) + \beta^2\mathbb{E}(Y^2) - 2\alpha\beta\,\mathbb{E}(XY).\end{aligned}$$

Hence

$$2\alpha\beta\,\mathbb{E}(XY) \le \alpha^2\mathbb{E}(X^2) + \beta^2\mathbb{E}(Y^2).$$

If $0 < \mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$, set $\alpha = (\mathbb{E}(X^2))^{-1/2}$ and $\beta = (\mathbb{E}(Y^2))^{-1/2}$. Then

$$\frac{2\,\mathbb{E}(XY)}{(\mathbb{E}(X^2))^{1/2}(\mathbb{E}(Y^2))^{1/2}} \le 1 + 1 = 2$$

or $E(XY) \le (\mathbb{E}(X^2))^{1/2}(\mathbb{E}(Y^2))^{1/2}$. Squaring both sides gives the result. If $\mathbb{E}(X^2) = 0$ then $X = 0$ almost surely, so $\mathbb{E}(XY) = 0$ and the inequality still holds. Similarly for $\mathbb{E}(Y^2) = 0$. If $\mathbb{E}(X^2) = \infty$ or $\mathbb{E}(Y^2) = \infty$ then the result is automatic. □

Finally, we show that for independent random variables, the variance is additive.

Lemma 1.1: *If X and Y are independent then*

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Proof: If X and Y are independent then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Thus

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}((X+Y)^2) - (\mathbb{E}(X+Y))^2 \\
&= \mathbb{E}(X^2+2XY+Y^2) - (\mathbb{E}(X)+\mathbb{E}(Y))^2 \\
&= \mathbb{E}(X^2+Y^2)+2\mathbb{E}(XY)-(\mathbb{E}(X))^2-2\mathbb{E}(X)\mathbb{E}(Y)-(\mathbb{E}(Y))^2 \\
&= \mathbb{E}(X^2) + \mathbb{E}(Y^2) - (\mathbb{E}(X))^2 - (\mathbb{E}(Y))^2 \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y).
\end{aligned}$$

□

1.5. *Probability Distributions*

Suppose X is a discrete random variable, so X takes only finitely or countably infinitely many values $\{t_1, t_2, \dots\}$. Then one can describe the probability of any event depending only on X just by specifying the probabilities $p_{t_i} = \mathbb{P}(X = t_i)$ for each t_i. We say that the *probability distribution* of X is given by the real numbers p_{t_i}, $i = 1, 2, \dots$. If X takes uncountably many values then this approach is not sufficient, since it may be the case that $\mathbb{P}(X = t) = 0$ for every $t \in \mathbb{R}$. One can however describe the probability of any event depending only on X in terms of the *cumulative probability distribution function*

$$F(c) = \mathbb{P}(X \le c).$$

The function $F(c)$ is not arbitrary, for example it is clear that $F(c)$ is an increasing function of c. Moreover, by continuity of probability

$$\begin{aligned}\lim_{x\to c^+} F(x) &= \lim_{n\to\infty} F(c+\tfrac{1}{n}) \\ &= \mathbb{P}(\textstyle\bigcap_n \{X \le c+\frac{1}{n}\}) \\ &= \mathbb{P}(X \le c),\end{aligned}$$

so

$$\lim_{x\to c^+} F(x) = F(c), \tag{1.11}$$

i.e., $F(c)$ is right-continuous. Now

$$\begin{aligned}\lim_{x\to c^-} F(x) &= \lim_{n\to\infty} F(c-\tfrac{1}{n}) \\ &= \mathbb{P}(\textstyle\bigcup_n \{X \le c-\frac{1}{n}\}) \\ &= \mathbb{P}(X < c).\end{aligned}$$

Thus

$$\mathbb{P}(X = c) = \mathbb{P}(X \le c) - \mathbb{P}(X < c) = F(c) - \lim_{x\to c^-} F(x),$$

so F is continuous at $x = c$ if and only if $\mathbb{P}(X = c) = 0$. Similarly, one has

$$\lim_{x\to-\infty} F(x) = 0 \qquad \text{and} \qquad \lim_{x\to+\infty} F(x) = 1. \tag{1.12}$$

Indeed, it can be shown that any function $F(x)$ satisfying (1.11) and (1.12) is the cumulative probability function of some random variable. Sometimes $F(x)$ is differentiable, in which case we call the derivative $f(x) = F'(x)$ the *probability density function* of X. If $f(x)$ exists then we can recover the probability of events depending of X by integration. For example

$$\mathbb{P}(a < X < b) = \int_a^b f(x)\,dx = F(b) - F(a).$$

It is worth noting however, that even if $\mathbb{P}(X = c) = 0$ for all c, it does not necessarily follow that the probability density function exists. It is possible that $F(c)$ is continuous, but not differentiable.

2. Galton-Watson Processes

Suppose at time 0 we have a single bacterium, and at each time step a bacterium randomly either divides or dies. How would we expect the population of bacteria to grow? Similarly, imagine an outbreak of an infectious

disease in a large population. Initially there is just one infected individual, and each infected individual infects a random number of other individuals. Would the disease spread, and if so, how quickly?

We can model the above questions with a *Galton-Watson branching process*: we specify some probability distribution $(p_k)_{k=0}^{\infty}$ on $\mathbb{N} = \{0, 1, 2, \dots\}$, so $p_k \geq 0$ and $\sum_{k=0}^{\infty} p_k = 1$. Now define random variables Z_n by setting $Z_0 = 1$ and for $n \geq 0$ letting Z_{n+1} be a sum of Z_n independent random variables ξ_i, $1 \leq i \leq Z_n$, where each ξ_i has the probability distribution given by $(p_k)_{k=0}^{\infty}$, i.e., $\mathbb{P}(\xi_i = k) = p_k$. The variable Z_n then represents the number of bacteria or infected individuals at time n.

Equivalently, we can represent the process as a random *Galton-Watson tree*. We start with a *node* (vertex) v_0 and then inductively declare each node to be joined to a number of *child nodes*. The number of such nodes is random with probability distribution $(p_k)_{k=0}^{\infty}$, and the number of children of a node is independent of the choice of the number of children of all the other nodes. The random variable Z_n is then just the number of nodes at *level* n, i.e., the number of nodes that are at graph distance n from v_0 (see Figure 1).

The process $(Z_n)_{n=0}^{\infty}$ we have constructed is a *Markov Chain*: the value of Z_{n+1} depends on the previous values $Z_0, \dots, Z_n$ only via the value of Z_n.

First we shall consider some qualitative arguments about this process using basic probability estimates and fairly simple but general methods.

There are two distinct possibilities: either $Z_n > 0$ for all n, in which case we say the process *survives*; or for some n, $Z_n = 0$, in which case $Z_m = 0$ for all $m \geq n$. In this case we say the process *dies out* or becomes *extinct*. The probabilities of these events obviously depend on the choice of the distribution $(p_k)_{k=0}^{\infty}$.

If we wish to estimate Z_n, our first attempt should be to look at the average, or mean $\mathbb{E}(Z_n)$, of Z_n.

Lemma 2.1: *$\mathbb{E}(Z_n) = \mu^n$ where $\mu = \sum_{k=0}^{\infty} k p_k$ is the mean of the distribution $(p_k)_{k=0}^{\infty}$.*

Proof: If we condition on the event $Z_{n-1} = k$ then Z_n is the sum $\xi_1 + \cdots + \xi_k$ of k random variables, each with mean μ. Hence,

$$\mathbb{E}(Z_n \mid Z_{n-1} = k) = \mathbb{E}(\xi_1 + \cdots + \xi_k) = \mathbb{E}(\xi_1) + \cdots + \mathbb{E}(\xi_k) = k\mu.$$

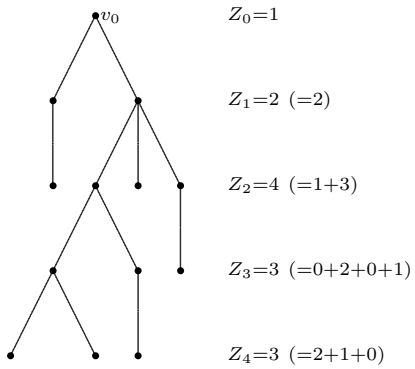


Fig. 1. A Galton-Watson tree.

Thus

$$\begin{aligned}\mathbb{E}(Z_n) &= \sum_{k=0}^{\infty} \mathbb{E}(Z_n \mid Z_{n-1} = k)\, \mathbb{P}(Z_{n-1} = k) \\ &= \sum_{k=0}^{\infty} \mu k\, \mathbb{P}(Z_{n-1} = k) \\ &= \mu\, \mathbb{E}(Z_{n-1}).\end{aligned}$$

The result follows by induction on n since $\mathbb{E}(Z_0) = 1$. □

Corollary 2.2: *If $\mu < 1$ then the process $(Z_n)_{n=0}^{\infty}$ almost surely dies out.*

Proof: If $(Z_n)_{n=0}^{\infty}$ survives then $Z_n > 0$ for all n, and for any value of Z_n, $\mathbb{1}_{\{Z_n>0\}} \le Z_n$. Hence, for all n,

$$\mathbb{P}(\text{survives}) \le \mathbb{P}(Z_n > 0) = \mathbb{E}(\mathbb{1}_{\{Z_n>0\}}) \le \mathbb{E}(Z_n) = \mu^n.$$

But $\mu^n \to 0$ as $n \to \infty$, so $\mathbb{P}(\text{survives}) = 0$. □

Let $T = \inf\{n : Z_n = 0\} \in [0, \infty]$ be the time to extinction. Corollary 2.2 says that if $\mu < 1$ then $T < \infty$ almost surely. In fact we can strengthen this to the following.

Corollary 2.3: *If $\mu < 1$ then $\mathbb{E}(T) < \infty$.*

Proof: Note that $Z_0, Z_1, \ldots, Z_{T-1} > 0$ and $Z_T, Z_{T+1}, \cdots = 0$, so

$$T = \mathbb{1}_{\{Z_0>0\}} + \mathbb{1}_{\{Z_1>0\}} + \cdots \le Z_0 + Z_1 + Z_2 + \cdots .$$

Since all the Z_i are non-negative, $\mathbb{E}(\sum_i Z_i) = \sum_i \mathbb{E}(Z_i)$ (by (1.3)). Thus

$$\begin{aligned}\mathbb{E}(T) &\leq \mathbb{E}(Z_0 + Z_1 + \dots) \\ &= \mathbb{E}(Z_0) + \mathbb{E}(Z_1) + \dots \\ &= 1 + \mu + \mu^2 + \dots \\ &= \tfrac{1}{1-\mu} < \infty.\end{aligned}$$

□

Now consider the case when $\mu \geq 1$. Then $\mathbb{E}(Z_n) \not\to 0$. However this does not imply that $Z_n \not\to 0$. Indeed, in the case when $\mu = 1$ we shall show that the process still dies out almost surely.

Lemma 2.4: *If $p_1 < 1$ then almost surely either $(Z_n)_{n=0}^{\infty}$ dies out or $Z_n \to \infty$ as $n \to \infty$.*

Note: if $p_1 = 1$ then each node in the Galton-Watson tree has exactly one child, so $Z_n = 1$ for all n. We shall usually exclude this trivial case.

Proof: Fix $k > 0$ and condition on the event $Z_n = k$. If $p_0 > 0$ then with probability p_0^k we will have $Z_{n+1} = 0$, and so $Z_m = 0 \neq k$ for all $m > n$. If $p_0 = 0$ then Z_n is non-decreasing in n, since every vertex in the Galton-Watson tree necessarily has at least one child. But with probability $1 - p_1$, the first vertex at level n will have more than one child, so $Z_{n+1} > Z_n$. But in this case $Z_m \neq k$ for all $m > n$. Thus in general there exists a $\gamma > 0$ such that

$$\mathbb{P}(\forall m > n \colon Z_m \neq k \mid Z_n = k) \geq \gamma. \tag{2.1}$$

Fix $k > 0$. Let E_i be the event that $Z_n = k$ occurs at least i times and E_∞ the event that $Z_n = k$ occurs infinitely often. Intuitively, (2.1) suggests that for any $i > 0$, if we have seen a value i times, then with probability at most $1 - \gamma$ will we ever see it again. Equivalently, $\mathbb{P}(E_{i+1} \mid E_i) \leq 1 - \gamma$. Let us check this more rigourously.

