1 The epidemic lambda-coalescent model

2 Xavier Didelot[1,2,*], Ian Roberts[2], ...

3 [1] School of Life Sciences, University of Warwick, United Kingdom

5 [2] Department of Statistics, University of Warwick, United Kingdom

7 [*] Corresponding author. Tel: 0044 (0)2476 572827. Email: `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 (Lloyd-Smith et al. 2005). For example SARS-CoV-2 superspreading (Wang et al. 2020; Lemieux et al. 2021; Gómez-Carballa et al. 2021). Coalescent model (Kingman 1982a,b). Work by Li and Fraser (Li et al. 2017; Fraser and Li 2017). Lambda-coalescent models (Pitman 1999; Sagitov 1999; Donnelly and Kurtz 1999). Beta-coalescent (Schweinsberg 2003) is a specific type of Lambda-coalescent. Was used in (Hoscheit and Pybus 2019) and (Menardo et al. 2021). David's paper on inference of multiple mergers while dating a pathogen phylogeny (Helekal et al. 2024).

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:

$$N_{t+1} = \sum_{i=1}^{N_t} s_{t,i} \tag{1}$$

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 $\mathbf{s}_t = (s_{t,1}, \ldots s_{t,N_t})$, we have

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

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

Full information $\{s_{t,i}\}$ yields the population size N_{t+1} but is not feasible to observe in practice. We

can instead express the inclusive coalescence probability conditioning on the next population size N_{t+1} by summing over possible offspring counts $\mathbf{s}_t = (s_{t,1}, \ldots s_{t,N_t})$ conditional on the total generation size

To-do/To-discuss:

- Almost certainly should be an appendix not a main-body derivation, doesn't fundamentally change anything but I think this looks horrific

- Sum subscripts over $\mathbf{s}_t \in \mathbb{N}_0^{N_t}$ rather than $\mathbf{s}_t : \ldots$ (c.f. exclusive probability derivation)? We have probability 0 if sum does not equal N_{t+1} anyway

- 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 ($S_{t,i} = s_i$ etc.)

- Define $S_t^{-(1)} = (S_{t,2}, \ldots, S_{t,N_t})$ notation

- Where to stop? Leave as conditional expectation of binomial coefficients, or convert to factorials etc? Factorials are slightly clearer that this is just a 'falling factorial moment' computed from probability generating function

$$p_{k,t}(N_t, N_{t+1}) = \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \mathbb{P}\left[\mathbf{S}_t = \mathbf{s}_t \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] p_{k,t}(\mathbf{s}_t, N_t) \tag{4}$$

$$= \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \mathbb{P}\left[\mathbf{S}_t = \mathbf{s}_t \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \sum_{i=1}^{N_t} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \tag{5}$$

$$= \sum_{i=1}^{N_t} \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \frac{\binom{s_{t,i}}{k}}{\binom{N_{t+1}}{k}} \mathbb{P}\left[S_{t,1} = s_{t,1}, \mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \tag{6}$$

$$= \frac{N_t}{\binom{N_{t+1}}{k}} \sum_{\mathbf{s}_t : \sum_{i=1}^{N_t} s_{i,t} = N_{t+1}} \binom{s_{t,1}}{k} \mathbb{P}\left[S_{t,1} = s_{t,1} \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]$$

$$\times \mathbb{P}\left[\mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \,\middle|\, S_{t,1} = s_{t,1}, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \tag{7}$$

$$= \frac{N_t}{\binom{N_{t+1}}{k}} \sum_{s_{t,1}=0}^{N_{t+1}} \binom{s_{t,1}}{k} \mathbb{P}\left[S_{t,1} = s_{t,1} \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]$$

$$\times \underbrace{\sum_{\mathbf{s}_t^{-(1)} : \sum_{i=2}^{N_t} s_{i,t} = N_{t+1} - s_{1,t}} \mathbb{P}\left[\mathbf{S}_t^{-(1)} = \mathbf{s}_t^{-(1)} \,\middle|\, \sum_{i=2}^{N_t} S_{t,i} = N_{t+1} - s_{1,t}\right]}_{=1} \tag{8}$$

$$= \frac{N_t}{\binom{N_{t+1}}{k}} \mathbb{E}\left[\binom{S_{t,1}}{k} \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right] \tag{9}$$

$$= N_t \frac{(N_{t+1} - k)!}{N_{t+1}!} \mathbb{E}\left[\frac{S_{t,1}!}{(S_{t,1} - k)!} \,\middle|\, \sum_{i=1}^{N_t} S_{t,i} = N_{t+1}\right]. \tag{10}$$

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 (10) can be readily obtained by
differentiating the probability generating function of $S_{t,1}|(\sum_{i=1}^{N_t}S_{t,i}=N_{t+1})$.

