

1 The epidemic lambda-coalescent model

2 Xavier Didelot<sup>1,2,\*</sup>, Ian Roberts<sup>2</sup>, ...

3 <sup>1</sup> School of Life Sciences, University of Warwick, United Kingdom

4

5 <sup>2</sup> Department of Statistics, University of Warwick, United Kingdom

6

7 \* Corresponding author. Tel: 0044 (0)2476 572827. Email: [xavier.didelot@gmail.com](mailto:xavier.didelot@gmail.com)

8 Running title: Epidemic lambda-coalescent

9 Keywords: lambda-coalescent model; pathogen phylogenetics; multiple mergers; offspring distribution

# 1 Introduction

Superspreading in infectious disease epidemiology (?). For example SARS-CoV-2 superspreading (???). Coalescent model (??). Work by Li and Fraser (??). Lambda-coalescent models (???). Beta-coalescent (?) is a specific type of Lambda-coalescent. Was used in (?) and (?). David's paper on inference of multiple mergers while dating a pathogen phylogeny (?).

## 2 Coalescence probabilities

### 2.1 General case

Discrete time  $t$ . Non-overlapping generations of infected individuals. At time  $t$  there are  $N_t$  infected individuals. Each of them creates a number  $s_{t,i}$  of secondary infections at time  $t + 1$ , following the offspring distribution  $\alpha_t(s)$ . The mean of this distribution is the basic reproduction number  $R_t$  and the variance is  $V_t$ . We have:

$$N_{t+1} = \sum_{i=1}^{N_t} s_{t,i} \quad (1)$$

Let  $p_{k,t}$  be the probability that  $k$  individuals at time  $t + 1$  have the same infector at time  $t$ .

### Inclusive Coalescence Probability

Inclusive coalescence probability  $p_{k,t}(N_t, N_{t+1})$  is the probability that exactly  $k$  randomly-sampled individuals from generation  $t + 1$  find a common ancestor in generation  $t$  (*conditioning on population sizes  $N_t$  and  $N_{t+1}$* ).

Given full information about offspring counts from individuals in generation  $t$   $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,N_t})$ , we have

Don't use all of these: -Ian

$$p_{k,t}(\mathbf{s}_t, N_t) = \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \quad (2)$$

$$= \sum_{i=1}^{N_t} \frac{s_{t,i}!}{(s_{t,i} - k)!} \frac{(N_{t+1} - k)!}{N_{t+1}!} \quad (3)$$

$$= \sum_{i=1}^{N_t} \frac{\Gamma(s_{t,i} + 1) \Gamma(N_{t+1} - k + 1)}{\Gamma(s_{t,i} - k + 1) \Gamma(N_{t+1})} \quad (4)$$

$$= \sum_{i=1}^{N_t} \frac{s_{t,i}(s_{t,i} - 1) \dots (s_{t,i} - k + 1)}{N_{t+1}(N_{t+1} - 1) \dots (N_{t+1} - k + 1)}. \quad (5)$$

29 Full information  $\{s_{t,i}\}$  yields the population size  $N_{t+1}$  but is not feasible to observe in practice. We  
 30 can instead express the inclusive coalescence probability conditioning on the next population size  $N_{t+1}$   
 31 by summing over possible offspring counts  $\mathbf{s}_t = (s_{t,1}, \dots, s_{t,N_t})$  conditional on the total generation size

32 To-do/To-discuss:

- 33 • Almost certainly should be an appendix not a main-body derivation, doesn't fundamentally  
 34 change anything but I think this looks horrific
- 35 • Sum subscripts over  $\mathbf{s}_t \in \mathbb{N}_0^{N_t}$  rather than  $\mathbf{s}_t : \dots$  (c.f. exclusive probability derivation)? We have  
 36 probability 0 if sum does not equal  $N_{t+1}$  anyway
- 37 • Change  $\mathbf{s}_t \mapsto \mathbf{s}$  and  $s_{t,i} \mapsto s_i$  for clarity/brevity? Should maintain  $S_{t,i}$  etc. for consistent notation  
 38 ( $S_{t,i} = s_i$  etc.)
- 39 • Define  $S_t^{-(1)} = (S_{t,2}, \dots, S_{t,N_t})$  notation
- 40 • Where to stop? Leave as conditional expectation of binomial coefficients, or convert to factorials  
 41 etc? Factorials are slightly clearer that this is just a 'falling factorial moment' computed from  
 42 probability generating function