The event E_i is a disjoint union of all the events of the form

$$E_{i,(a_1,\dots,a_{n-1})} = \{Z_n = k, Z_{n-1} = a_{n-1}, \dots, Z_1 = a_1\},$$

where exactly $i-1$ of the numbers $a_1, \dots, a_{n-1}$ are equal to k and n ranges over all non-negative integers. Conditioned on any one of these events, E_{i+1} occurs with probability at most $1 - \gamma$. Thus

$$\mathbb{P}(E_{i+1} \cap E_{i,(a_1,\dots,a_{n-1})}) \leq (1-\gamma)\mathbb{P}(E_{i,(a_1,\dots,a_{n-1})}).$$

Adding these inequalities over all the (disjoint) sub-events $E_{i,(a_1,\ldots,a_{n-1})}$ gives

$$\mathbb{P}(E_{i+1} \cap \bigcup_{n,(a_1,\ldots,a_{n-1})} E_{i,(a_1,\ldots,a_{n-1})}) \leq (1-\gamma)\mathbb{P}(\bigcup_{n,(a_1,\ldots,a_{n-1})} E_{i,(a_1,\ldots,a_{n-1})}).$$

or more simply

$$\mathbb{P}(E_{i+1}) = \mathbb{P}(E_{i+1} \cap E_i) \leq (1-\gamma)\mathbb{P}(E_i). \tag{2.2}$$

Hence indeed it is true that $\mathbb{P}(E_{i+1} \mid E_i) \leq 1-\gamma$. But $\mathbb{P}(E_1) \leq 1$, so by (2.2) and induction on i, $\mathbb{P}(E_i) \leq (1-\gamma)^{i-1}$. Now $\mathbb{P}(E_\infty) \leq \mathbb{P}(E_i) \leq (1-\gamma)^{i-1}$ for all $i > 0$ and so $\mathbb{P}(E_\infty) = 0$. Now consider $\liminf_{n\to\infty} Z_n$. If $\liminf_{n\to\infty} Z_n = k$ then $Z_n = k$ occurs infinitely often, which we know cannot happen if $k > 0$. Thus with probability 1, $\liminf_{n\to\infty} Z_n \in \{0, \infty\}$. If $\liminf_{n\to\infty} Z_n = 0$ then $Z_n = 0$ for some n, and so the process dies out. If $\liminf_{n\to\infty} Z_n = \infty$ then $Z_n \to \infty$. The result follows. □

Corollary 2.5: *If $p_1 < 1$ and $k > 0$ then $\mathbb{P}(Z_n = k) \to 0$ as $n \to \infty$.*

Proof: $Z_n = k$ for infinitely many n if and only if, for all n, there is an $m \geq n$ such that $Z_m = k$. In terms of events this gives

$$\{Z_n = k \text{ infinitely often}\} = \bigcap_{n=0}^{\infty} \{Z_m = k \text{ for some } m \geq n\}.$$

Since $\mathbb{P}(Z_n = k \text{ infinitely often}) = 0$, continuity of probability gives

$$\mathbb{P}(Z_m = k \text{ for some } m \geq n) \to 0 \quad \text{as} \quad n \to \infty.$$

But $\{Z_n = k\} \subseteq \{Z_m = k \text{ for some } m \geq n\}$ so the result follows. □

Corollary 2.6: *If $\mu = 1$ and $p_1 < 1$ then $(Z_n)_{n=0}^{\infty}$ dies out almost surely.*

Proof: Fix $k > 0$. Then by Markov's inequality, $\mathbb{P}(Z_n > k) \leq \frac{\mathbb{E}(Z_n)}{k} = \frac{1}{k}$. But for each $i = 1, \ldots, k$, $\mathbb{P}(Z_n = i) \to 0$ as $n \to \infty$, so for sufficiently large n (depending on k), $\mathbb{P}(Z_n = i) \leq \frac{1}{k^2}$ for each $i = 1, 2, \ldots, k$. Then

$$\mathbb{P}(Z_n > 0) = \textstyle\sum_{i=1}^{k} \mathbb{P}(Z_n = i) + \mathbb{P}(Z_n > k) \leq \frac{k}{k^2} + \frac{1}{k} = \frac{2}{k}.$$

Since $\mathbb{P}(\text{survives}) \leq \mathbb{P}(Z_n > 0)$ for all n, $\mathbb{P}(\text{survives}) \leq \frac{2}{k}$. Since this holds for all k, $\mathbb{P}(\text{survives}) = 0$. □

We shall see later that the expected time to extinction, $\mathbb{E}(T)$, is usually infinite when $\mu = 1$.

In the case when $\mu = 1$ and $p_1 < 1$ we have seen that $Z_n \to 0$ almost surely, while $\mathbb{E}(Z_n) = 1$ for all n. In particular $\mathbb{E}(\lim_{n\to\infty} Z_n) \neq \lim_{n\to\infty} \mathbb{E}(Z_n)$. Hence we have an example where swapping limits and expectations is not valid. The question is, why did this happen? If we look at the distribution of Z_n then it is very lopsided (see Figure 2). The probability that $Z_n > 0$ is very small, but when $Z_n > 0$, Z_n is typically very large. Thus $\mathbb{E}(Z_n) = 1$, even though $\mathbb{P}(Z_n = 0)$ is close to 1.

Fig. 2. $\mathbb{P}(Z_n = 0)$ large although $\mathbb{E}(Z_n) = 1 \not\to 0$.

Now consider the situation when $\mu > 1$. Of course, if $p_0 > 0$ then the process could die out in the first step. Thus all we can hope for is that the probability of survival is strictly positive. To show that this happens however, we need to show that the distribution of Z_n is not too spread out (in contrast to the $\mu = 1$ case). Write σ^2 for the variance of the distribution $(p_k)_{k=0}^{\infty}$, so that if ξ is distributed according to $(p_k)_{k=0}^{\infty}$ then $\mathrm{Var}(\xi) = \mathbb{E}(\xi^2) - (\mathbb{E}(\xi))^2 = \sigma^2$.

Lemma 2.7: *If $\sigma^2 < \infty$ then*

$$\mathrm{Var}(Z_n) = \sigma^2\mu^{n-1}(1 + \mu + \mu^2 + \cdots + \mu^{n-1}).$$

Hence if $\mu \neq 1$ then

$$\mathrm{Var}(Z_n) = \sigma^2\mu^{n-1}\left(\tfrac{\mu^n - 1}{\mu - 1}\right),$$

and if $\mu = 1$ then

$$\mathrm{Var}(Z_n) = n\sigma^2.$$

Proof: Conditioning on $Z_{n-1} = k$ we have $Z_n = \xi_1 + \cdots + \xi_k$ where ξ_i are independent with distribution $(p_k)_{k=0}^{\infty}$. Thus $\mathbb{E}(\xi_i\xi_j) = \mathbb{E}(\xi_i)\mathbb{E}(\xi_j) = \mu^2$ if

$i \neq j$ and $\mathbb{E}(\xi_i\xi_j) = \mathbb{E}(\xi_i^2) = \mathrm{Var}(\xi_i) + (\mathbb{E}(\xi_i))^2 = \sigma^2 + \mu^2$ if $i = j$. Then

$$\begin{aligned}\mathbb{E}(Z_n^2 \mid Z_{n-1} = k) &= \mathbb{E}(\textstyle\sum_{i,j=1}^n \xi_i\xi_j) \\ &= \textstyle\sum_{i,j=1}^n \mathbb{E}(\xi_i\xi_j) \\ &= \textstyle\sum_{i=j=1}^n \sigma^2 + \sum_{i,j=1}^n \mu^2 \\ &= k\sigma^2 + k^2\mu^2.\end{aligned}$$

Thus

$$\mathbb{E}(Z_n^2) = \sum_{k=0}^{\infty}(k\sigma^2 + k^2\mu^2)\,\mathbb{P}(Z_{n-1} = k) = \sigma^2\,\mathbb{E}(Z_{n-1}) + \mu^2\,\mathbb{E}(Z_{n-1}^2)$$

Since $\mathbb{E}(Z_n) = \mu^n$ we get

$$\begin{aligned}\mu^{-2n}\mathrm{Var}(Z_n) &= \mu^{-2n}\mathbb{E}(Z_n^2) - \mu^{-2n}(\mathbb{E}(Z_n))^2 \\ &= \sigma^2\mu^{-2n}\mathbb{E}(Z_{n-1}) + \mu^{-2n}\mu^2\mathbb{E}(Z_{n-1}^2) - 1 \\ &= \sigma^2\mu^{-2n}\mu^{n-1} + \mu^{-2(n-1)}\mathbb{E}(Z_{n-1}^2) - \mu^{-2(n-1)}(\mathbb{E}(Z_{n-1}))^2 \\ &= \sigma^2\mu^{-n-1} + \mu^{-2(n-1)}\mathrm{Var}(Z_{n-1}).\end{aligned}$$

Using $\mathrm{Var}(Z_0) = 0$, we obtain by induction

$$\mu^{-2n}\mathrm{Var}(Z_n) = \sigma^2(\mu^{-n-1} + \cdots + \mu^{-3} + \mu^{-2}).$$

Thus $\mathrm{Var}(Z_n) = \sigma^2\mu^{n-1}(1 + \mu + \cdots + \mu^{n-1})$. The last part of the lemma follows by summing the geometric series $1 + \mu + \cdots + \mu^{n-1}$. □

Lemma 2.8: *If* $\mu > 1$ *then* $\mathbb{P}(Z_n \to \infty) > 0$.

Proof: First, assume that $\mu, \sigma^2 < \infty$. Now by Lemma 2.7

$$\mathbb{E}(Z_n^2) = \mathrm{Var}(Z_n) + (\mathbb{E}(Z_n))^2 = \sigma^2\tfrac{\mu^n(\mu^n-1)}{\mu(\mu-1)} + \mu^{2n} \leq \Big(\tfrac{\sigma^2}{\mu(\mu-1)} + 1\Big)\mu^{2n}$$

In particular, $\mathbb{E}(Z_n^2) \leq C(\mathbb{E}(Z_n))^2$ for some constant $C = \frac{\sigma^2}{\mu(\mu-1)} + 1$ that is independent of n. Applying the Cauchy-Schwarz inequality (1.10) to $X = Z_n$ and $Y = \mathbb{1}_{\{Z_n>0\}}$ gives

$$(\mathbb{E}(Z_n))^2 = (\mathbb{E}(Z_n\mathbb{1}_{\{Z_n>0\}}))^2 \leq \mathbb{E}(Z_n^2)\mathbb{E}(\mathbb{1}_{\{Z_n>0\}}^2) = \mathbb{E}(Z_n^2)\mathbb{P}(Z_n > 0).$$

Hence $\mathbb{P}(Z_n > 0) \geq \frac{1}{C}$. The events $\{Z_n > 0\}$ are decreasing and $\bigcap_n\{Z_n > 0\} = \{\text{survives}\}$, so by continuity of probability $\mathbb{P}(\text{survives}) \geq \frac{1}{C}$. Finally, by Lemma 2.4, $\mathbb{P}(Z_n \to \infty) = \mathbb{P}(\text{survives}) > 0$.

In the case that $\sigma^2 = \infty$ or $\mu = \infty$, choose a finite μ' with $1 < \mu' < \mu$. Then, for sufficiently large N, $\sum_{k=0}^N kp_k > \mu'$. Replace the random variables

ξ_i in the definition of Z_n by ξ'_i, where $\xi'_i = \xi_i$ if $\xi_i \le N$ and $\xi'_i = 0$ otherwise. Equivalently, delete all the descendants of a vertex of the Galton-Watson tree when that vertex has more than N children. Now ξ'_i has mean at least $\mu' > 1$ and finite variance, so this new process survives with positive probability. But this means that the original process must also survive with positive probability. □

Assume that $1 < \mu < \infty$ and $\sigma^2 < \infty$. For any c, $0 < c < 1$, $\mathbb{E}(Z_n \mathbb{1}_{\{Z_n < c\mu^n\}}) \le c\mu^n$ and $\mathbb{E}(Z_n) = \mathbb{E}(Z_n \mathbb{1}_{\{Z_n < c\mu^n\}}) + \mathbb{E}(Z_n \mathbb{1}_{\{Z_n \ge c\mu^n\}})$. Thus $\mathbb{E}(Z_n \mathbb{1}_{\{Z_n \ge c\mu^n\}}) \ge \mu^n(1-c)$. If we replace $\mathbb{1}_{\{Z_n>0\}}$ by $\mathbb{1}_{\{Z_n \ge c\mu^n\}}$ in the above proof we obtain that $\mathbb{P}(Z_n \ge c\mu^n) \ge \frac{(1-c)^2}{C}$, so the size of Z_n is exponential in n with positive probability. Indeed, we shall see later that $\mathbb{P}(\forall n\colon Z_n \ge c\mu^n) > 0$. Note that this is stronger, since it says that often we have $Z_n \ge c\mu^n$ for *all* n, while $\mathbb{P}(Z_n \ge c\mu^n) \ge \frac{(1-c)^2}{C}$ does not exclude the possibility that in each Galton-Watson tree, Z_n may just oscillate between being above and below $c\mu^n$, spending on average a reasonable amount of time above $c\mu^n$. In the $\sigma^2 = \infty$ case it may not necessarily be true that $\mathbb{P}(Z_n \ge c\mu^n)$ is bounded away from zero, but for any $\mu' < \mu$, $\mathbb{P}(Z_n \ge c\mu'^n)$ is bounded away from zero by the same truncation argument used in the proof of Lemma 2.8.

3. Generating Functions

For more quantitative results, the main tool we use for studying Galton-Watson processes is the *generating function* of a distribution. Given any random variable ξ with values in $\mathbb{N} = \{0, 1, 2, \dots\}$, the generating function of ξ is given by

$$f_\xi(x) = \mathbb{E}(x^\xi) = \sum_{k=0}^{\infty} \mathbb{P}(\xi = k) x^k.$$

If ξ has the distribution $(p_k)_{k=0}^\infty$ of child nodes in the Galton-Watson process, then we shall write $f(x)$ for $f_\xi(x)$.

Since $f(1) = \sum_{k=0}^\infty p_k = 1$ is absolutely convergent, the series for $f(x)$ converges for all complex x with $|x| \le 1$. Thus for $|x| < 1$, $f(x)$ is an analytic function and we can, for example, differentiate term by term to obtain $f'(x)$, $f''(x)$, etc.. We shall normally consider the function just on the real interval $[0, 1]$ where it is well-behaved for all $x < 1$. However, if ξ is unbounded then $f(x)$ may fail to be analytic, or even real-differentiable, at $x = 1$.

Table 1: Generating functions for some simple distributions.

Distribution	Notation	p_k	$f(x)$
"Split or die"	$2B(p)$	$p_2 = p,\ p_0 = 1-p$	$(1-p)+px^2$
Binomial	$Bin(n,p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	$((1-p)+px)^n$
Poisson	$Po(\mu)$	$e^{-\mu}\,\frac{\mu^k}{k!}$	$e^{(x-1)\mu}$
Geometric	$Geom(r)$	$(1-r)r^k$	$\frac{1-r}{1-rx}$

We list some simple properties of $f(x)$.

Lemma 3.1: *If $f(x)$ is the generating function of some distribution then*

G1. *$f(x)$ and all its derivatives exist and are non-negative on $[0,1)$;*
G2. *$f(x)$ is continuous, increasing, and convex on $[0,1]$;*
G3. *$f(0) = p_0$ and $f(1) = 1$;*
G4. *the mean μ is equal to $f'(1)$, or $+\infty$ if $f'(1)$ does not exist;*
G5. *the variance σ^2 is equal to $f''(1) + \mu(1-\mu)$, or $+\infty$ if $f''(1)$ does not exist.*

Technical Note: For G4 and G5, if $f(x)$ does not converge for $x > 1$ then we define the derivatives $f'(1)$ and $f''(1)$ as the limit of $f'(x)$ or $f''(x)$ as $x \to 1^-$. By G1 these functions are increasing, so the limit either exists or is $+\infty$. Alternatively we could define "left-derivatives" such as $f'(1^-) = \lim_{h\to 0^-} \frac{f(1+h)-f(1)}{h}$. By the Mean Value Theorem, this gives the same values for $f'(1)$ and $f''(1)$ as taking limits $x \to 1^-$.

