# STRUCTURED GENEALOGY PROCESSES

QIANYING LIN, AARON A. KING, AND EDWARD L. IONIDES

ABSTRACT. This is a followup paper on Markov Genealogy Process. We construct a continuous time Markov Process, called Structure Genealogy Process, to include structures in the focal population with a moderate adjustments in the setting of individuals, genealogies, and genealogy processes. We derive the exact expression of likelihood of a structured genealogy and develop simulation-based algorithms to conduct statistical inferences. Numerical simulation and real-world examples are included.

## 1. Introduction.

## 2. Mathematical settings.

**Markov genealogy process.** A population process $\mathcal{X}_t \in \mathbb{Z}^d$ as a non-explosive Markov jump process is defined, where $t \in \mathbb{R}_+$ indicates the time and the population is divided into $J \in \mathbb{N}$ classes, with initial-state distribution $p_0(x)$. We then define the jump rate functions $\alpha_u(t, x) \in \mathbb{R}_+$, where $u, x \in \mathbb{Z}^d$ are the jump event and states, respectively. Thus, the population process $\mathcal{X}_t$ is a stochastic process by defining a sequence of jumps, $\omega = (t_k, u_k, n_k)_{k=1}^{K}$, where $K \in \mathbb{N} \cup \{\infty\}$ is the total number of jumps and for the $k$-th jump, $t_k \in \mathbb{R}_+$ is the time, $u_k \in \mathbb{Z}^d$ is the event type, and $n_k \in \mathbb{N}$ is the label of the individual who conducts $u_k$. The full construction can be referred to King et al. (2021). We define the following functions on $\omega \in \Omega$:

$$T_k(\omega) := t_k, \qquad U_k(\omega) := u_k, \qquad N_k(\omega) := n_k, \qquad K(\omega) := K \tag{1}$$

and

$$T(\omega) := (T_k(\omega))_{k=0}^{K(\omega)}, \qquad U(\omega) := (U_k(\omega))_{k=0}^{K(\omega)}, \qquad N(\omega, j) := (N_k(\omega))_{k=0}^{K(\omega)}. \tag{2}$$

**Births.** Within a structured infective population, the parent and the child, at a given birth event, can be in different classes. That is, an individual in class $i$ can give birth to a newborn individual in Class $j$, where $i, j = 1, \dots, J$. We then define the birth function $B_{ij} : \mathbb{Z}^d \to \mathbb{N}$ for any $i, j = 1, \dots, J$. The set of birth events where an individual in Class $i$ gives birth to $n$ individuals in Class $j$ is $\mathbf{B}_{ij} := B_{ij}^{-1}(\{1\})$. Trivially, the set of birth events in Class $j$ is $\mathbf{B}_j := \bigcup_i \mathbf{B}_{ij}$.

**Transitions.** Infective classes are interchangeable and we define the transition function from Class $i$ to $j$ for any $i \neq j$ and $i, j = 1, \dots, J$ as $M_{ij} : \mathbb{Z}^d \to \mathbb{N}$ to represent the number of infectives transiting to Class $j$ from Class $i$. Thus the set or transition is $\mathbf{M}_{ij} := M_{ij}^{-1}(\{1\})$ since we assume single transition at a single transition event. Similarly, we disallow conninciding transition events and other events.

*Date*: January 5, 2022.

$$\begin{tikzcd}
\mathcal{X}_t & & & \mathcal{A}_t \arrow[lll] \arrow[rrr] \arrow[dll] \arrow[ddl] \arrow[dd] \arrow[ddr] \arrow[drr] & & & \mathcal{W}_i \\
& \mathcal{H}_t \arrow[ul] \arrow[urrrrr, dashed] \arrow[rrrr, dashed] \arrow[dr, dashed] \arrow[drr, dashed] & & & & \mathcal{V}_t \arrow[ur] & \\
& & \mathcal{I}_t & \mathcal{G}_t \arrow[l, "\text{ext}"'] \arrow[r, "\text{prune}"'] & \mathcal{U}_t \arrow[ur, "\text{obs}"'] & &
\end{tikzcd}$$