$$p_{k,t}(N_t, N_{t+1}) = \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \mathbb{P} \left[ \mathbf{S}_t = \mathbf{s}_t \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] p_{k,t}(\mathbf{s}_t, N_t) \quad (6)$$

$$= \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \mathbb{P} \left[ \mathbf{S}_t = \mathbf{s}_t \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \quad (7)$$

$$= \sum_{i=1}^{N_t} \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \frac{\binom{s_{t,i}}{\binom{N_{t+1}}{k}}} \mathbb{P} \left[ S_{t,1} = s_{t,1}, \mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \quad (8)$$

$$= \frac{N_t}{\binom{N_{t+1}}{k}} \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \binom{s_{t,1}}{k} \mathbb{P} \left[ S_{t,1} = s_{t,1} \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \\ \times \mathbb{P} \left[ \mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \mid S_{t,1} = s_{t,1}, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \quad (9)$$

$$= \frac{N_t}{\binom{N_{t+1}}{k}} \sum_{s_{t,1}=0}^{N_{t+1}} \binom{s_{t,1}}{k} \mathbb{P} \left[ S_{t,1} = s_{t,1} \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \\ \times \underbrace{\sum_{\mathbf{s}_t^{-(1)} : \sum_{i=2}^{N_t} s_{i,t} = N_{t+1} - s_{1,t}} \mathbb{P} \left[ \mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \mid \sum_{i=2}^{N_t} S_{t,i} = N_{t+1} - s_{1,t} \right]}_{=1} \quad (10)$$

$$= \frac{N_t}{\binom{N_{t+1}}{k}} \mathbb{E} \left[ \binom{S_{t,1}}{k} \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right] \quad (11)$$

$$= N_t \frac{(N_{t+1} - k)!}{N_{t+1}!} \mathbb{E} \left[ \frac{S_{t,1}!}{(S_{t,1} - k)!} \mid \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right]. \quad (12)$$

43 The falling factorial moments  $\mathbb{E}\left[\frac{S_{t,1}!}{(S_{t,1}-k)!} \middle| \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]$  in (12) can be readily obtained by  
 44 differentiating the probability generating function of  $S_{t,1} | (\sum_{i=1}^{N_t} S_{t,i} = N_{t+1})$ .

## 45 Exclusive Coalescence Probability

46 Generally, we observe a sample of individuals from each generation rather than the entire population.  
 47 In this case, we are interested in the exclusive coalescence probability  $p_{nkt}(N_t, N_{t+1})$  that exactly  $k$   
 48 individuals from a sample of  $n$  arose from a common ancestor one generation in the past given knowlege  
 49 of the total population sizes  $N_t$  and  $N_{t+1}$ .

50 Given full information about offspring counts of the parents of sampled individuals at the present,  
 51  $\mathbf{x}_t = (x_{t,1}, \dots, x_{t,N_t})$ , we have

$$p_{nkt}(\mathbf{x}_t, N_t) = \sum_{i=1}^{N_t} \frac{\binom{x_{t,i}}{k}}{\binom{n}{k}} \mathbb{I}\{x_{t,i} = k\} \quad (13)$$

$$= \sum_{i=1}^{N_t} \frac{x_{t,i}!}{(x_{t,i} - k)!} \frac{(n - k)!}{n!} \mathbb{I}\{x_{t,i} = k\}. \quad (14)$$

52 Similarly to the exclusive coalescence probability, we can use this to evaluate the exclusive probability  
 53 given  $N_t$  and  $N_{t+1}$  by summing over possible parent offspring configurations (for  $k \leq n$ ),

$$p_{nkt}(N_t, N_{t+1}) = \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^n X_{t,i} = n\right] p_{n,k,t}(N_t, N_{t+1}) \quad (15)$$

$$= \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^n X_{t,i} = n\right] \sum_{i=1}^{N_t} \frac{\binom{x_{t,i}}{k}}{\binom{n}{k}} \mathbb{I}\{x_{t,i} = k\} \quad (16)$$

$$= \frac{N_t}{\binom{n}{k}} \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \binom{x_{t,1}}{k} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \mathbb{I}\{x_{t,1} = k\} \quad (17)$$