Proof:
G1. Follows from the fact that $f(x)$ is analytic for $|x| < 1$, so all derivatives are defined by power series $f^{(r)}(x) = \sum_{n=r}^{\infty} n(n-1)\dots(n-r+1)p_n x^{n-r}$ with all terms non-negative when $x \in [0,1)$.
G2. By G1, $f(x)$ is continuous on $[0,1)$ and by the Monotone Convergence Theorem $f(x) = \mathbb{E}(x^\xi) \to \mathbb{E}(1^\xi) = f(1)$ as $x \to 1^-$, so $f(x)$ is continuous at $x = 1$. Monotonicity and convexity follow from the fact that $f'(x), f''(x) \geq 0$ on $(0,1)$.
G3. Clear.
G4. $f'(x) = \sum_k p_k k x^{k-1} = \mathbb{E}(\xi x^{\xi-1}) \to \mathbb{E}(\xi)$ as $x \to 1^-$ by MCT.
G5. $f''(x) = \sum_k p_k k(k-1)x^{k-2} = \mathbb{E}(\xi(\xi-1)x^{\xi-2}) \to \mathbb{E}(\xi(\xi-1))$ as $x \to 1^-$ by MCT, so $\mathrm{Var}(\xi) = \mathbb{E}(\xi^2) - (\mathbb{E}(\xi))^2 = \mathbb{E}(\xi(\xi-1)) + \mathbb{E}(\xi) - (\mathbb{E}(\xi))^2 =$

$f''(1) + \mu - \mu^2$. □

Now we apply the idea of generating functions to the Galton-Watson process $(Z_n)_{n=0}^{\infty}$. Let $f_n(x)$ be the generating function for Z_n, i.e.,

$$f_n(z) = f_{Z_n}(z) = \mathbb{E}(x^{Z_n}) = \sum_{k=0}^{\infty} \mathbb{P}(Z_n = k)x^k.$$

Lemma 3.2: *$f_0(x) = x$ and $f_{n+1}(x) = f(f_n(x))$.*

Proof: Since $Z_0 = 1$ we have $f_0(x) = x$. If we have k independent random variables $Y_1, \ldots, Y_k$ each distributed according to Z_n then

$$\begin{aligned}
\mathbb{E}(x^{Y_1+\cdots+Y_k}) &= \mathbb{E}(x^{Y_1}x^{Y_2}\ldots x^{Y_k}) \\
&= \mathbb{E}(x^{Y_1})\mathbb{E}(x^{Y_2})\ldots\mathbb{E}(x^{Y_k}) && \text{(Independence)} \\
&= f_n(x)f_n(x)\ldots f_n(x) && \text{(Distributed according to } Z_n\text{)} \\
&= (f_n(x))^k.
\end{aligned}$$

We can consider Z_{n+1} as the sum of Z_1 independent copies of Z_n since each of the Z_1 children of v_0 starts its own independent Galton-Watson process which we need to run for another n steps. Thus

$$\begin{aligned}
f_{n+1}(x) = \mathbb{E}(x^{Z_{n+1}}) &= \sum_{k=0}^{\infty} \mathbb{E}(x^{Z_{n+1}} \mid Z_1 = k)\mathbb{P}(Z_1 = k) \\
&= \sum_{k=0}^{\infty} (f_n(x))^k p_k \\
&= f(f_n(x)).
\end{aligned}$$

□

Note that Lemma 3.2 provides quick proofs of the mean and variance formulae for Z_n that we obtained in Section 2. For example, $f'_{n+1}(x) = \frac{d}{dx}f(f_n(x)) = f'(f_n(x))f'_n(x)$, so

$$\mathbb{E}(Z_{n+1}) = f'_{n+1}(1) = f'(f_n(1))f'_n(1) = f'(1)f'_n(1) = \mu\,\mathbb{E}(Z_n),$$

giving $\mathbb{E}(Z_n) = \mu^n$ by induction. Similarly one can calculate the variance (Lemma 2.7) using $f''_n(x)$.

The following theorem allows us to calculate the *exact* probability that $(Z_n)_{n=0}^{\infty}$ survives.

Theorem 3.1: *The probability $p_e = \mathbb{P}((Z_n)_{n=0}^{\infty}$ dies out$)$ is the smallest solution $p_e \in [0, 1]$ of the equation $f(p_e) = p_e$.*

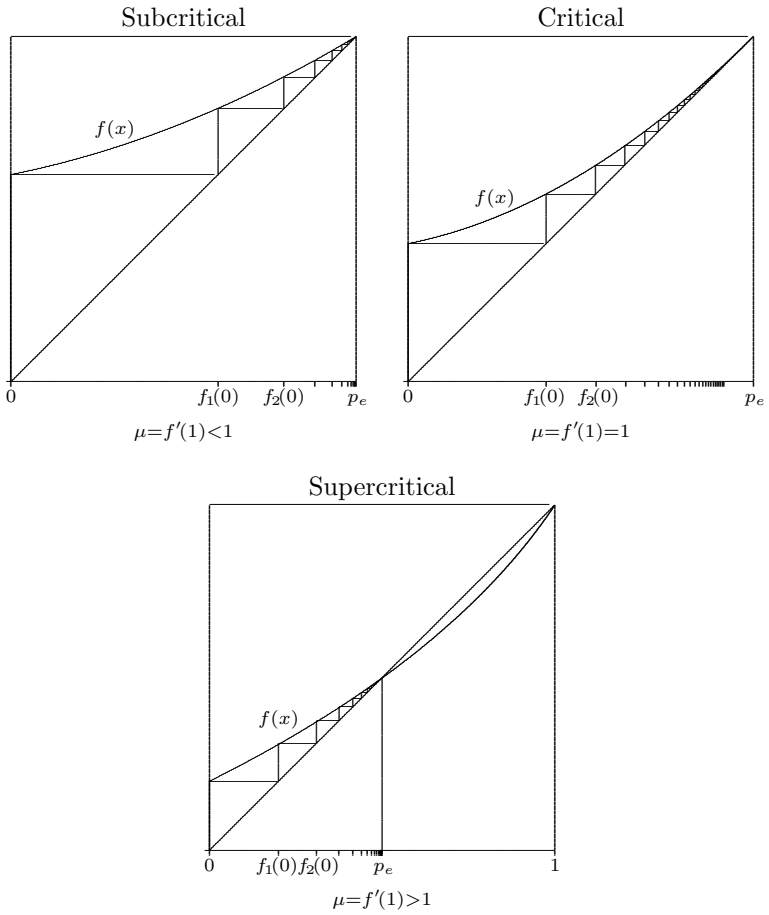


Fig. 3. Examples of subcritical, critical and supercritical generating functions.

Proof: Let E_n be the event that $Z_n = 0$. Then $E_1 \subseteq E_2 \subseteq \ldots$ and the event that the process dies out is $\bigcup_{n=0}^{\infty} E_n$. Since $\mathbb{P}(E_n) = f_n(0)$, continuity of probability implies that the probability that $(Z_n)_{n=0}^{\infty}$ dies out is $p_e = \lim_{n\to\infty} f_n(0)$ (see Figure 3). Let p_r be the smallest solution to $f(p_r) = p_r$. This exists by the Intermediate Value Theorem since $f(x)$ is continuous, $f(0) \geq 0$ and $f(1) \leq 1$. We now show that $f_n(0) \leq p_r$ by induction on n. Clearly $f_0(0) = 0 \leq p_r$, and $f(x)$ is an increasing function, so if $f_n(0) \leq p_r$ then $f_{n+1}(0) = f(f_n(0)) \leq f(p_r) = p_r$. Taking limits gives $p_e \leq p_r$. By continuity of $f(x)$, $f(p_e) = f(\lim_n f_n(0)) = \lim_n f(f_n(0)) = \lim_n f_{n+1}(0) =$

p_e. Thus p_e is a solution of $f(p) = p$. Since p_r was the smallest such solution and $p_e \leq p_r$, we must have $p_e = p_r$. □

Theorem 3.2: $\mathbb{P}((Z_n)_{n=0}^{\infty}$ *survives*$) > 0$ *if and only if* $\mu > 1$ (*or* $p_1 = 1$).

Proof: If $\mu = f'(1) > 1$ then for some $\varepsilon > 0$, $f(1-\varepsilon) < 1-\varepsilon$ (even if $\mu = \infty$). But $f(0) \geq 0$ so, by the Intermediate Value Theorem, there is some $x \in [0, 1-\varepsilon)$ with $f(x) = x$. Thus by Theorem 3.1, $p_e = \mathbb{P}(\text{dies out}) < 1-\varepsilon$, and so $\mathbb{P}(\text{survives}) = 1 - p_e > 0$. If $p_1 = 1$ then $Z_n = 1$ for all n, so $(Z_n)_{n=0}^{\infty}$ survives. Conversely, suppose $\mu \leq 1$ and $p_e < 1$. Then $f(p_e) = p_e$ and $f(1) = 1$. But $f'(x) \leq f'(1) \leq 1$ for $x \in (p_e, 1)$. If $f(x) < x$ for any $x \in (p_e, 1)$ then by the Mean Value Theorem, $f'(y) > 1$ for some $y \in (x, 1)$, while if $f(x) > x$ for any $x \in (p_e, 1)$ then $f'(y) > 1$ for some $y \in (p_e, x)$. Hence we must have $f(x) = x$ for all $x \in (p_e, 1)$. But then $f'(x) = 1$ and $f''(x) = \sum_{k=2}^{\infty} k(k-1)p_k x^{k-2} = 0$ for all $x \in (p_e, 1)$. Thus $p_k = 0$ for all $k \geq 2$ and $p_1 = \mu = f'(1) = 1$. □

4. Decomposing the supercritical process

Suppose we have a supercritical Galton-Watson process with $p_0 > 0$, so that the extinction probability p_e lies strictly between 0 and 1. Colour the vertices of the Galton-Watson tree red if they have an infinite number of descendants, and blue if they only have a finite number of descendants. If nothing is known about the descendants of a vertex v then it will be blue with probability p_e and red with probability $1 - p_e$. Hence if the colour and descendants of a vertex v are unknown, then the probability that there are r red and b blue children is given by

$$p(r, b) = p_{r+b}\tbinom{r+b}{r} p_e^b (1 - p_e)^r.$$

(There are $r + b$ children with probability p_{r+b}, and conditioning on this, the probability that a fixed subset of size r of these are red is $p_e^b(1-p_e)^r$ by independence. But there are $\binom{r+b}{r}$ choices for which subset should be red.)

Suppose we condition on the event that $Z_n \to 0$, i.e., on the event that v_0 is blue. Then the distribution of the number of children of a vertex v can be found simply by conditioning on the event that all these children are blue. (The event that v_0 is blue is the intersection of the event that the children of v are blue, and another event depending on other nodes of the tree that are independent of the subtree starting at v.) Thus the conditional probability of k children is $\frac{p(k,0)}{\sum_b p(b,0)} = \frac{p_k p_e^k}{p_e} = p_k p_e^{k-1}$, where we have used

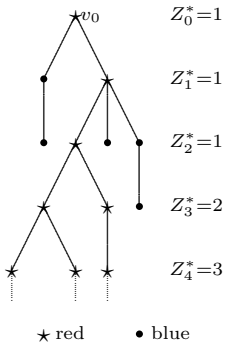


Fig. 4. Red and Blue trees.

the fact that $\sum_b p(b,0) = \mathbb{P}(v \text{ is blue}) = p_e$. The generating function of this distribution is $f_b(x) = \sum p_k p_e^{k-1} x^k = f(p_e x)/p_e$. Note that the number of children is still independent of the rest of the tree and only depends on the existence of a (blue) vertex v. Thus, conditioning on $Z_n \to 0$ we obtain a new Galton-Watson process with generating function $f_b(x)$. Graphically we see that $f_b(x)$ is obtained by taking the graph of $f(x)$ in $[0, p_e] \times [0, p_e]$ and scaling it up to a $[0,1] \times [0,1]$ square (see Figure 5). Also note that the new process is a subcritical Galton-Watson process since $f_b'(1) = f'(p_e) < 1$.

Now condition on v_0 being red, i.e., on $(Z_n)_{n=0}^{\infty}$ surviving. Suppose we now ignore all blue vertices and just look at the red subtree. Let Z_n^* be the number of red vertices at level n. Fix some red vertex v and consider the number of red children it has. Conditioning on v being red is equivalent to conditioning on it having at least one red child. The conditional probability of having $r > 0$ red children is $\frac{\sum_b p(r,b)}{1-p_e}$. Once again this is independent of the rest of the tree, so we obtain a new Galton-Watson process with

generating function

$$\begin{aligned} f_r(x) &= \tfrac{1}{1-p_e} \sum_{r>0,\, b\geq 0} p(r,b)x^r \\ &= \tfrac{1}{1-p_e} \sum_{r>0\, b\geq 0} p_{r+b}\tbinom{r+b}{b} p_e^b (1-p_e)^r x^r \\ &= \tfrac{1}{1-p_e} \sum_{k=0}^{\infty} p_k \sum_{r=1}^{k} \tbinom{k}{r} p_e^{k-r}(1-p_e)^r x^r \\ &= \tfrac{1}{1-p_e} \sum_{k=0}^{\infty} p_k ((p_e+(1-p_e)x)^k - p_e^k) \\ &= \tfrac{1}{1-p_e}(f(p_e+(1-p_e)x) - f(p_e)) \\ &= \frac{f(p_e+(1-p_e)x)-p_e}{1-p_e}. \end{aligned}$$

Graphically we see that $f_r(x)$ is obtained by taking the graph of $f(x)$ in $[p_e,1]\times[p_e,1]$ and scaling it up to a $[0,1]\times[0,1]$ square. Also note that the new process is a supercritical Galton-Watson process with $p_0 = 0$.

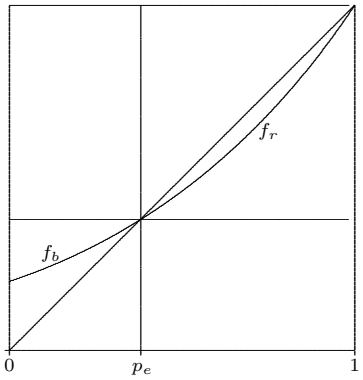


Fig. 5. Decomposing the supercritical process.

To summarise, we can consider a supercritical Galton-Watson process as being either a subcritical process with probability p_e, or a supercritical process with $p_0 = 1$ with probability $1 - p_e$, where in the second case we have various subcritical trees hanging off the main supercritical tree.