Exclusive Coalescence Probability

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

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

$$p_{nkt}(\mathbf{x}_t, N_t) = \sum_{i=1}^{N_t} \frac{\binom{x_{t,i}}{k}}{\binom{n}{k}} \mathbb{I}\{x_{t,i} = k\} \tag{11}$$

$$= \sum_{i=1}^{N_t} \frac{x_{t,i}!}{(x_{t,i}-k)!} \frac{(n-k)!}{n!} \mathbb{I}\{x_{t,i} = k\}. \tag{12}$$

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

$$p_{nkt}(N_t, N_{t+1}) = \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{n} X_{t,i} = n\right] p_{n,k,t}(\mathbf{x}_t, N_t) \tag{13}$$

$$= \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{n} X_{t,i} = n\right] \sum_{i=1}^{N_t} \frac{\binom{x_{t,i}}{k}}{\binom{n}{k}} \mathbb{I}\{x_{t,i} = k\} \tag{14}$$

$$= \frac{N_t}{\binom{n}{k}} \sum_{\mathbf{x}_t \in \mathbb{N}_0^{N_t}} \binom{x_{t,1}}{k} \mathbb{P}\left[\mathbf{X}_t = \mathbf{x}_t \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \mathbb{I}\{x_{t,1} = k\} \tag{15}$$

$$= \frac{N_t}{\binom{n}{k}} \sum_{\mathbf{x}_t^{-(1)}:\sum_{i=2}^{n} x_{t,i}=n-k} \binom{k}{k} \mathbb{P}\left[X_{t,1} = k, \mathbf{X}_t^{-(1)} = \mathbf{x}_t^{-(1)} \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \tag{16}$$

$$= \frac{N_t}{\binom{n}{k}} \mathbb{P}\left[X_{t,1} = k \middle| \sum_{i=1}^{N_t} X_{t,i} = n\right] \sum_{\mathbf{x}_t^{-(1)}:\sum_{i=2}^{n} x_{t,i}=n-k} \mathbb{P}\left[\mathbf{X}_t^{-(1)} = \mathbf{x}_t^{-(1)} \middle| \sum_{i=1}^{N_t} X_{t,i} = n, X_{t,1} = k\right] \tag{17}$$

$$= \frac{N_t}{\binom{n}{k}} \mathbb{P}\left[X_{t,1} = k \middle| \sum_{i=1}^{N_t} X_{i,t} = n\right]. \tag{18}$$

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

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

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

$$p_{2,t} = \frac{1}{N_t}, \tag{19}$$

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

$$p_{k,t} = \frac{1}{N_t^{k-1}}. \tag{20}$$

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

$$p_{n,2,t} = \frac{(N_t - 1)^{n-2}}{N_t^{n-1}}, \tag{21}$$

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

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

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:

$$p_{2,t} = \frac{r + 1}{N_t r + 1} \tag{23}$$

The inclusive probability of coalescence for k lines is:

$$p_{k,t} = \prod_{i=1}^{k-1} \frac{r + i}{N_t r + i} = \frac{\text{B}(N_t r + 1, r + k)}{\text{B}(r + 1, N_t r + k)} \tag{24}$$

69 The exclusive probability of coalescence for k lines is:

$$p_{nkt} = \frac{N_t \mathrm{B}(k+r, n-k+N_t r - r)}{\mathrm{B}(r, N_t r - r)} \tag{25}$$

70 Somewhere need to note that for any offspring distribution we have:

$$\sum_{k=1}^{n} p_{nkt} \binom{n-1}{k-1} = 1 \tag{26}$$

$$\sum_{k=1}^{n} \binom{n-1}{k-1} p_{nkt} = \sum_{k=1}^{n} \binom{n-1}{k-1} \frac{N_t}{\binom{n}{k}} \mathbb{P}[X_1 = k | \sum_{i=1}^{N_t} X_i = n] \tag{27}$$

$$= \sum_{k=1}^{n} N_t \frac{k}{n} \mathbb{P}[X_1 = k | \sum_{i=1}^{N_t} X_i = n] \tag{28}$$

$$= \frac{N_t}{n} \sum_{k=0}^{n} k \mathbb{P}[X_1 = k | \sum_{i=1}^{N_t} X_i = n] \quad \text{(lower limit } k=0 \text{ makes no difference overall)}$$