$$= \frac{N_t}{\binom{n}{k}} \sum_{\mathbf{x}_t^{-(1)}: \sum_{i=2}^n x_{t,i} = n-k} \binom{k}{k} \mathbb{P}\left[X_{t,1} = k, \mathbf{X}_t^{-(1)} = \mathbf{x}_t^{-(1)} \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \quad (18)$$

$$= \frac{N_t}{\binom{n}{k}} \mathbb{P}\left[X_{t,1} = k \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \sum_{\mathbf{x}_t^{-(1)}: \sum_{i=2}^n x_{t,i} = n-k} \mathbb{P}\left[\mathbf{X}_t^{-(1)} = \mathbf{x}_t^{-(1)} \middle| \sum_{i=1}^{N_t} X_{t,i} = n, X_{t,1} = k\right] \quad (19)$$

$$= \frac{N_t}{\binom{n}{k}} \mathbb{P}\left[X_{t,1} = k \middle| \sum_{i=1}^{N_t} X_{i,t} = n\right]. \quad (20)$$

## 2.2 Poisson case

Here the offspring distribution is  $\text{Poisson}(R_t)$ . In this case, we have

$$\sum_{i=1}^{N_t} S_{t,i} \sim \text{Poisson}(N_t R_t)$$

and

$$S_{t,1} | \left( \sum_{i=1}^{N_t} S_{t,i} = N_{t+1} \right) \sim \text{Binomial}(N_{t+1}, \frac{1}{N_t}).$$

Analogously to the Wright-Fisher model, individuals select a parent uniformly at random from the previous generation. The inclusive probability of coalescence for two lines is

$$p_{2,t} = \frac{1}{N_t}, \quad (21)$$

and more generally the inclusive probability of coalescence for  $n$  lines is

$$p_{k,t} = \frac{1}{N_t^{k-1}}. \quad (22)$$

The exclusive probability of coalescence for two lines from a sample of  $n$  ( $n \geq 2$ ) is

$$p_{n,2,t} = \frac{(N_t - 1)^{n-2}}{N_t^{n-1}}, \quad (23)$$

and more generally the exclusive probability of coalescence for  $k$  lines from a sample of  $n$  ( $n \geq k$ ) is

$$p_{n,k,t} = \frac{(N_t - 1)^{n-k}}{N_t^{n-1}}. \quad (24)$$

## 2.3 Negative-Binomial case

Here the offspring distribution is Negative-Binomial with mean  $R_t$  and variance  $V_t$ . The parameters of this distribution are  $r = R_t^2/(V_t - R_t)$  and  $p = R_t/V_t$ .

The inclusive probability of coalescence for two lines is:

$$p_{2,t} = \frac{r+1}{N_t r + 1} \quad (25)$$

The inclusive probability of coalescence for  $k$  lines is:

$$p_{k,t} = \prod_{i=1}^{k-1} \frac{r+i}{N_t r + i} \quad (26)$$

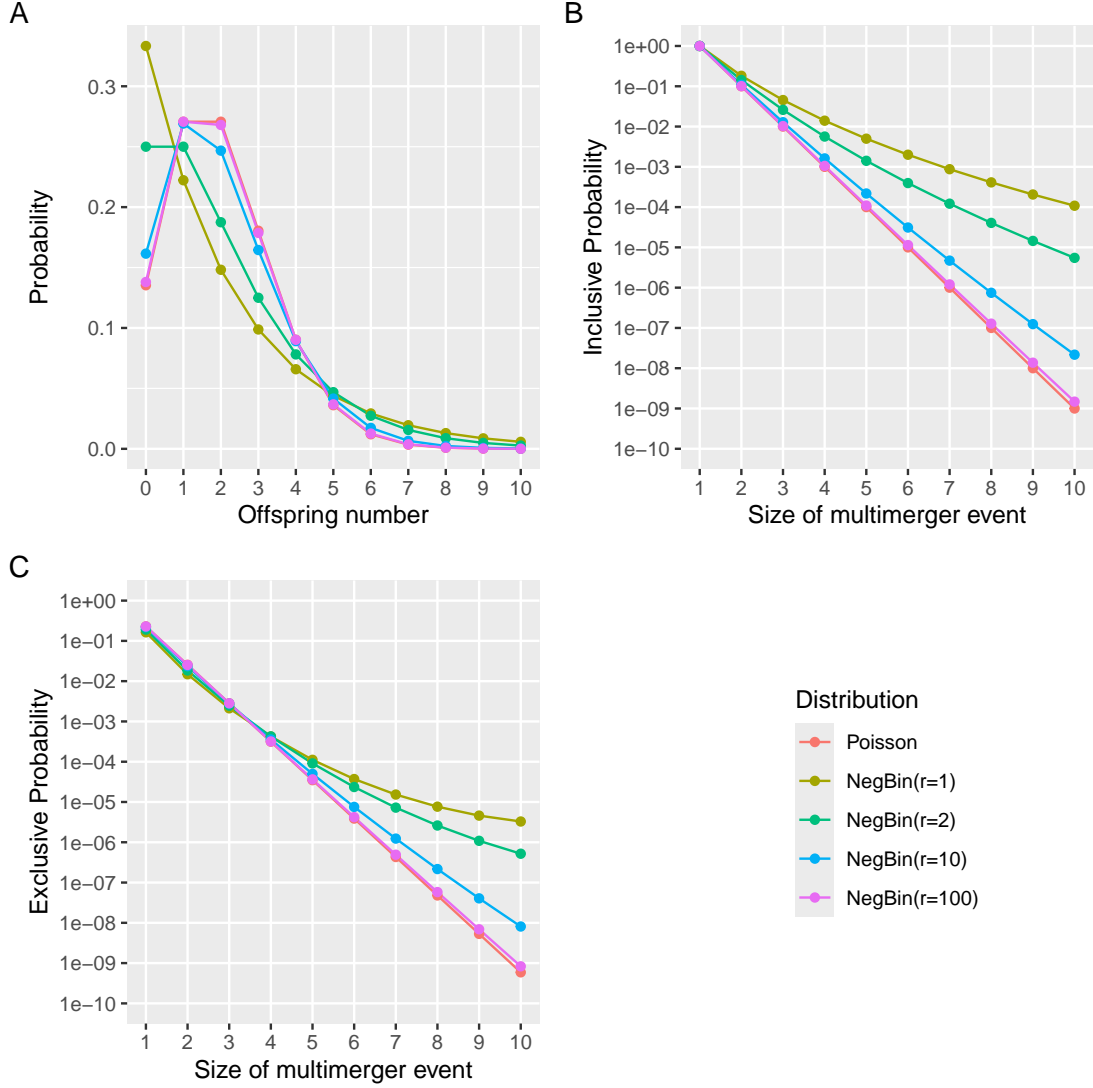


Figure 1: (A) Offspring distribution. (B) Inclusive probability of coalescence. (C) Exclusive probability of coalescence.

## 2.4 Example

Let the offspring distribution have a mean of  $R_t = 2$ . In the Poisson case the offspring distribution is Poisson(2). We consider NegBin cases with the same mean and varying dispersion parameters  $r$ . When  $r$  is high the dispersion is low and the NegBin behaves almost like a Poisson. See Figure 1.

### 3 Lambda-coalescent

A lambda-coalescent model is defined by a probability measure  $\Lambda(dx)$  on the interval  $[0, 1]$ , from which we can deduce the rate  $\lambda_{n,k}$  at which any subset of  $k$  lineages within a set of  $n$  observed lineages coalesce:

$$\lambda_{n,k} = \int_0^1 x^{k-2} (1-x)^{n-k} \Lambda(dx) \quad (27)$$

#### 3.1 Beta-coalescent

Let the Beta function be denoted as  $B(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ . The Beta( $2-\alpha, \alpha$ )-coalescent model (?) has a single parameter  $\alpha \in [0, 2]$  and is defined as:

$$\Lambda(dx) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha, \alpha)} dx \quad (28)$$

from which we can deduce that:

$$\lambda_{n,k} = \frac{B(k-\alpha, n-k+\alpha)}{B(2-\alpha, \alpha)} \quad (29)$$

Special cases include  $\alpha = 2$  corresponding to the Kingman coalescent,  $\alpha = 1$  which is known as the Bolthausen-Sznitman coalescent and  $\alpha = 0$  for which the phylogeny is always star-shaped.

### 4 Implementation

We implemented the analytical methods described in this paper in a new R package entitled *EpiLambda* which is available at <https://github.com/xavierdidelot/EpiLambda> for R version 3.5 or later. All code and data needed to replicate the results are included in the “run” directory of the *EpiLambda* repository.

### 5 Discussion

### Acknowledgements

We acknowledge funding from the National Institute for Health Research (NIHR) Health Protection Research Unit in Genomics and Enabling Data.