5. Subcritical Limit Law

In this section, and in Sections 7 and 8, we investigate the limiting behaviour of a Galton-Watson process in much more detail. As a result the arguments that we shall use are somewhat more technical in nature. Firstly we prove a general result on the limits of generating functions.

Theorem 5.1: *Suppose that for each n, $f_n(x) = \sum_{k=0}^{\infty} p_k^{(n)} x^k$ is a generating function of some probability distribution, and suppose that for $0 \le x < 1$, $f(x) = \lim_{n\to\infty} f_n(x)$ exists. Then provided $\lim_{x\to 1^-} f(x) = 1$, $f(x)$ is the generating function of some probability distribution. Moreover, if $f(x) = \sum_{k=0}^{\infty} p_k x^k$, then for each k, $\lim_{n\to\infty} p_k^{(n)} = p_k$.*

Proof: We shall weaken the hypothesis that the $p_k^{(n)}$ are probability distributions. Instead we shall just assume that there is some polynomial $P(k)$, independent of n, such that $|p_k^{(n)}| \le P(k)$ for all n and k. We shall show that all derivatives $f_n^{(k)}(x)$ converge pointwise for $x \in [0,1)$. Consider $f_n'(x) = \sum_{k=0}^{\infty}(k+1)p_{k+1}^{(n)}x^k$. The coefficients of $f_n'(x)$ are bounded by the polynomial $\tilde{P}(k) = (k+1)P(k+1)$. Also, if $x_0 \in (0,1)$ and ε is sufficiently small, say $x_0+2\varepsilon < 1$, then $|f_n''(x)| \le C = \sum_{k=0}^{\infty} P(k+2)(k+2)(k+1)(\frac{x_0+1}{2})^k < \infty$ for all $x \in [x_0, x_0+\varepsilon]$ and all n. Thus by Taylor's Theorem

$$|f_n(x_0+\varepsilon) - f_n(x_0) - \varepsilon f_n'(x_0)| \le C\varepsilon^2/2.$$

Hence for any n and m,

$$|(f_n(x_0+\varepsilon) - f_n(x_0) - \varepsilon f_n'(x_0)) - (f_m(x_0+\varepsilon) - f_m(x_0) - \varepsilon f_m'(x_0))| \le C\varepsilon^2,$$

so

$$\varepsilon|f_n'(x_0) - f_m'(x_0)| \le C\varepsilon^2 + |f_n(x_0+\varepsilon) - f_m(x_0+\varepsilon)| + |f_n(x_0) - f_m(x_0)|.$$

But $f_n(x_0)$ and $f_n(x_0+\varepsilon)$ converge as $n \to \infty$. Thus there is an N such that for all $n, m > N$,

$$|f_n'(x_0) - f_m'(x_0)| \le 2C\varepsilon.$$

Since this holds for all sufficiently small ε, $f_n'(x_0)$ converges as $n \to \infty$. By induction, for any $k \ge 0$, $f_n^{(k)}(x)$ converges pointwise for $0 \le x < 1$. In particular, $f_n^{(k)}(0) = k!p_k^{(n)}$ converges. Thus $p_k = \lim_{n\to\infty} p_k^{(n)}$ exists for all $k \ge 0$.

Now since $p_k^{(n)}$ are probability distributions, we have $p_k^{(n)} \in [0,1]$, so $p_k \in [0,1]$ for all k. Let $f(x) = \sum_{k=0}^{\infty} p_k x^k$. Since the $p_k^{(n)}$ are bounded, it is clear that $f_n(x) \to f(x)$ for $0 \le x < 1$. Moreover, $\lim_{x\to 1^-} f(x) = \sum_{k=0}^{\infty} p_k$,

so if this limit is 1 then $f(x)$ is the generating function of a probability distribution. □

Note that it is possible that $\lim_{x\to 1^-} f(x) < 1$. For example, if $f_n(x) = x^n$, then $f(x) = \lim_{n\to\infty} f_n(x) = 0$ for $0 \le x < 1$.

We now return to the Galton-Watson process. Recall that μ is the mean number of children of a given node and $f_n(x)$ is the generating function of the number of nodes Z_n at time n.

Lemma 5.1: *Suppose $0 < \mu < 1$. Then $(1 - f_n(t))/\mu^n$ decreases monotonically to a limit $K \ge 0$ for any t with $0 \le t < 1$. Moreover, K is strictly positive if and only if $\mathbb{E}(\xi \log \xi) < \infty$.*

Note: we consider $\xi \log \xi$ to be zero if $\xi = 0$.

Proof: Let $u_n = 1 - f_n(t)$ so that $f(1 - u_n) = 1 - u_{n+1}$. Note that by the Mean Value Theorem, $u_{n+1}/u_n = f'(x)$ for some $x \in (1 - u_n, 1)$, so in particular $u_{n+1}/u_n \le f'(1) = \mu$. Thus $(u_{n+1}/\mu^{n+1})/(u_n/\mu^n) \le 1$. Hence $(1 - f_n(t))/\mu^n = u_n/\mu^n$ is a decreasing sequence of positive real numbers, and so has a limit $K \ge 0$. The limit K is strictly positive provided $K/u_0 = \prod_{n=0}^{\infty}(u_{n+1}/\mu u_n) > 0$, which occurs if and only if $\sum_{n=0}^{\infty} \log(u_{n+1}/\mu u_n)$ converges. As $n \to \infty$, $u_n \to 1$, so the ratio $u_{n+1}/u_n = f'(x)$ converges to μ. In particular, $u_{n+1}/\mu u_n$ is bounded away from zero. In general, if $x_i \ge 0$ are real numbers with $1 - x_i$ bounded away from zero, then $x_i \le -\log(1 - x_i) \le Cx_i$ for some constant C. Thus convergence of $\sum_i \log(1 - x_i)$ is equivalent to the convergence of $\sum_i x_i$. Thus convergence of $\sum_{n=0}^{\infty} \log(u_{n+1}/\mu u_n)$ is equivalent to the convergence of

$$\sum_{n=0}^{\infty}(1 - \tfrac{u_{n+1}}{\mu u_n}) = \sum_{n=0}^{\infty} \tfrac{\mu u_n - (1 - f(1-u_n))}{\mu u_n}$$
$$= \sum_{n=0}^{\infty} \tfrac{f(1-u_n) - (1-\mu u_n)}{u_n^2}(u_n - u_{n+1})\tfrac{u_n}{\mu(u_n - u_{n+1})}.$$

But $\frac{u_n}{\mu(u_n - u_{n+1})} = \frac{1}{\mu(1 - u_{n+1}/u_n)} \to \frac{1}{\mu(1-\mu)} \ne 0$ as $n \to \infty$, so this is equivalent to convergence of

$$\sum_{n=0}^{\infty} \tfrac{f(1-u_n) - (1-\mu u_n)}{u_n^2}(u_n - u_{n+1}).$$

But $\frac{f(1-u_n) - (1-\mu u_n)}{u_n^2}$ is monotonically increasing to $f''(1)/2$ as $n \to \infty$, and this sum is a Riemann sum of the integral $\int_0^{u_0} \frac{f(1-u) - (1-\mu u)}{u^2}\, du$ obtained

by taking u_n as division points of the interval $[0, u_0]$. Thus convergence of the sum is equivalent to convergence of this integral. (The upper limit u_0 is unimportant for convergence, the question is what happens near $u = 0$.) Now $f(1-u) - (1-\mu u) = \sum_{k=0}^{\infty} p_k((1-u)^k - (1-ku))$ and all the terms $(1-u)^k - (1-ku)$ are non-negative (and zero for $k = 0, 1$). Thus

$$\int_0^1 \tfrac{f(1-u)-(1-\mu u)}{u^2}\, du = \sum_{k=2}^{\infty} p_k \int_0^1 \tfrac{(1-u)^k-(1-ku)}{u^2}\, du.$$

We estimate the inner integral by

$$\begin{aligned}
\int_0^1 \tfrac{(1-u)^k-(1-ku)}{u^2}\, du &= \int_0^{1/k} \tfrac{(1-u)^k-(1-ku)}{u^2}\, du + \int_{1/k}^1 \tfrac{(1-u)^k-(1-ku)}{u^2}\, du \\
&= \int_0^{1/k} \tbinom{k}{2} - \tbinom{k}{3}u + \ldots\, du + \int_{1/k}^1 (\tfrac{k}{u} + O(\tfrac{1}{u^2}))\, du \\
&= \int_{1/k}^1 \tfrac{k}{u}\, du + \int_0^{1/k} O(k^2)\, du + \int_{1/k}^1 O(\tfrac{1}{u^2})\, du \\
&= k \log k + O(k).
\end{aligned}$$

Since $\sum_{k=0}^{\infty} p_k k = \mu$ converges, the integral converges if and only if $\sum_{k=2}^{\infty} p_k(k \log k) = \mathbb{E}(\xi \log \xi)$ converges. □

Theorem 5.2: (Yaglom [7]**)** *If $\mu < 1$ then $\mathbb{P}(Z_n = k \mid Z_n > 0)$ converges as $n \to \infty$ to a probability distribution $(\tilde{p}_k)_{k=0}^{\infty}$. Moreover, $\mathbb{P}(Z_n > 0)/\mu^n$ decreases monotonically to $1/\tilde{\mu}$ as $n \to \infty$, where $\tilde{\mu} = \sum_{k=0}^{\infty} k\tilde{p}_k \in [1, \infty]$ is the mean of this limiting distribution. Also, $\tilde{\mu} < \infty$ if and only if $\mathbb{E}(\xi \log \xi) < \infty$.*

Proof: The generating function of $\mathbb{P}(Z_n = k \mid Z_n > 0) = \frac{\mathbb{P}(Z_n=k)}{\mathbb{P}(Z_n>0)}$ is given by

$$g_n(x) = \frac{f_n(x) - f_n(0)}{1 - f_n(0)} = 1 - \frac{1 - f_n(x)}{1 - f_n(0)}.$$

Define $h(x)$ by

$$h(x) = \frac{1 - f(x)}{1 - x}.$$

Then by convexity of $f(x)$, $h(x)$ increases monotonically to $f'(1) = \mu$ at $x = 1$. Now

$$\frac{1 - g_{n+1}(x)}{1 - g_n(x)} = \frac{1 - f_{n+1}(x)}{1 - f_{n+1}(0)} \cdot \frac{1 - f_n(0)}{1 - f_n(x)} = \frac{h(f_n(x))}{h(f_n(0))} \geq 1$$

Thus $g_n(x)$ is a decreasing function of n, and so has a limit, say $g(x)$, for all $0 \le x \le 1$. Now

$$1 - g_n(f(x)) = \frac{1 - f_{n+1}(x)}{1 - f_n(0)} = h(f_n(0))(1 - g_{n+1}(x)).$$

Taking limits as $n \to \infty$ gives

$$1 - g(f(x)) = \mu(1 - g(x)).$$

But $g(x)$ is increasing, so if $c = \lim_{x\to 1^-} g(x)$ then $1 - c = \mu(1 - c)$. Since $\mu < 1$, $c = 1$, so applying Theorem 5.1 we see that $\mathbb{P}(Z_n = k \mid Z_n > 0)$ converges to a distribution $(\tilde{p}_k)_{k=0}^{\infty}$ with generating function $g(x)$. We now estimate the mean of this distribution. By Lemma 5.1, $g_n'(1) = \mu^n/(1 - f_n(0))$ increases monotonically to $1/K$. But by Theorem 5.1,

$$\sum_{k=1}^{N} k\tilde{p}_k = \lim_{n\to\infty} \sum_{k=1}^{N} k\mathbb{P}(Z_n = k \mid Z_n > 0) \le \liminf_{n\to\infty} g_n'(1),$$

so letting $N \to \infty$, $g'(1) \le \liminf_{n\to\infty} g_n'(1)$. However, $g_n(x)$ is decreasing in n for all $x \le 1$ and $g_n(1) = 1$, so

$$g'(1) = \lim_{x\to 1^-} \frac{1 - g(x)}{1 - x} \ge \lim_{x\to 1^-} \frac{1 - g_n(x)}{1 - x} = g_n'(1)$$

for all n. Hence $g'(1) \ge \limsup_{n\to\infty} g_n'(x)$ and so $g'(1) = \lim_{n\to\infty} g_n'(x) = 1/K$. The last part now follows from the last part of Lemma 5.1 □

6. Moment Generating Functions

Before we examine the critical and supercritical cases in more detail, it will help to introduce the concept of the moment generating function of a distribution. This is very similar to the generating functions defined above, but applies more generally to random variables which take arbitrary real values, rather than just integer values. If X is a random variable that takes values in $[0, \infty)$, the *moment generating function* is defined by

$$L_X(\lambda) = \mathbb{E}(e^{-\lambda X})$$

for all $\lambda \ge 0$. We note that for integer valued X,

$$L_X(\lambda) = f_X(e^{-\lambda})$$

where f_X is the usual generating function of X. Also, $L_X(\lambda) \in (0, 1]$ for all $\lambda \ge 0$, $L_X(0) = 1$, and $L_X(\lambda)$ is a decreasing function of λ. If X has a probability density function, then L_X is just the Laplace transform of this

density function. Theorem 6.1 below is therefore essentially a result about the inverse Laplace transform of a function. First we prove that we can approximate a step function with a polynomial.