FIGURE 1. Relations among the various Markov processes discussed in the paper. Deterministic maps are indicated with solid arrows; random maps are shown as dashed arrows. All the maps shown commute. $\mathcal{X}_t$ is the *population process*, a model of the dynamics of some system, which we take as a starting point. $\mathcal{H}_t$ is the *history process*, which records the full history of $\mathcal{X}_t$. $\mathcal{I}_t$ is the *inventory process*: at each time $t$, $\mathcal{I}_t$ is an inventory of all extant individuals in the population, each of which has a globally unique name. $\mathcal{G}_t$ is the *genealogy process*, which captures the precise genealogical relationships among all individuals in $\mathcal{I}_t$, as well as among any samples that have been taken from the population. $\mathcal{V}_t$ is the *visible genealogy process*, which is $\mathcal{G}_t$ pruned so that only relationships among samples remain. Finally $\mathcal{W}_i$ is the *embedded chain of the visible genealogy process*, which is $\mathcal{V}_{s_i}$, $s_i$ being the time of the $i$-th sample. All of these processes can be obtained via deterministic procedures applied to the *master process* $\mathcal{A}_t$, as described in the text.

**Samples.** We assume samples are taken serially and no births or deaths are coincided. We then define the sampling function on the probability space of the population process $\mathcal{X}_t$: $G : \mathbb{Z}^d \to \mathbb{N}_0$ is the number of samples taken at event $u$, then the set of sampling events is $\mathbf{G} := G^{-1}(\{1\})$. Of course, the non-coincidence assumption can be easily relaxed while it's not the focus in this paper.

**Deaths and population size.** The population, the set of infected individuals in the context of epidemiology, is structured, categorized into interchangeable classes. We can then define population size function and death functions in Class $j$, respectively, as $I_j, D_j : \mathbb{Z}^d \to \mathbb{N}$ for any $j = 1, \cdots, J$.

$$\alpha_u(t,x) > 0 \Longrightarrow I_j(x+u) - I_j(x) = \sum_i B_{ij}(u) + \sum_i M_{ij}(u) - D_j(u) - \sum_i M_{ji}(u), \tag{3}$$

for all $x, u \in \mathbb{Z}^d$. We disallow simultaneous death at a single death event at this stage and define the set of death events in Class $j$ as $\mathbf{D}_j := D^{-1}(\{1\})$. We also insist, birth and death events don't coincide for any classes, $i, j$, that is, $\mathbf{B}_j \cap \mathbf{D}_i = \emptyset$. Eventually, we can have the function for total population size by letting $I(x) := \sum_{j=1}^{J} I_j(x)$ and it's trivial that

$$\alpha_u(t,x) > 0 \Longrightarrow I(x+u) - I(x) = \sum_{i,j} B_{i,j}(u) - \sum_j D_j(u), \tag{4}$$

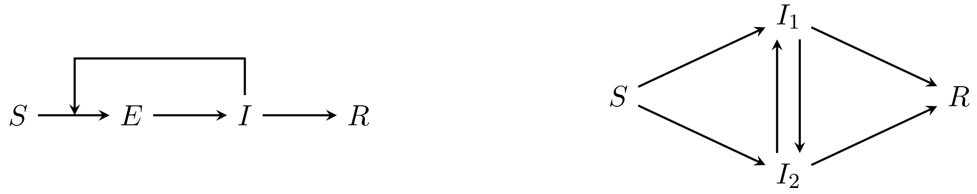

FIGURE 2.

## 3. Examples.

Some commonly used compartmental models in epidemiology are in fact structured, either explicitly or implicitly. Here, we demonstrate below how these representatives fit within the population processes we described previously.

**SEIR model.** The SEIR model is a simple extension of the most basic SIR model, by adding the latent infected compartment "E" into the system, where susceptible individual first become "exposed" when being infected, then turn into being infectious and eventually recover or get removed from the system. The infected population, which we are mostly interested in, is divided into two classes: (1) "E" represents those that are not able to spread the disease though infected and (2) "I" are those that are infectious. To fit within the aforementioned population process, we take $d = 4$ so that the state vector is $\mathcal{X} = (s, e, i, g)$, where $s, e, i, g$ are the number of susceptibles, exposeds, infectives, and the cumulative number of genomic samples collected, respectively. We then summarize four types of jumps, with rate functions:

$$\alpha_u = \begin{cases} b(t)\, s\, i, & u = (-1, 1, 0, 0)\,,\ s > 0,\ i > 0 \\ \sigma\, e, & u = (0, -1, 1, 0)\,,\ e > 0, \\ \gamma\, i, & u = (0, 0, -1, 0)\,,\ i > 0 \\ \psi(t, s, e, i, g), & u = (0, 0, 0, 1)\,,\ e + i > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The above shows the rates of being infected, progression from exposed to infectious, recovery, and sampling, respectively. As in King et al. (2021), the transmission $b(t)$ is time-dependent, and the sampling rate $\psi$ is any function, as long as the constraints in Section 2 are satisfied. Therefore, $I_1(\mathcal{X}) = e$, $I_2(\mathcal{X}) = i$, $I(\mathcal{X}) = e + i$, $\mathbf{B}_{21} = \{(-1, 1, 0, 0)\}$, $\mathbf{M}_{12} = \{(0, -1, 1, 0)\}$, $\mathbf{D}_2 = \{(0, 0, -1, 0)\}$, $\mathbf{G} = \{(0, 0, 0, 1)\}$, and $\mathbf{B}_{11} = \mathbf{B}_{12} = \mathbf{B}_{22} = \mathbf{M}_{21} = \mathbf{D}_1 = \emptyset$.

**SI$^2$R model.** We can customize the state and event vector to fit a complex system. In SI$^2$R model, we have two different while interchangeable classes of infectious individuals, the *per capita* transmissibility therefore in the first class is $b_1(t)$ whilst in the second class is $b_2(t)$. Transition rates between Class $i$ and $j$ are also defined: Class $i$ infection turns into Class $j$ at rate $r_{ij}$ for $i, j = 1, 2$ and $i \neq j$. Furthermore, we also suppose that, once being infected, a susceptible individual becomes Class 1 infectious with probability $\rho$ while Class 2 infectious with probability $1 - \rho$. To specify each event and the number of individuals involved, we define a relatively complex state vector by setting $d = 10$ and $\mathcal{X} = (s, i_1, i_2, b_{11}, b_{12}, b_{21}, b_{22}, m_{12}, m_{21}, g)$, for $s$, $i_1$, and $i_2$ being the population size of the susceptible, the infectives in Class 1, and those in Class2, respectively, and $b_{11}$, $b_{12}$, $b_{21}$, $b_{22}$, $m_{12}$, and $m_{21}$ being the cumulative number of events of parents in Class 1 giving to children in Class 1 and 2, parents in Class 2 giving birth to children in Class 1 and 2, individuals transiting from Class 1 to Class 2, and vice versa, and sampling, respectively. In this case, we can summarize the event rates as:

$$\alpha_u = \begin{cases} b_1(t)\, s\, i_1\, \rho, & u = (-1,1,0,1,0,0,0,0,0,0)\,, s > 0, i_1 > 0 \\ b_1(t)\, s\, i_1\, (1-\rho), & u = (-1,0,1,0,1,0,0,0,0,0)\,, s > 0, i_1 > 0 \\ b_2(t)\, s\, i_2\, \rho, & u = (-1,1,0,0,0,1,0,0,0,0)\,, s > 0, i_2 > 0 \\ b_2(t)\, s\, i_2\, (1-\rho), & u = (-1,0,1,0,0,0,1,0,0,0)\,, s > 0, i_2 > 0 \\ r_{12}\, i_1, & u = (0,-1,1,0,0,0,0,1,0,0)\,, i_1 > 0 \\ r_{21}\, i_2, & u = (0,1,-1,0,0,0,0,0,1,0)\,, i_2 > 0 \\ \gamma\, i_1, & u = (0,-1,0,0,0,0,0,0,0,0)\,, i_1 > 0 \\ \gamma\, i_2, & u = (0,0,-1,0,0,0,0,0,0,0)\,, i_2 > 0 \\ \psi(t,s,i_1,i_2,g), & u = (0,0,0,0,0,0,0,0,0,1)\,, i_1 + i_2 > 0 \\ 0, & \text{otherwise.} \end{cases}$$

The nine non-zero rates above represent those of infection into Class 1 infected by individuals in Class 1, infection into Class 2 infected by individuals in Class 1, infection in Class 1 infected by individuals in Class 2, infection into Class 2 infected by individuals in Class 1, transition from Class 1 to Class 2, transition from Class 2 to Class 1, recovery from Class 1, recovery from Class 2, and sampling, respectively. We then summarize the population functions and sets as follows:

$$\begin{array}{lll} I_1(\mathcal{X}) = i_1, & I_2(\mathcal{X}) = i_2, & I(\mathcal{X}) = i_1 + i_2, \\ \mathbf{B}_{11} = \{(-1,1,0,1,0,0,0,0,0,0)\}, & \mathbf{B}_{12} = \{(-1,0,1,0,1,0,0,0,0,0)\}, & \\ \mathbf{B}_{21} = \{(-1,1,0,0,0,1,0,0,0,0)\}, & \mathbf{B}_{22} = \{(-1,0,1,0,0,0,1,0,0,0)\}, & \\ \mathbf{M}_{21} = \{(0,-1,1,0,0,0,0,1,0,0)\}, & \mathbf{M}_{12} = \{(0,1,-1,0,0,0,0,0,1,0)\}, & \\ \mathbf{D}_1 = \{(0,-1,0,0,0,0,0,0,0,0)\}, & \mathbf{D}_2 = \{(0,0,-1,0,0,0,0,0,0,0)\}, & \\ \mathbf{G} = \{(0,0,0,0,0,0,0,0,0,1)\}. & & \end{array}$$

**Migration.** One special case of the SI$^2$R model is that with one-time migration or introduction. In other words, an infectious individual from another population join our interested population at a certain point of time $\tau$ and then stay and spread the disease. In this case, if this specific individual is defined as Class 1 while other infectious individuals as Class 2, the transition rate $r_{12}$ is a Dirac delta function $\delta(t)$, where $\delta(t) = \infty$ for $t = \tau$ and $\delta(t) = 0$ for $t \neq \tau$, and $r_{21} = 0$ all the time.

**Superspreading.** The definition of superspreading is ambiguous in epidemiology. Some researchers view it as a certain individual who has extremely strong transmissibility with high reproduction rate and call this individual *superspreader*; others avoid to invidualize this phenomenon and consider it as an event where a cluster of infections appears in a single event within a small amount of time and define it as a *superspreading event*. The first definition can be represented by the SI$^2$R model with two classes of infectious individuals: one with extremely strong transmissibility and the other with weak transmissibility. That is, $b_1(t) \ggg b_2(t)$ if Class 1 is the set of highly infective individuals and Class 2 is the set of ordinary infective. In terms of the second definition, the involvement of polytomies and multifurcating trees adds an extra layer of complexity in the population model, along with the issues of measurement error and model robustness, such that we would like to postpone the development of a new framework to a future study.

## 4. Structure genealogy processes.

First, we define a global set $\mathcal{Q}$, which is a collection of infinite unique names. For any Class $j$ of infective population, we define an inventory $\mathcal{I}_t^j(\omega) = \mathrm{inven}\,(\omega|_t, j)$, a list to document all the extant infected individuals in Class $j$ at time $t$, by their names, where $\omega|_t$ is the sequence of jumps up to time $t$. On top of the global set $\mathcal{Q}$ and the $\mathrm{inven}$ procedure, we can then define six operators:

(1) $\mathrm{select}(\mathcal{Q})$ returns a randomly selected name, call it $n$, from the set $\mathcal{Q}$, and update the global set $\mathcal{Q} = \mathcal{Q} \setminus \{n\}$;
(2) $\mathrm{initialize}(I) \coloneqq \bigcup_{i=1}^{I} \{\mathrm{select}(\mathcal{Q})\}$;
(3) $\mathrm{add}(\mathcal{I}) = \mathcal{I} \cup \{\mathrm{select}(\mathcal{Q})\}$;
(4) $\mathrm{drop}(\mathcal{I}, n) = \mathcal{I} \setminus \{n\}$, while update the global set $\mathcal{Q} = \mathcal{Q} \cup \{n\}$;
(5) $\mathrm{movein}(\mathcal{I}, m) = \mathcal{I} \cup \{m\}$; and
(6) $\mathrm{moveout}(\mathcal{I}, m) = \mathcal{I} \setminus \{m\}$.

Now the deterministic and recursive procedure $\mathrm{inven}$ is described as follows, for $\omega$ is a finite jump sequence:

$$\mathrm{inven}(\omega, j) \coloneqq \begin{cases} \mathrm{initialize}(I_j(\mathcal{X}_0(\omega)), & \text{if } K(\omega) = 0; \\ \mathrm{add}(\mathrm{inven}(\omega\text{-})), & \text{if } K(\omega) > 0 \text{ and } U_K(\omega) \in \mathbf{B}_j; \\ \mathrm{drop}(\mathrm{inven}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_K(\omega) \in \mathbf{D}_j; \\ \mathrm{movein}(\mathrm{inven}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_K(\omega) \in \bigcup_i \mathbf{M}_{ij}; \\ \mathrm{moveout}(\mathrm{inven}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_K(\omega) \in \bigcup_i \mathbf{M}_{ji}; \\ \mathrm{inven}(\omega\text{-}), & \text{otherwise.} \end{cases} \tag{5}$$

Eventually, we can define $\mathrm{inven}(\omega) \coloneqq \{\mathrm{inven}(\omega, j)\}_{j=1}^{J}$ as a list of inventories.