$$= \frac{N_t}{n} \mathbb{E}[X_1 | \sum_{i=1}^{N_t} X_i = n] \tag{29}$$

$$= p_{1,t}(N_t, n) = 1 \quad \text{(see below)}$$

71 Taken from stackexchange (link)

$$\mathbb{E}[X_1 | \sum_{i=1}^{N_t} X_i = n] = \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbb{E}[X_i | \sum_{i=1}^{N_t} X_i = n] \quad (X_i \text{ i.i.d.})$$

$$= \frac{1}{N_t} \mathbb{E}[\sum_{i=1}^{N_t} X_i | \sum_{i=1}^{N_t} X_i = n] \quad \text{(Swap expectation, sum \textit{(possibly questionable?)})}$$

$$= \frac{n}{N_t} \tag{30}$$

72 -Ian

2.4 Example

74 Let the offspring distribution have a mean of $R_t = 2$. In the Poisson case the offspring distribution
75 is Poisson(2). We consider NegBin cases with the same mean and varying dispersion parameters r.
76 When r is high the dispersion is low and the NegBin behaves almost like a Poisson. See Figure 1.
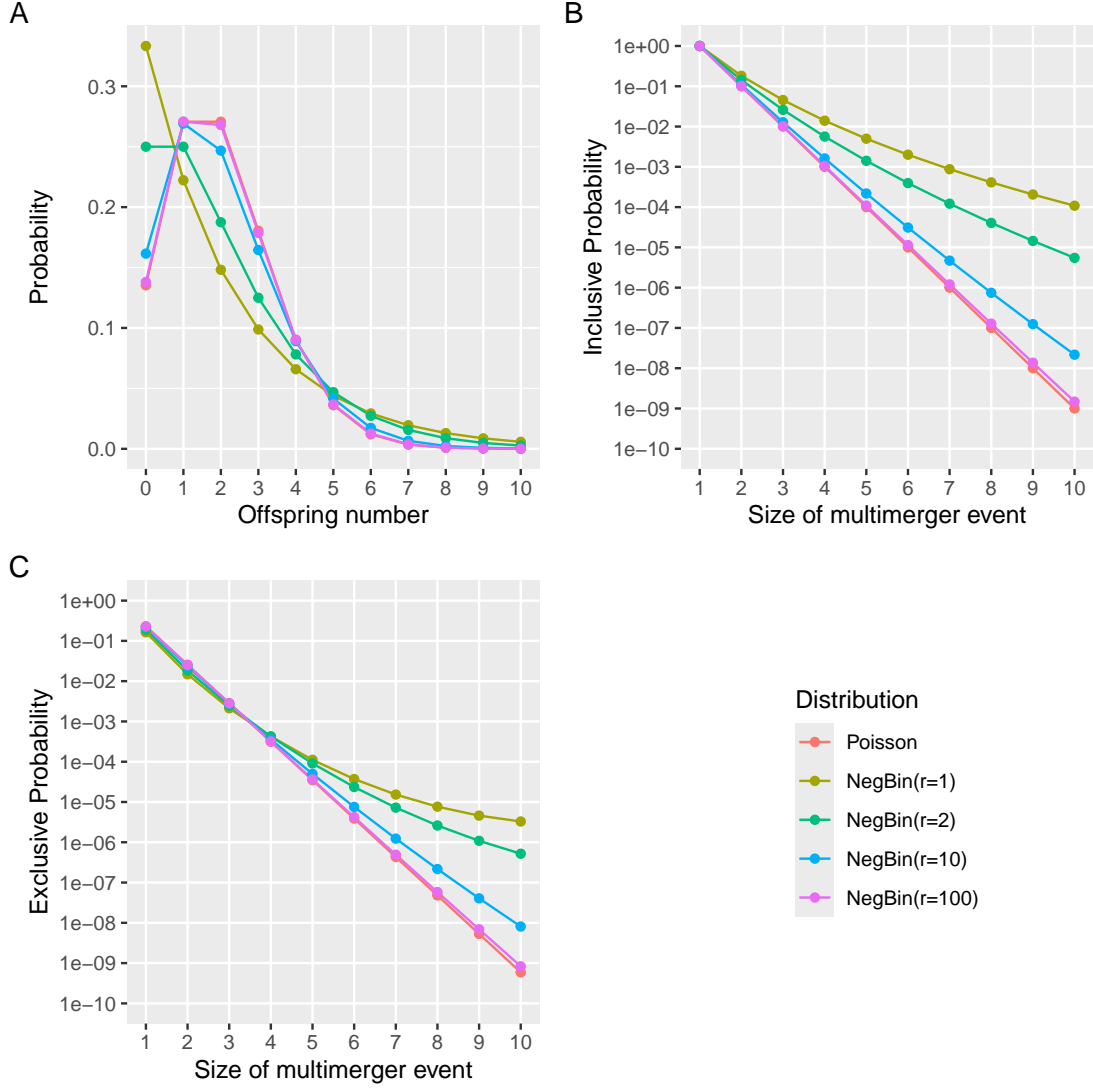
6