Lemma 6.1: *For any $\varepsilon > 0$ and $0 \le \alpha \le 1$ there exists a polynomial $P(x)$ such that*

$$\begin{aligned} 0 \le P(x) < \varepsilon \qquad & \text{for} \quad 0 \le x < \alpha - \varepsilon, \text{ and} \\ 1 - \varepsilon < P(x) \le 1 \qquad & \text{for} \quad \alpha + \varepsilon < x \le 1. \end{aligned}$$

Proof: Define for any $n \ge 0$,

$$P_n(x) = \sum_{r=\lceil \alpha n \rceil}^{n} \binom{n}{r} x^r (1-x)^{n-r},$$

so if $X = X_1 + \cdots + X_n$ is a binomial variable which is the sum of n independent 0-1 random variables X_i with $\mathbb{P}(X_i = 1) = x$, then

$$P_n(x) = \mathbb{P}(X \ge \alpha n).$$

Now $\mathbb{E}(X) = \sum_i \mathbb{E}(X_i) = nx$ and $\mathrm{Var}(X) = \sum_i \mathrm{Var}(X_i) = nx(1-x) \le n$. Thus by Markov's inequality

$$\mathbb{P}(|X - nx| \ge \varepsilon n) = \mathbb{P}((X - \mathbb{E}(X))^2 \ge \varepsilon^2 n^2) \le \frac{\mathrm{Var}(X)}{\varepsilon^2 n^2} \le \frac{1}{n\varepsilon^2},$$

so $P_n(x) = \mathbb{P}(X \ge \alpha n) \le \frac{1}{n\varepsilon^2}$ for $x < \alpha - \varepsilon$ and $1 - P_n(x) = \mathbb{P}(X < \alpha n) \le \frac{1}{n\varepsilon^2}$ for $x > \alpha + \varepsilon$. Hence we can take $P(x) = P_n(x)$ for any $n > 1/\varepsilon^3$. □

Theorem 6.1: *Suppose X_n, and X are non-negative random variables and*

$$\lim_{n\to\infty} \mathbb{E}(e^{-\lambda X_n}) = \mathbb{E}(e^{-\lambda X})$$

for all $\lambda \ge 0$. Suppose further that X is continuous, i.e., $\mathbb{P}(X \le c)$ is a continuous function of c. Then for all $c \ge 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_n \le c) = \mathbb{P}(X \le c).$$

Proof: For any polynomial $P(x) = \sum_{r=0}^{N} a_r x^r$ and any $\lambda \ge 0$ we have

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}(P(e^{-\lambda X_n})) &= \lim_{n\to\infty} \sum_{r=0}^{N} a_r \mathbb{E}(e^{-r\lambda X_n}) \\ &= \sum_{r=0}^{N} a_r \mathbb{E}(e^{-r\lambda X}) \\ &= \mathbb{E}(P(e^{-\lambda X})). \end{aligned}$$

Now for any $\alpha \in (0,1)$ and sufficiently small $\varepsilon > 0$, we can choose $P(x)$ as in Lemma 6.1. Then for any random variable $Y \in [0,1]$ we have

$$\mathbb{1}_{\{Y \geq \alpha+\varepsilon\}} - \varepsilon \leq P(Y) \leq \mathbb{1}_{\{Y \geq \alpha-\varepsilon\}} + \varepsilon,$$

so by taking expectations

$$\mathbb{P}(Y \geq \alpha+\varepsilon) - \varepsilon \leq \mathbb{E}(P(Y)) \leq \mathbb{P}(Y \geq \alpha-\varepsilon) + \varepsilon.$$

Hence for sufficiently large n and $\lambda = 1$

$$\begin{aligned}\mathbb{P}(e^{-X_n} \geq \alpha+\varepsilon) &\leq \mathbb{E}(P(e^{-X_n})) + \varepsilon \\ &\leq \mathbb{E}(P(e^{-X})) + 2\varepsilon \\ &\leq \mathbb{P}(e^{-X} \geq \alpha-\varepsilon) + 3\varepsilon.\end{aligned}$$

Setting $\alpha = e^{-c} - \varepsilon$ gives

$$\mathbb{P}(X_n \leq c) \leq \mathbb{P}(X \leq -\log(e^{-c} - 2\varepsilon)) + 3\varepsilon.$$

Similarly, taking $\alpha = e^{-c} + \varepsilon$ and $\lambda = 1$ we have for sufficiently large n

$$\begin{aligned}\mathbb{P}(e^{-X_n} \geq \alpha-\varepsilon) &\geq \mathbb{E}(P(e^{-X_n})) - \varepsilon \\ &\geq \mathbb{E}(P(e^{-X})) - 2\varepsilon \\ &\geq \mathbb{P}(e^{-X} \geq \alpha+\varepsilon) - 3\varepsilon,\end{aligned}$$

so

$$\mathbb{P}(X_n \leq c) \geq \mathbb{P}(X \leq -\log(e^{-c} + 2\varepsilon)) - 3\varepsilon.$$

Letting ε tend to 0 and by continuity of $\mathbb{P}(X \leq c)$ we have

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq c) = \mathbb{P}(X \leq c). \qquad \square$$

7. Critical Limit Law

Theorem 7.1: *If $\mu = 1$ and $0 < \sigma^2 < \infty$ then $\mathbb{P}(Z_n > 0) = \frac{2}{n\sigma^2}(1+o(1))$ and hence $\mathbb{E}(Z_n \mid Z_n > 0) = \frac{n\sigma^2}{2}(1+o(1))$. Moreover*

$$\mathbb{P}(Z_n/n \geq x \mid Z_n > 0) \to e^{-2x/\sigma^2} \quad \text{as} \quad n \to \infty,$$

i.e., conditioned on $Z_n > 0$, Z_n is approximately exponentially distributed with mean $\frac{n\sigma^2}{2}$.

Proof: Since $\sigma^2 < \infty$ we have $f''(1) = \sigma^2 - \mu(1-\mu) = \sigma^2$. Thus by Taylor's theorem,

$$f(1-u) = 1 - u + \tfrac{\sigma^2}{2}u^2 - o(u^2).$$

Indeed, since $f''(x)$ is increasing on $[0,1]$, the $o(u^2)$ term above is non-negative. Fix t, $0 \le t < 1$, and let $u_n = 1 - f_n(t)$. Then $f(1-u_n) = 1 - u_{n+1}$, so

$$u_{n+1} = u_n - \tfrac{\sigma^2}{2}u_n^2 + o(u_n^2).$$

By comparing with the differential equation

$$\tfrac{du}{dn} = -\tfrac{\sigma^2}{2}u^2$$

which has solution $u = \frac{2}{\sigma^2 n}$, one would expect that $1/u_n$ would grow like $\frac{\sigma^2 n}{2}$, i.e., linearly in n. To prove this we estimate

$$\frac{1}{u_{n+1}} - \frac{1}{u_n} = \frac{u_n - u_{n+1}}{u_n u_{n+1}} = \frac{\sigma^2 u_n^2/2 - o(u_n^2)}{u_n^2(1 - \sigma^2 u_n/2 + o(u_n))} = \frac{\sigma^2}{2} + o(1),$$

where the $o(1)$ term tends to zero as $u_n \to 0$, or equivalently (since $u_n \le 1 - f_n(0) = \mathbb{P}(Z_n > 0)$), as $n \to \infty$, uniformly in u_0. Hence

$$\frac{1}{u_n} - \frac{1}{u_0} = \sum_{i=0}^{n-1}\Big(\frac{1}{u_{i+1}} - \frac{1}{u_i}\Big) = \frac{\sigma^2}{2}n + o(n) = \frac{\sigma^2 n}{2}(1 + o(1)), \qquad (7.1)$$

where the $o(1)$ term tends to zero as $n \to \infty$, uniformly in u_0. Letting $t = 0$ we get $\mathbb{P}(Z_n > 0) = u_n = \frac{2}{n\sigma^2}(1 + o(1))$. Thus

$$\mathbb{E}(Z_n \mid Z_n > 0) = \frac{\mathbb{E}(Z_n \mathbb{1}_{\{Z_n > 0\}})}{\mathbb{P}(Z_n > 0)} = \frac{\mathbb{E}(Z_n)}{u_n} = \frac{1}{u_n} = \frac{\sigma^2 n}{2}(1 + o(1)).$$

Now let $t = e^{-\lambda/n}$ for some fixed $\lambda > 0$. Then

$$\frac{1}{u_0} = \frac{1}{1-t} = \frac{n}{\lambda} + O(1) \quad \text{as} \quad n \to \infty,$$

so by (7.1)

$$\frac{1}{u_n} = \frac{n\sigma^2}{2}(1 + o(1)) + \frac{n}{\lambda} + O(1) = n\Big(\frac{1}{\lambda} + \frac{\sigma^2}{2}\Big)(1 + o(1)).$$

But $u_n = 1 - f_n(t) = 1 - f_n(e^{-\lambda/n}) = \mathbb{E}(1 - e^{-\lambda Z_n/n})$. Also, $1 - e^{-\lambda Z_n/n} = 0$ when $Z_n = 0$, so

$$\mathbb{E}(1 - e^{-\lambda Z_n/n} \mid Z_n > 0) = \frac{\mathbb{E}(1 - e^{-\lambda Z_n/n})}{\mathbb{P}(Z_n > 0)} = \Big(\frac{1}{\lambda} + \frac{\sigma^2}{2}\Big)^{-1}\frac{\sigma^2}{2} + o(1).$$

Hence

$$\lim_{n\to\infty} \mathbb{E}(e^{-\lambda Z_n/n} \mid Z_n > 0) = 1 - \frac{\sigma^2/2}{1/\lambda + \sigma^2/2} = \frac{1}{1+\lambda\sigma^2/2},$$

but this last expression is equal to $\mathbb{E}(e^{-\lambda\sigma^2 X/2})$ where X is an exponential random variable with mean 1: $\mathbb{P}(X \geq c) = e^{-c}$. Hence by Theorem 6.1,

$$\mathbb{P}(Z_n/n \geq x) \to \mathbb{P}(\sigma^2 X/2 \geq x) = e^{-2x/\sigma^2}. \qquad \square$$

Corollary 7.1: *If $\mu = 1$ and $\sigma^2 < \infty$ then $\mathbb{E}(T) = \infty$, where $T = \inf\{n : Z_n = 0\}$ is the time to extinction.*

Proof: If $\sigma^2 = 0$ then $p_1 = 1$, so $T = \infty$ almost surely. Hence we may assume $\sigma^2 > 0$. Recall that $T = \mathbb{1}_{\{Z_0>0\}} + \mathbb{1}_{\{Z_1>0\}} + \dots$ so

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(Z_n > 0).$$

But by Theorem 7.1,

$$\mathbb{P}(Z_n > 0) = \frac{2}{n\sigma^2}(1 + o(1)),$$

But then $\mathbb{E}(T) = \sum_n \mathbb{P}(Z_n > 0)$ diverges, so $\mathbb{E}(T) = \infty$. $\square$

It is possible to construct distributions with $\mu = 1$ and $\mathbb{E}(T) < \infty$, however, by Corollary 7.1, all such distributions have infinite variance. For example, if we set $p_{4^r} = 8^{-r}$ for $r \geq 1$, and $p_0 = 1 - \sum_{r=1}^{\infty} 8^{-r} = \frac{6}{7}$. Then

$$\mu = \sum_{k=0}^{\infty} k p_k = \sum_{r=1}^{\infty} 4^r 8^{-r} = \sum_{r=1}^{\infty} 2^{-r} = 1.$$

Write $u_n = 1 - f_n(0) = \mathbb{P}(Z_n > 0)$. Then $f(1-u_n) = 1 - u_{n+1}$, so (since $\mu = 1$),

$$u_n - u_{n+1} = f(1-u_n) - (1-\mu u_n) = \sum_{k=0}^{\infty} p_k((1-u_n)^k - (1-ku_n)).$$

Since all the terms $(1-u_n)^k - (1-ku_n)$ are positive,

$$u_n - u_{n+1} \geq \sum_{n:\, ku_n \geq 1} (ku_n - 1) \geq \tfrac{1}{2} \sum_{n:\, ku_n \geq 2} ku_n.$$

Thus

$$u_n - u_{n+1} \geq \tfrac{1}{2} \sum_{r:\, 4^r u_n \geq 2} 8^{-r} 4^r u_n = 2^{-r_0} u_n \geq \sqrt{u_n/8}\, u_n = 8^{-1/2} u_n^{3/2},$$

where r_0 is chosen so that $\frac{2}{u_n} \leq 4^{r_0} < \frac{8}{u_n}$. However, one can prove by induction that $u_n \leq \frac{32}{n^2}$. Indeed, this clearly holds for $n \leq 4$, and

$$\frac{32}{n^2} - \frac{32}{(n+1)^2} = \frac{32(2n+1)}{n^2(n+1)^2} \leq \frac{64}{n^3}.$$

Now $x - 8^{-1/2}x^{3/2}$ is increasing for $0 \leq x \leq 2$, so

$$u_{n+1} \leq u_n - 8^{-1/2}u_n^{3/2} \leq \tfrac{32}{n^2} - 8^{-1/2}(\tfrac{32}{n^2})^{3/2} = \tfrac{32}{n^2} - \tfrac{64}{n^3} \leq \tfrac{32}{(n+1)^2}.$$

But then

$$\mathbb{E}(T) = \sum_{n=0}^{\infty} \mathbb{P}(Z_n > 0) = u_0 + \sum_{n=1}^{\infty} u_n \leq 1 + \sum_{n=1}^{\infty} \frac{32}{n^2} < \infty.$$

8. Supercritical Limit Law

Now assume $\mu > 1$. For simplicity we shall just consider the case when $\mu < \infty$.

Let $W_n = Z_n/\mu^n$. Then $\mathbb{E}(W_n) = 1$ for all n, and

$$\mathrm{Var}(W_n) = \frac{\sigma^2}{\mu(\mu-1)}(1 - \mu^{-n}) \to \frac{\sigma^2}{\mu(\mu-1)}$$

provided $\sigma^2 < \infty$. Indeed, we can say more. If W_{n-1} (and hence Z_{n-1}) is given, then $\mathbb{E}(W_n \mid W_{n-1}) = W_{n-1}$. Since Z_n, and hence W_n, depends on $Z_{n-1}, Z_{n-2}, \ldots, Z_0$ only via the value of Z_{n-1}, one can write

$$\mathbb{E}(W_n \mid W_{n-1}, W_{n-2}, \ldots, W_0) = W_{n-1}.$$

This says that W_n is a *martingale*. The Martingale Convergence Theorem (see [6]) says that if W_n is a non-negative martingale then W_n converges, i.e., there is a random variable W taking values in $[0, \infty)$ such that $W_n \to W$ almost surely. Note that this is a very strong condition analogous to the Strong Law of Large Numbers: if we look at any instance of the sequence Z_n/μ^n then this almost surely converges, so Z_n is almost surely of the form $(c + o(1))\mu^n$ for some (random) $c \geq 0$.

Unfortunately, this leaves many questions unanswered. For example it is entirely possible that W is identically zero. Indeed this can happen.

Theorem 8.1: (Kesten-Stigum [2, 3]**)** *If $\mu > 1$ and $\mathbb{E}(\xi \log \xi) < \infty$ then $\mathbb{E}(W) = 1$ and $\mathbb{P}(W = 0) = \mathbb{P}(Z_n \to 0) < 1$. If in addition $p_k \neq 1$ for all k then W has a continuous and strictly positive density function for all $W > 0$. Conversely, if $\mathbb{E}(\xi \log \xi) = \infty$ then $W = 0$ almost surely.*

The situation when $\mathbb{E}(\xi \log \xi) = \infty$ indicates that the normalisation μ^n in $W_n = Z_n/\mu^n$ was not a good choice. In fact it can be shown that there is always a normalising sequence C_n such that Z_n/C_n converges to a non-trivial random variable.

In the case when $\sigma^2 < \infty$ one can give the following heuristic argument for Theorem 8.1. If $Z_{n-1} = k$ then Z_n is a sum of k independent identically distributed random variables. By the Central Limit Theorem, we expect that Z_n is approximately normal $N(k\mu, k\sigma^2)$, with mean $k\mu$ and variance $k\sigma^2$. Normalising, we get that, conditional on $W_{n-1} = c$, $W_n \sim N(c, c\sigma^2/\mu^n)$, so that $W_n \sim W_{n-1} + N(0, c\sigma^2/\mu^n)$. But $c\sigma^2/\mu^n$ is decreasing very rapidly and so one expects that $|W_n - W_{n-1}|$ is exponentially small. Thus W_n converges to some W. Moreover, the distribution of W should be very smooth, since at each step we are convoluting the distribution of W_{n-1} with something that looks very much like a normal distribution.

We shall not give a complete proof Theorem 8.1 here. In particular we shall not show that W has a strictly positive density function. But we will show how the $\mathbb{E}(\xi \log \xi)$ condition appears. Indeed, it appears for much the same reason as it did in the subcritical case, by looking at a limiting ratio u_n/μ^n. In this case u_0 should be taken very close to 0 since u_n is increasing geometrically. Let

$$L_n(\lambda) = \mathbb{E}(e^{-\lambda W_n})$$

be the moment generating function of W_n. Fix $\lambda > 0$ and $N > 0$ and let $u_n = 1 - f_n(e^{-\lambda/\mu^N})$ so that

$$u_0 = 1 - e^{-\lambda/\mu^N} \approx \lambda/\mu^N, \qquad \text{and} \qquad 1 - u_N = \mathbb{E}(e^{-\lambda W_N}) = L_N(\lambda).$$

We wish to estimate u_N, or equivalently the ratio

$$R = (u_N/\mu^N)/(u_0/\mu^0) \approx u_N/\lambda.$$

To do this we consider the product $R = \prod_{n=0}^{N-1}(u_{n+1}/\mu u_n)$. As in the subcritical case, $u_{n+1}/\mu u_n \leq 1$, so as $N \to \infty$ either R tends to a positive limit, or it tends to 0. Now

$$\log(R) = \sum_{n=0}^{N-1} \log(u_{n+1}/\mu u_n),$$

which given $u_{n+1}/\mu u_n$ is bounded away from zero is within a constant

factor of

$$\sum_{n=0}^{N-1}(1-u_{n+1}/\mu u_n).$$

As in Lemma 5.1, this can be estimated in terms of the integral

$$\int_{u_0}^{u_N}\frac{f(1-u)-(1-\mu u)}{u^2}\,du.$$

If we let $N\to\infty$ then $u_0=1-e^{-\lambda/\mu^N}$ tends to 0. Hence $\log(R)$ converges if and only if the integral $\int_0^c \frac{f(1-u)-(1-\mu u)}{u^2}\,du$ converges, which by the same argument as in Lemma 5.1 is if and only if $\mathbb{E}(\xi\log\xi)<\infty$. Thus if $\mathbb{E}(\xi\log\xi)<\infty$ then as $N\to\infty$, u_N is bounded away from zero and so $\mathbb{E}(e^{-\lambda W_N})$ converges to some value less than 1. On the other hand, if $\mathbb{E}(\xi\log\xi)=\infty$ then $u_N\to 0$ and so $\mathbb{E}(e^{-\lambda W_N})\to 1$.

Now using the Martingale Convergence Theorem, $W_n\to W$ almost surely. Hence $e^{-\lambda W_n}\to e^{-\lambda W}$ almost surely. But $e^{-\lambda W_n}\in[0,1]$ so by the Dominated Convergence Theorem (dominated by the constant 1),

$$\lim_{n\to\infty}L_n(\lambda)=\lim_{n\to\infty}\mathbb{E}(e^{-\lambda W_n})=\mathbb{E}(e^{-\lambda W})=L(\lambda)$$

where $L(\lambda)$ is the moment generating function of W. Thus if $\mathbb{E}(\xi\log\xi)=\infty$, $\mathbb{E}(e^{-\lambda W})=1$, and so by Theorem 6.1, $W=0$ almost surely. On the other hand, if $\mathbb{E}(\xi\log\xi)<\infty$, then $\mathbb{E}(e^{-\lambda W})<1$ for $\lambda>0$. Hence W is not identically zero. But

$$f(L_n(\lambda))=f(f_n(e^{-\lambda/\mu^n}))=f_{n+1}(e^{-\lambda/\mu^n})=L_{n+1}(\mu\lambda).$$

Letting $n\to\infty$ and using continuity of $f(x)$ we get

$$f(L(\lambda))=L(\mu\lambda). \tag{8.1}$$

If we let $\lambda\to+\infty$, $L(\lambda)=\mathbb{E}(e^{-\lambda W})$ decreases monotonically to a limit, which by the Dominated Convergence Theorem is just $\mathbb{E}(\mathbb{1}_{\{W=0\}})=\mathbb{P}(W=0)$. But by (8.1), this limit must be a solution to $f(x)=x$. Thus $\mathbb{P}(W=0)$ is either 1 or the extinction probability p_e. In the case that $\mathbb{E}(\xi\log\xi)<\infty$ we know $\mathbb{P}(W=0)\neq 1$, so $\mathbb{P}(W=0)=p_e=\mathbb{P}(Z_n\to 0)$.

9. Total Number of Nodes

In this section we consider the critical and subcritical cases $\mu\le 1$, $p_1<1$, and ask about the distribution of the total number of nodes in the Galton-Watson tree. Write $S=\sum_{n=0}^{\infty}Z_n$ for the total number of nodes. We know

that $\mathbb{P}(S < \infty) = 1$ when $\mu \leq 1$, so S gives a probability distribution on $\mathbb{N} = \{0, 1, \dots\}$.

Lemma 9.1: *For $\mu \leq 1$, the generating function $f_S(x)$ of the total number of nodes $S = \sum_{n=0}^{\infty} Z_n$ satisfies the following equation:*

$$f_S(x) = xf(f_S(x)).$$

Proof: Let v_0 be the root node of the Galton-Watson tree and consider each child node $v_1, \dots, v_k$. If we let S_i be the total number of nodes in the tree starting at v_i then $S = 1 + S_1 + S_2 + \cdots + S_k$. Moreover, the S_i are independent and have the same distribution as S. Thus in terms of generating functions

$$\mathbb{E}(x^S \mid Z_1 = k) = \mathbb{E}(x^{1+S_1+\cdots+S_k}) = x\mathbb{E}(x^{S_1})\mathbb{E}(x^{S_2})\cdots = xf_S(x)^k.$$

Hence

$$f_S(x) = \sum_{k=0}^{\infty} \mathbb{E}(x^S \mid Z_1 = k)\mathbb{P}(Z_1 = k) = x\sum_{k=0}^{\infty} p_k f_S(x)^k = xf(f_S(x)). \quad \square$$

Thus $f_S(x)$ is a solution to the equation $y = xf(y)$. Put more simply, let $f_S^{-1}(y)$ be the inverse function of $f_S(x)$ on $[0, 1]$. Then

$$f_S^{-1}(y) = y/f(y).$$

Note that since $p_0 > 0$, and by convexity of $f(x)$, the function $y/f(y)$ increases from 0 at $y = 0$ to 1 at $y = 1$.

Lemma 9.2: *For $\mu < 1$, $\mathbb{E}(S) = \frac{1}{1-\mu}$. If in addition we have $\sigma^2 < \infty$, then $\mathrm{Var}(S) = \frac{\sigma^2}{(1-\mu)^3}$.*

Proof: Implicitly differentiating the equation $y = xf(y)$ gives

$$\begin{aligned} y' &= xf'(y)y' + f(y) \\ y'' &= xf''(y)y'^2 + xf'(y)y'' + 2f'(y)y' \end{aligned}$$

Setting $x = y = 1$, $f'(1) = \mu$, gives $y' = \mu y' + 1$, so $y' = \frac{1}{1-\mu}$. In addition, setting $f''(1) = \sigma^2 - \mu(1-\mu)$ gives

$$y'' = (\sigma^2 - \mu(1-\mu))y'^2 + \mu y'' + 2\mu y'.$$

Substituting $y' = \frac{1}{1-\mu}$ gives

$$(1-\mu)y'' = \tfrac{\sigma^2}{(1-\mu)^2} - \tfrac{\mu}{1-\mu} + \tfrac{2\mu}{1-\mu} = \tfrac{\sigma^2}{(1-\mu)^2} + \tfrac{\mu}{1-\mu},$$

hence

$$\mathrm{Var}(S) = y'' + y'(1-y') = \tfrac{\sigma^2}{(1-\mu)^3} + \tfrac{\mu}{(1-\mu)^2} - \tfrac{\mu}{(1-\mu)^2} = \tfrac{\sigma^2}{(1-\mu)^3}. \qquad \square$$

For $\mu = 1$ we see that $\mathbb{E}(S) = \infty$. But $S < \infty$ almost surely. So what does the asymptotic distribution of S look like?

As an example, suppose the number of child nodes is Poisson with mean μ, so the generating function for the process is $f(x) = e^{\mu(x-1)}$. We shall calculate the exact distribution of S in this case. By Lemma 9.1, the generating function for S is the solution to the equation $x = ye^{\mu(1-y)}$. We shall find an explicit form of this function.

Lemma 9.3: *Suppose $P(k)$ is a polynomial in k of degree less than n. Then $\sum_{k=0}^{n}(-1)^k\binom{n}{k}P(k) = 0$.*

Proof: If $n = 1$ then $P(k) = c$ is a constant and $\sum_{k=0}^{n}(-1)^k\binom{n}{k}P(k) = c - c = 0$. For $n > 1$ we use the identity $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ to get

$$\begin{aligned}\sum_{k=0}^{n}(-1)^k\tbinom{n}{k}P(k) &= \sum_{k=0}^{n-1}(-1)^k\tbinom{n-1}{k}P(k) + \sum_{k=1}^{n}(-1)^k\tbinom{n-1}{k-1}P(k)\\ &= \sum_{k=0}^{n-1}(-1)^k\tbinom{n-1}{k}P(k) - \sum_{j=0}^{n-1}(-1)^j\tbinom{n-1}{j}P(j+1)\\ &= \sum_{k=0}^{n-1}(-1)^k\tbinom{n-1}{k}(P(k) - P(k+1))\end{aligned}$$

But $P(k)-P(k+1)$ is a polynomial of degree less than $n-1$ so, by induction on n, the last expression above is zero. $\square$

Lemma 9.4: *The solution to $x = ye^{-y}$, $y \in [0,1]$, $x \in [0, e^{-1}]$, is given by the power series $y = \sum_{k=1}^{\infty}\frac{k^{k-1}}{k!}x^k$.*

Proof: By Stirling's formula, $k! \sim (k/e)^k\sqrt{2\pi k}$, so $\frac{k^{k-1}}{k!} \sim e^k\sqrt{\frac{1}{2\pi k^3}}$. Hence the power series converges for all $x \in [0, e^{-1}]$. Substituting $x = ye^{-y}$ into the power series gives

$$\begin{aligned}\sum_{k=1}^{\infty}\frac{k^{k-1}}{k!}y^ke^{-ky} &= \sum_{k=1}^{\infty}\frac{k^{k-1}}{k!}\sum_{n=k}^{\infty}\frac{(-k)^{n-k}}{(n-k)!}y^n\\ &= \sum_{n=1}^{\infty}(-1)^n\frac{y^n}{n!}\sum_{k=1}^{n}(-1)^k\tbinom{n}{k}k^{n-1}. \qquad (9.1)\end{aligned}$$

(The interchange of summation is justified by absolute convergence of the double sum when $ye^y < e^{-1}$.) However, by Lemma 9.3, $\sum_{k=0}^{n}(-1)^k\binom{n}{k}k^{n-1} = 0$ for all $n \geq 1$, and the $k = 0$ term is zero for $n > 1$. Thus the inner sum in (9.1) vanishes for all $n \neq 1$. Hence

$$\sum_{k=1}^{\infty} \frac{k^{k-1}}{k!} y^k e^{-ky} = y$$

for sufficiently small y. Thus the power series gives the solution to $x = ye^{-y}$ for small y. Since both the power series and the inverse function $y = y(x)$ are analytic for $x \in [0, e^{-1})$, they agree for $x \in [0, e^{-1})$. Continuity of both functions imply that they also agree at $x = e^{-1}$. □

Corollary 9.5: *For the branching process with child distribution given by a Poisson distribution with mean $\mu \leq 1$, the probability that $\sum_{n=0}^{\infty} Z_n = k$ is given exactly for $k \geq 1$ by*

$$\mathbb{P}(S = k) = \frac{(\mu k)^{k-1}}{k!} e^{-\mu k}.$$

Proof: If $y = f_S(x)$ is the generating function for $S = \sum_{n=0}^{\infty} Z_n$ then $x = ye^{\mu(1-y)}$. Thus $\mu e^{-\mu} x = \mu y e^{-\mu y}$. Hence $\mu y = \sum_{k=1}^{\infty} \frac{k^{k-1}}{k!}(\mu e^{-\mu} x)^k$. Taking the coefficient of x^k gives the result. □

Using Stirling's Formula, $k! \sim (\frac{k}{e})^k \sqrt{2\pi k}$, we can approximate

$$\frac{(\mu k)^{k-1}}{k!} e^{-\mu k} \sim \frac{(\mu e^{1-\mu})^k}{\sqrt{2\pi k^3}}.$$

For $\mu = 1$ this simplifies to $\frac{1}{\sqrt{2\pi k^3}}$. In particular it is now clear why $\mathbb{E}(S) = \infty$. Indeed, this is due to the fact that the sum $\sum_k k\frac{1}{\sqrt{2\pi k^3}} = \sum_k ck^{-1/2}$ diverges.