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

3 Lambda-coalescent

A lambda-coalescent model is defined by a probability measure $\Lambda(\mathrm{d}x)$ 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:

$$\lambda_{n,k} = \int_0^1 x^{k-2}(1-x)^{n-k}\,\Lambda(\mathrm{d}x) \tag{31}$$

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 (Schweinsberg 2003) has a single parameter $\alpha \in [0, 2]$ and is defined as:

$$\Lambda(\mathrm{d}x) = \frac{x^{1-\alpha}(1-x)^{\alpha-1}}{B(2-\alpha, \alpha)}\mathrm{d}x \tag{32}$$

from which we can deduce that:

$$\lambda_{n,k} = \frac{B(k-\alpha, n-k+\alpha)}{B(2-\alpha, \alpha)} \tag{33}$$

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

References

Donnelly, P., Kurtz, T.G., 1999. Particle Representations for Measure-Valued Population Models. The Annals of Probability 27. doi:10.1214/aop/1022677258.

Fraser, C., Li, L.M., 2017. Coalescent models for populations with time-varying population sizes and arbitrary offspring distributions. bioRxiv , 10.1101/131730doi:10.1101/131730.

Gómez-Carballa, A., Pardo-Seco, J., Bello, X., Martinón-Torres, F., Salas, A., 2021. Superspreading in the emergence of COVID-19 variants. Trends in Genetics 37, 1069–1080. doi:10.1016/j.tig.2021.09.003.

Helekal, D., Koskela, J., Didelot, X., 2024. Inference of multiple mergers while dating a pathogen phylogeny. bioRxiv , 2023.09.12.557403doi:10.1101/2023.09.12.557403.

Hoscheit, P., Pybus, O.G., 2019. The multifurcating skyline plot. Virus Evolution 5, 1–10. doi:10.1093/ve/vez031.

Kingman, J., 1982a. The coalescent. Stochastic Processes and their Applications 13, 235–248. doi:10.1016/0304-4149(82)90011-4.

Kingman, J.F.C., 1982b. On the genealogy of large populations. Journal of Applied Probability 19, 27–43. doi:10.2307/3213548.

Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., DeRuff, K.C., Rudy, M., Bauer, M.R., Lagerborg, K.A., Normandin, E., Chapman, S.B., Reilly, S.K., Anahtar, M.N., Lin, A.E., Carter, A., Myhrvold, C., Kemball, M.E., Chaluvadi, S., Cusick, C., Flowers, K., Neumann, A., Cerrato, F., Farhat, M., Slater, D., Harris, J.B., Branda, J.A., Hooper, D., Gaeta, J.M., Baggett, T.P., O'Connell, J., Gnirke, A., Lieberman, T.D., Philippakis, A., Burns, M., Brown, C.M., Luban, J., Ryan, E.T., Turbett, S.E., LaRocque, R.C., Hanage, W.P., Gallagher, G.R., Madoff, L.C., Smole, S., Pierce, V.M., Rosenberg, E., Sabeti, P.C., Park, D.J., MacInnis, B.L., 2021. Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science 371, eabe3261. doi:10.1126/science.abe3261.

Li, L.M., Grassly, N.C., Fraser, C., 2017. Quantifying Transmission Heterogeneity Using Both Pathogen Phylogenies and Incidence Time Series. Molecular Biology and Evolution 34, 2982–2995. doi:10.1093/molbev/msx195.

Lloyd-Smith, J., Schreiber, S., Kopp, P., Getz, W., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438, 355–9. doi:10.1038/nature04153.

Menardo, F., Gagneux, S., Freund, F., 2021. Multiple Merger Genealogies in Outbreaks of Mycobacterium tuberculosis. Molecular Biology and Evolution 38, 290–306. doi:10.1093/molbev/msaa179.

Pitman, J., 1999. Coalescents with multiple collisions. The Annals of Probability 27, 1870–1902.

Sagitov, S., 1999. The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability 36, 1116–1125. doi:10.1239/jap/1032374759.

Schweinsberg, J., 2003. Coalescent processes obtained from supercritical Galton–Watson processes. Stochastic Processes and their Applications 106, 107–139. doi:10.1016/S0304-4149(03)00028-0.

Wang, L., Didelot, X., Yang, J., Wong, G., Shi, Y., Liu, W., Gao, G.F., Bi, Y., 2020. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nature Communications 11, 5006. doi:10.1038/s41467-020-18836-4.