10. Trees and Branching numbers

Let T be a *locally finite* tree, i.e., a (possibly infinite) connected graph with no cycles and for which each vertex has finite degree. Note that the degrees of the vertices need not be bounded over the whole tree. Fix a root vertex $v_0 \in V(T)$. We define the *level* $\ell(v)$ of a vertex $v \in V(T)$ to be the graph distance in T from v to v_0. For any vertex $v \in V(T)$, the *children* of v are the vertices u such that $\ell(u) = \ell(v) + 1$ and $uv \in E(T)$. For any node v, define T_v to be the subtree of T consisting of v and all of its descendants,

i.e., all nodes u such that the unique path from u to v_0 contains v. We consider v to be the root of T_v.

Definition 10.1: A *flow* on an infinite (but locally finite) tree T is a non-negative function on the vertices, $f\colon V(T) \to \mathbb{R}$, such that if $v_1, \ldots, v_r$ are the children of a node v then $f(v) = \sum_{i=1}^{r} f(v_i)$. We call the flow *non-trivial* if f is not identically zero. Equivalently f is non-trivial if and only if $f(v_0) \neq 0$.

Definition 10.2: A *cut* of T is a finite subset $V_0 \subseteq V(T)$ of vertices whose removal makes the root v_0 part of a finite component (or such that $v_0 \in V_0$). If $b > 0$ then we define the b-weight $w_b(V_0)$ of a cut V_0 to be $\sum_{v \in V_0} b^{-\ell(v)}$.

Lemma 10.1: *For all $b > 0$, the following are equivalent.*

1. *There exists a non-trivial flow with $b^{\ell(v)} f(v)$ bounded on T.*
2. *There is a constant $c > 0$ such that all cuts have b-weight at least c.*

Proof:
$1 \Rightarrow 2$: Fix a non-trivial flow such that $b^{\ell(v)} f(v) \leq C$ for all vertices v. We claim that for any cut V_0, $\sum_{v \in V_0} f(v) \geq f(v_0) > 0$. We prove this by induction on the maximum level of any node in V_0. If $v_0 \in V_0$ then the result is clear. Otherwise let $v_1, \ldots, v_k$ be the children of v_0. Then we can decompose $V_0 = V_1 \cup \cdots \cup V_k$ where V_i, $i = 1, \ldots, k$, are cuts of the subtrees T_{v_i}. By induction $\sum_{v \in V_i} f(v) \geq f(v_i)$, and so $\sum_{v \in V_0} f(v) \geq \sum_{i=1}^{k} f(v_i) = f(v_0)$. Now $f(v) \leq C b^{-\ell(v)}$, so $\sum_{v \in V_0} f(v) \leq C \sum_{v \in V_0} b^{-\ell(v)} = C w_b(V_0)$. Hence $w_b(V_0) \geq f(v_0)/C > 0$.
$2 \Rightarrow 1$: Let $g(v) = \inf_{V_v} w_b(V_v)$ where V_v runs over all cuts of the subtree T_v. If $v_1, \ldots, v_k$ are the children of v, and $V_1, \ldots, V_k$ are cuts of $T_{v_1}, \ldots, T_{v_k}$ respectively, then $V_1 \cup \cdots \cup V_k$ is a cut of T_v. Hence

$$g(v) \leq \sum_{i=1}^{k} g(v_i).$$

The inequality may be strict, since $\{v\}$ cuts T_v and may give a lower weight than any cut obtained as a union $V_1 \cup \cdots \cup V_k$. However, we can define a flow bounded by $g(v)$ by setting $f(v_0) = g(v_0)$ and inductively defining f on the children of a node v by

$$\frac{f(v_i)}{f(v)} = \frac{g(v_i)}{\sum_{i=1}^{k} g(v_i)}$$

(or $f(v_i) = 0$ if $f(v) = 0$). Then $f(v_i)/f(v) \le g(v_i)/g(v)$, so by induction on $\ell(v)$, $f(v) \le g(v)$ for all v. Clearly f is a flow and $f(v_0) = g(v_0) \ge c > 0$. But $V_v = \{v\}$ is a cut of T_v. Hence $f(v) \le g(v) \le w_b(\{v\}) = b^{-\ell(v)}$ and so $b^{\ell(v)} f(v) \le 1$. □

Definition 10.3: The *branching number* of T is given by

$$\begin{aligned}\mathrm{br}(T) &= \sup\{\, b : \exists \text{ a non-trivial flow } f \text{ such that } b^{\ell(v)} f(v) \text{ is bounded} \,\} \\ &= \inf\{\, b : \exists \text{ cuts with arbitrarily small } b\text{-weight} \,\}\end{aligned}$$

Note that by Lemma 10.1, these are the same.

It is clear that $\mathrm{br}(T) \ge 1$ for any infinite tree, since for such a tree we can define a flow that is 1 on some infinite path from v_0 and zero elsewhere. Furthermore, $\mathrm{br}(T)$ is independent of the choice of the root.

If we let $T^n = \{v : \ell(v) = n\}$ be the vertices of T at level n, then the number $|T^n|$ of such vertices may behave very erratically with n. It is clear that

$$\mathrm{br}(T) \le \liminf_{n\to\infty} |T^n|^{1/n}.$$

(Use the cut $V_0 = T^n$ in Definition 10.3), however this inequality may be strict. For example, consider the T in Figure 6. At level n there are $2^{n+1} - 1$ vertices which are linearly ordered. The first 2^n vertices of T^n all have 3 children (making up the first $3 \cdot 2^n$ vertices of T^{n+1}), but all the remaining vertices of T^n have exactly one child. Clearly $\liminf_{n\to\infty} |T^n|^{1/n} = 2$. However, every vertex other than the first one in T^n has a bounded number of descendants at each level $m > n$. This is because all its descendants at level $m > n$ which have 3 children must have 3^{m-n} vertices before them in T^m, and for large m $3^{m-n} > 2^n$. Thus ultimately all the descendants have one child each. Now suppose we have a flow on T with $b^{\ell(v)} f(v)$ bounded for some $b > 1$. Each vertex $v \in T^n$ other than the first must then have $f(v) = 0$. But then the first vertex must have $f(v) = f(v_0)$. But since $b^{\ell(v)} f(v)$ must be bounded, we have $f(v_0) = 0$ and the flow is trivial. Thus $\mathrm{br}(T) = 1$.

For a regular tree of degree $k+1$ (so each vertex has exactly k children), we have $\mathrm{br}(T) = k$. Indeed, $\mathrm{br}(T) \le \liminf |T^n|^{1/n} = k$, while the flow defined by $f(v) = k^{-\ell(v)}$ for all v shows that $\mathrm{br}(T) \ge k$.

We now introduce the concept of *percolation* on T. Fix $p \in [0, 1]$. For each node v of T, we randomly declare v to *open* with probability p. Otherwise v will be *closed*. Moreover, the choice of the state of each node is

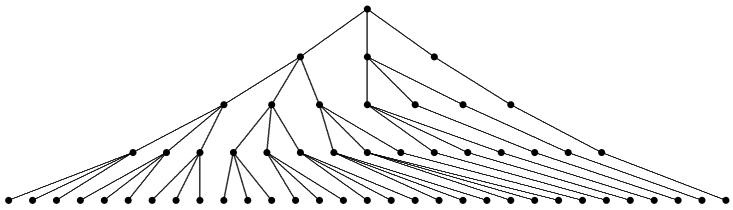


Fig. 6. A tree with $\mathrm{br}(T) < \liminf_{n\to\infty} |T^n|^{1/n}$.

made independently of the states of every other node of T. A path in T is then said to be *open* if all the vertices of P are open. Hence any fixed path on n vertices will be open with probability p^n.

Theorem 10.1: (Theorem 6.2 of [5]**)** *Suppose each node of a locally finite infinite tree T is declared to be open with probability p, independently of all other nodes. If $p < 1/\mathrm{br}(T)$ there is almost surely no infinite open path from v_0, and if $p > 1/\mathrm{br}(T)$ then an infinite open path from v_0 exists with positive probability.*

Proof: Suppose first that $p < \frac{1}{\mathrm{br}(T)}$. Then $\frac{1}{p} > \mathrm{br}(T)$, so there exist cuts with arbitrarily small $\frac{1}{p}$-weight. Fix a cut V_0 with $w_{1/p}(V_0) < \varepsilon$. Suppose P is an infinite open path from v_0. If v is any vertex in P then all the vertices on the unique path from v_0 to v must be open, and there are $\ell(v)+1$ such vertices (including v_0 and v themselves). Hence

$$\mathbb{P}(\exists \text{ open path } P \text{ containing } v) \le p^{\ell(v)+1}.$$

But V_0 is a cut, so any infinite path must meet some vertex of V_0. Hence

$$\mathbb{P}(\exists \text{ infinite open path from } v_0) \le \sum_{v\in V_0} p^{\ell(v)+1} = p w_{1/p}(V_0) < p\varepsilon.$$

Since this holds for all $\varepsilon > 0$, there is almost surely no infinite open path from v_0.

Now suppose that $p > \frac{1}{\mathrm{br}(T)}$. Choose $b > \mathrm{br}(T)$ and $\varepsilon > 0$ so that $p = \frac{1}{b} + \varepsilon$. Let f be a flow such that $b^{\ell(v)} f(v) \le \varepsilon$ for all v. (By scaling, we can assume such a flow exists). Fix $N > 0$ and let p_v be the probability that there exists an open path from v down to some vertex at level N. We shall show by reverse induction on the level that $p_v \ge f(v) b^{\ell(v)}$. At level N, $p_v = p > \varepsilon \ge f(v) b^{\ell(v)}$. Now suppose that the result is true if $\ell(v) = n$ and v is a node at level $n-1$. Let $v_1, \ldots, v_k$ be the children of v. Then

$p_{v_i} \geq f(v_i)b^n$. If there is no open path to level N from v, then either v is not open or there is no open path to level N from v_i for each v_i. Hence

$$1 - p_v/p = \prod_{i=1}^{k}(1 - p_{v_i}) \leq \prod_{i=1}^{k}(1 - f(v_i)b^n)$$
$$\leq \exp\Big(-\sum_{i=1}^{k} f(v_i)b^n\Big) = \exp(-f(v)b^n).$$

Now $f(v)b^n \leq \varepsilon b$ and

$$\exp(-x) \leq \tfrac{1}{1+x} = 1 - \tfrac{x}{1+x} \leq 1 - \tfrac{x}{1+\varepsilon b} = 1 - \tfrac{x}{pb}$$

for $x < \varepsilon b$. Hence $1 - p_v/p \leq 1 - f(v)b^n/pb$. Thus $p_v \geq f(v)b^{n-1}$. Hence $p_v \geq f(v)b^{\ell(v)}$ for all v, and in particular $p_{v_0} \geq f(v_0)$. If we let E_N be the event that there exists an open path from v_0 to some vertex at level N, then $\mathbb{P}(E_N) = p_{v_0} \geq f(v_0)$. Thus by continuity of probability, with probability at least $f(v_0)$ there exists open paths down to any level. However, in this case there is an infinite open path from v_0. (If there are arbitrarily long open paths down from v then this must also hold for at least one of the children of v. Thus one can construct an open path by always taking the next vertex as one such child.) □

Theorem 10.2: *If T is a Galton-Watson tree with distribution $(p_k)_{k=0}^{\infty}$ of mean $\mu > 1$, then conditional on survival,* $\mathrm{br}(T) = \mu$ *almost surely.*

Proof: Let T be a Galton-Watson tree with generating function $f(x)$. First we show that if $\mathbb{P}(\mathrm{br}(T) \geq b) > 0$, then conditional on survival, $\mathrm{br}(T) \geq b$ almost surely. Let E_b be the event that $\mathrm{br}(T) < b$. It is easy to see that $\mathrm{br}(T) < b$ if and only if $\mathrm{br}(T_v) < b$ for every child v of v_0. Thus if $\rho = \mathbb{P}(E_b)$, then

$$\rho = \sum_{k=0}^{\infty} p_k \rho^k = f(\rho).$$

If $\mathbb{P}(\mathrm{br}(T) \geq b) > 0$ then $\rho = \mathbb{P}(E_b) < 1$. But $\rho = f(\rho)$, so $\rho = p_e$, the extinction probability of the process. Since $E_b \supseteq \{\text{extinction}\}$, we have $\mathbb{P}(\mathrm{br}(T) \geq b \mid \text{survival}) = (1 - \rho)/(1 - p_e) = 1$.

Now fix $p \in [0, 1]$ and delete vertices (and all their descendants) from T independently with probability $1 - p$ (and independently of the process that gave rise to T). The result is a new Galton-Watson process T' with mean $p\mu$. Indeed, the number of surviving children of a surviving node is given by the generating function $f((1 - p) + px)$. On the other hand, the

deletion of vertices is equivalent to declaring vertices open with probability p and ignoring any vertex that does not lie on an open path from v_0. Thus the pruned process T' survives if and only if there is an infinite open path from v_0 in T.

The pruned tree T' survives with positive probability if and only if $p\mu > 1$. But we want to know whether the pruned version survives given a *fixed* instance of T. For this we need the conditional probability $Y = \mathbb{P}(\text{survives} \mid T)$. Now

$$\mathbb{P}(\text{survives}) = \mathbb{E}(\mathbb{1}_{\{\text{survives}\}}) = \mathbb{E}(\mathbb{E}(\mathbb{1}_{\{\text{survives}\}} \mid T)) = \mathbb{E}(\mathbb{P}(\text{survives} \mid T)).$$

If $p\mu \leq 1$ then $\mathbb{E}(Y) = \mathbb{P}(\text{survives}) = 0$, so, since $Y \geq 0$, $Y = 0$ almost surely. In other words, for almost all instances of the tree T, $\mathbb{P}(\text{survives} \mid T) = 0$, and so by Theorem 10.1, $\text{br}(T) \leq 1/p$. Hence $\text{br}(T) \leq \mu$ almost surely. Conversely, if $p\mu > 1$ then $\mathbb{E}(Y) = \mathbb{P}(\text{survives}) > 0$, so $Y > 0$ with positive probability. Thus $\text{br}(T) \geq 1/p$ with positive probability, and hence $\text{br}(T) \geq 1/p$ almost surely given survival. Since this holds whenever $p\mu > 1$, $\text{br}(T) \geq \mu$ almost surely given survival. □

11. Multi-type Galton-Watson Processes

In this last section we generalise the concept of a Galton-Watson process to include nodes of different types. More specifically we specify that each node is one of a finite number of types $\{1, 2, \ldots, N\}$ and for each type i, a node of type i has a random number $\xi_j^{(i)}$ of child nodes of type j. We do not assume that $\xi_1^{(i)}, \ldots, \xi_N^{(i)}$ are independent, so for example, the number of children of type 2 may be dependent on the number of children of type 1. However, as before, the number and type of children of distinct nodes are independent. We replace Z_n, the number of nodes at time n, with a vector $\mathbf{Z}_n^{(i)} = (Z_{n,1}^{(i)}, \ldots, Z_{n,N}^{(i)})$ of the number of nodes of each type at time n, assuming that we start with just one type i node at time 0. Define a vector-valued generating function $\mathbf{f}(\mathbf{x}) = (f_1(x_1, \ldots, x_N), f_2(x_1, \ldots, x_N), \ldots)$ where

$$f_i(x_1, \ldots, x_N) = \sum_{k_1, k_2, \ldots, k_N} \mathbb{P}(\xi_1^{(i)} = k_1,\, \xi_2^{(i)} = k_2,\, \ldots) x_1^{k_1} x_2^{k_2} \ldots$$

Define $\mathbf{f}_n(\mathbf{x})$ similarly using the random variables $Z_{n,j}^{(i)}$ in place of $\xi_j^{(i)}$. Let $M = (\mu_{ij})$ be the matrix of the average number of type j children of a type i node, so $\mu_{ij} = \mathbb{E}(\xi_j^{(i)})$.

Example: Suppose that each type i node has a Poisson $Po(\mu_{ij})$ number of children of type j, independently for each j. Then $M = (\mu_{ij})$ and

$$\begin{aligned} f_i(\mathbf{x}) &= \sum_{k_1,k_2,\dots,k_N} \mathbb{P}(\xi_1^{(i)} = k_1,\, \xi_2^{(i)} = k_2,\, \dots) x_1^{k_1} x_2^{k_2} \dots \\ &= \sum_{k_1,k_2,\dots,k_N} \mathbb{P}(\xi_1^{(i)} = k_1)\mathbb{P}(\xi_2^{(i)} = k_2) \dots x_1^{k_1} x_2^{k_2} \dots \qquad \text{(Independence)} \\ &= \sum_{k_1} \mathbb{P}(\xi_1^{(i)} = k_1) x_1^{k_1} \sum_{k_2} \mathbb{P}(\xi_2^{(i)} = k_2) x_2^{k_2} \dots \\ &= e^{\mu_{i1}(x_1-1)} e^{\mu_{i2}(x_2-1)} \dots = \exp\left(\textstyle\sum_j \mu_{ij}(x_j - 1)\right). \qquad \text{(Poisson)} \end{aligned}$$

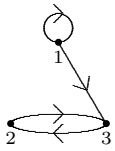
So in vector notation, $\mathbf{f}(\mathbf{x}) = e^{M(\mathbf{x}-\mathbf{1})}$, where $\mathbf{1} = (1, 1, \dots, 1)$ and both $\mathbf{x}$ and $\mathbf{1}$ are regarded as column vectors.

We wish to generalise the results of the previous sections, determining whether or not the process becomes extinct in terms of the matrix M.

To simplify the results, lets make a few assumptions on the process. Construct a directed graph G on vertices $\{1, \dots, N\}$ with an edge from i to j whenever $\mu_{ij} > 0$, i.e., whenever it is possible for a type i node to have a type j child.

Example: Denoting positive entries by $+$,

$$M = \begin{pmatrix} + & 0 & + \\ 0 & 0 & + \\ 0 & + & 0 \end{pmatrix} \qquad \Rightarrow \qquad G = \quad$$



Suppose at time 0 we start with a single node of type 2. Then we can never get a node of type 1 as a descendent. Type 1 then becomes redundant, and can effectively be ignored.

To avoid problems such as this, we shall assume that you can get from any type j to any type i in some number of steps. In other words we shall assume that G is *strongly connected*. The ij entry in M^2 is $\sum_k \mu_{ik}\mu_{kj}$, so is positive precisely when there is a trail $i \to k \to j$ of length 2 from i to j in G. Similarly, the ij entry in M^n is non-zero precisely when there is a trail of length n from i to j. Thus any type node can have any type of descendent provided that for all i and j, there exists an n such that $(M^n)_{ij} > 0$. We call such an M *irreducible*. Note that if M is *not* irreducible then it is possible to divide the types into two classes A and B such that it is impossible to

ever go from type A to type B. By permuting the types, we can then put M into the following form

$$M = \left(\begin{array}{c|c} A & B \\ \hline 0 & C \end{array}\right) \quad \text{with } A \text{ and } C \text{ square matrices.}$$

An irreducible matrix M is therefore any matrix that cannot be written in this form.

However, even ignoring type 1 in the above example, if you start with a node of type 2, then in each generation you alternate between type 2 and type 3. In particular, if the process survives you will have no type 2 nodes in every odd generation, but many in every even generation. To avoid this situation we shall insist that you can get from any type i to any type j in n steps for *all* sufficiently large n. Equivalently, there is some fixed $n > 0$ such that the matrix M^n has *all* entries strictly positive. If this condition holds then we shall call the process $(\mathbf{Z}_n^{(i)})_{n=0}^{\infty}$ *positive regular*.

Another pathological case is when every node has exactly one child. This corresponds to the $p_1 = 1$ case for the single type Galton-Watson process and occurs if and only if $\mathbf{f}(\mathbf{x}) = M\mathbf{x}$. In this case we say that the process is *singular*.

The following result from linear algebra will be useful.

Theorem 11.1: (Perron-Frobenius) *Suppose that M is a matrix with non-negative entries, and for some $n > 0$, M^n has all entries strictly positive. Then there exists a unique largest eigenvalue $\lambda > 0$, a left (row) eigenvector $\mathbf{u} = (u_1, \ldots, u_N)$, $\mathbf{u}M = \lambda\mathbf{u}$, and a right (column) eigenvector $\mathbf{v} = (v_1, \ldots, v_N)$, $M\mathbf{v} = \lambda\mathbf{v}$, with all entries in $\mathbf{u}$ and $\mathbf{v}$ strictly positive. Moreover, if we normalise the eigenvectors so that $\mathbf{u}.\mathbf{v} = 1$, then $(M^n)_{ij} = \lambda^n v_i u_j + O(\alpha^n)$ for some $\alpha < \lambda$.*

From this result we see that if $\mathbf{x}$ is any non-zero vector with non-negative entries, then $\mathbf{x}M^n \sim c\lambda^n\mathbf{u}$ as $n \to \infty$, where $c = \mathbf{x}.\mathbf{v} > 0$.

The main results for single-type processes have very natural generalisations to multi-type processes.

Lemma 11.1: $\mathbb{E}(\mathbf{Z}_n^{(i)}) = \mathbf{Z}_0^{(i)} M^n$. *More generally,* $\mathbb{E}(\mathbf{Z}_n^{(i)} \mid \mathbf{Z}_{n-1}^{(i)} = \mathbf{k}) = \mathbf{k}M$.

Hence for positive regular processes, $\mathbb{E}(\mathbf{Z}_n^{(i)}) \sim c\lambda^n\mathbf{u}$ as $n \to \infty$, where $c = v_i > 0$. We also note that if the process is singular then we must have $\lambda = 1$ since the entries in $\mathbb{E}(\mathbf{Z}_n^{(i)})$ must sum to 1 for all n.

Lemma 11.2: $\frac{\partial f_i}{\partial x_j}(\mathbf{1}) = \mu_{ij}$.

Lemma 11.3: $\mathbf{f}_0(\mathbf{x}) = \mathbf{x}$ *and* $\mathbf{f}_{n+1}(\mathbf{x}) = \mathbf{f}(\mathbf{f}_n(\mathbf{x}))$.

Lemma 11.4: *If* $\mathbf{Z}_n^{(i)}$ *is positive regular and non-singular then almost surely either eventually* $\mathbf{Z}_n^{(i)} = \mathbf{0}$*, or, for each* j*,* $Z_{n,j}^{(i)} \to \infty$*.*

Note that for Lemma 11.4 we need *both* positive regularity *and* non-singularity. In this case there is once again a sharp distinction between survival and extinction of the process.

For any two vectors $\mathbf{x} = (x_1, \dots, x_N)$ and $\mathbf{y} = (y_1, \dots, y_N)$, write $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ for all i. Define the vector $\mathbf{p}_e = (p_{e,1}, \dots, p_{e,N})$ to be the extinction probabilities starting with one node of each type: $p_{e,i} = \mathbb{P}(\mathbf{Z}_n^{(i)} \to \mathbf{0})$.

Theorem 11.2: *The vector* $\mathbf{p}_e$ *is the smallest solution of* $\mathbf{f}(\mathbf{p}_e) = \mathbf{p}_e$*.*

Proof: As in Theorem 2.1, $\mathbf{f}_n(\mathbf{0}) = (\mathbb{P}(\mathbf{Z}_n^{(1)} = \mathbf{0}), \mathbb{P}(\mathbf{Z}_n^{(2)} = \mathbf{0}), \dots)$ increases to a limit $\mathbf{p}_e$, which must then satisfy $\mathbf{f}(\mathbf{p}_e) = \mathbf{p}_e$. Moreover, for any solution $\mathbf{p}_r$ of this equation, $\mathbf{f}_n(\mathbf{0}) \leq \mathbf{p}_r$, so $\mathbf{p}_e \leq \mathbf{p}_r$. In particular, the "smallest" solution is well-defined. (In fact, $\mathbf{1}$ is always one solution, and there is at most one other, which would then be the smallest if it exists.)□

Example: For the $Po(\mu_{ij})$ process described above, $\mathbf{p}_e$ is the smallest solution to $\mathbf{p}_e = e^{M(\mathbf{p}_e - \mathbf{1})}$, or in terms of survival probabilities, $\mathbf{p}_s = \mathbf{1} - \mathbf{p}_e$, $\mathbf{p}_s = \mathbf{1} - e^{-M\mathbf{p}_s}$.

Recall that $\lambda > 0$ is the largest eigenvalue of M.

Theorem 11.3: *Suppose the process* $\mathbf{Z}_n^{(i)}$ *is positive regular and non-singular. If* $\lambda \leq 1$ *then* $\mathbf{Z}_n^{(i)}$ *becomes extinct almost surely, while if* $\lambda > 1$ *then* $\mathbf{Z}_n^{(i)}$ *survives with positive probability.*

Limit Theorems for Multi-type processes

We give (without proofs) several more precise results about the limit of the distribution of $\mathbf{Z}_n^{(i)}$. These are mostly analogous to the single type limit theorems already described.

Subcritical case: $\lambda < 1$*.*

Theorem 11.4: *Suppose* $\mathbf{Z}_n^{(i)}$ *is positive regular and* $\lambda < 1$*. Then* $\mathbb{P}(\mathbf{Z}_n^{(i)} = \mathbf{k} \mid \mathbf{Z}_n^{(i)} \neq \mathbf{0})$ *converges as* $n \to \infty$ *to a probability distribution* $(\tilde{p}_{\mathbf{k}})$*. Moreover,* $\mathbb{P}(\mathbf{Z}_n^{(i)} \neq \mathbf{0})/\lambda^n$ *tends to a limit as* $n \to \infty$*, which is non-zero if and*

only if the mean of the limiting distribution $(\tilde{p}_{\mathbf{k}})$ *is finite, which in turn occurs if and only if* $\mathbb{E}(\xi_j^{(i)} \log \xi_j^{(i)}) < \infty$ *for all* i *and* j.

Critical Case: $\lambda = 1$.

Theorem 11.5: *Suppose* $\mathbf{Z}_n^{(i)}$ *is positive regular and non-singular,* $\lambda = 1$, *and* $\mathrm{Var}(\xi_j^{(i)}) < \infty$ *for all* i *and* j. *Then conditioned on* $\mathbf{Z}^{(i)} \neq \mathbf{0}$, $\mathbf{Z}_n^{(i)}/n$ *tends in distribution to* $W\mathbf{u}$ *where* W *is an exponential random variable and* $\mathbf{u}$ *is the left eigenvector of* M *defined in Theorem 11.1.*

Supercritical case: $\lambda > 1$.

Theorem 11.6: *Suppose* $\mathbf{Z}_n^{(i)}$ *is positive regular and* $\lambda > 1$. *Then almost surely* $\mathbf{Z}_n^{(i)}/\lambda^n \to W\mathbf{u}$, *where* W *is a random variable and* $\mathbf{u}$ *is the left eigenvector of* M *defined in Theorem 11.1. Moreover, if* $\mathbb{E}(\xi_j^{(i)} \log \xi_j^{(i)}) < \infty$ *for all* i *and* j, *then* $\mathbb{E}(W) = v_i > 0$, $\mathbb{P}(W = 0) = \mathbb{P}(\mathbf{Z}_n^{(i)} \to 0) < 1$ *and, provided the number of child nodes is not almost surely* λ, W *has a continuous and strictly positive density function for all* $W > 0$. *Conversely, if* $\mathbb{E}(\xi_j^{(i)} \log \xi_j^{(i)}) = \infty$ *for some* i *and* j, *then* $W = 0$ *almost surely.*

Acknowledgements

The author wishes to thank Béla Bollobás, Stefanie Gerke, and Amites Sarkar for their help in preparing this article.

References

1. K.B. Athreya and P.E. Ney, Branching Processes, Dover Publications Inc. (1972).
2. H. Kesten and B.P. Stigum, A limit theorem for multidimensional Galton-Watson processes. Annals of Mathematical Statisics 37 (1966), 1211–1223.
3. H. Kesten and B.P. Stigum, Additional limit theorems for indecomposable multidimensional Galton-Watson processes Annals of Mathematical Statisics 37 (1966), 1463–1481.
4. N. Levinson, Limiting theorems for Galton-Watson branching processes. Illinois Journal of Mathematics 3 (1959), 554–565.
5. R. Lyons, Random walks and percolation on trees, Annals of Probability 18 (1990), 931–958.
6. D. Williams, Probability with Martingales, Cambridge University Press (1991).
7. A.M. Yaglom, Certain limit theorems of the theory of branching processes. Dokl. Acad. Nauk SSSR 56 (1947), 795–798.