**Structured genealogies.** Types of nodes are defined in a structured genealogy: *Leaves* are the extant members at current time and *Internal Nodes* represent ancestors to a certain subset of extant members, which we further define *Sampling Nodes* and *Transition Nodes*, whose differences will be clear in a moment.

For the sake of distinction and visualization, we define a set of *colored balls*, $(f, n) \in \mathsf{F} \times \mathbb{N}$, where $n$ is the label and $f$ is the color of the ball, and $\mathsf{F} \coloneqq \{\mathrm{green}, \mathrm{black}, \mathrm{blue}, \mathrm{red}, \mathrm{purple}\}$. Therefore, we can define a genealogy *node* as a quadruple $(n, t, w, z)$, where $n \in \mathbb{N}$ is the node's *name*, $t \in \mathbb{R}_+$ is its time, $w$ is the node's *pocket* (*i.e.,* unordered pair of colored balls), and $z \in \mathbb{N}$ is the node's *class*. For a node $\mathsf{p}$, we will use $n(\mathsf{p})$, $t(\mathsf{p})$, $w(\mathsf{p})$, and $z(\mathsf{p})$ to denote the name, time, pocket, and class of $\mathsf{p}$, respectively. We then assign black, green, blue, and purple balls to serve as pointers to the extant population at time $t$, to the internal nodes, to the samples, and to the transitions between classes, respectively.

Eventually, the current time $t$ and the sequence of nodes, $\mathcal{G} = \left(t, (\mathsf{p}_k)_{k=0}^{K-1}\right)$, explicitly define a structured genealogy $\mathcal{G}$, where $t \in \mathbb{R}_+$, $K \in \mathbb{N}$, and node $\mathsf{p}_k$ are subjected to conditions described in King et al. (2021). We then use $t(\mathcal{G})$, $K(\mathcal{G})$, and $\mathsf{p}_k(\mathcal{G})$ to represent the time, the length (*i.e.,* the number of nodes), and the $k$-th node of the genealogy $\mathcal{G}$, respectively.

**Effect of births, death, sampling and transistions.** Births, deaths, samples, and transitions change the topology of a structured genealogy. A death at time $t$ indicates the removal of a leaf in the genealogy and the dismiss of a node. The detailed definition can be found in King et al. (2021) and here we continue to use the $\mathrm{drop}(\mathcal{G})$ to denote the resulting sequence of nodes.

A birth at time $t$ in a structured genealogy is similar to that in a unstructured one (King et al., 2021), while the new member introduced to the queue of sequence is in the same class as the parent. Let $\mathsf{b}$ be the $n$-th black ball, the parent, and there exists a node $\mathsf{p} \in \mathcal{G}$ with $\mathsf{b} \in w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$ where $\mathsf{b}'$ is the other ball held by $\mathsf{p}$. We then can produce a new node $\mathsf{p}' = (c, t, \{\mathsf{g}, \mathsf{b}''\}, z(\mathsf{p}))$, where $c = \mathrm{Select}(\mathcal{Q})$, $\mathsf{g} = (\mathrm{green}, c)$, and $\mathsf{b}'' = (\mathrm{black}, c)$. By swapping balls between $\mathsf{p}$ and $\mathsf{p}'$, it ends up that $w(\mathsf{p}) = \{\mathsf{g}, \mathsf{b}'\}$ and $w(\mathsf{p}') = \{\mathsf{b}, \mathsf{b}''\}$. We denote the resulting sequence of nodes as $\mathrm{add}(\mathsf{P}(\mathcal{G}), n)$.

A sample at time $t$ results in the participation of a new node with a new blue ball. The procedure is similar to the previous sampling in King et al. (2021) and the birth mentioned above: the $n$-th black ball held by $\mathsf{p}$ (*i.e.,* $w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$) is selected, then a new node $\mathsf{p}' = (c, t, \{\mathsf{g}, \mathsf{b}''\}, z(\mathsf{p}))$ is generated, where $c = \mathrm{Select}(\mathcal{Q})$, $\mathsf{g} = (\mathrm{green}, c)$, and $\mathsf{b}'' = (\mathrm{blue}, q)$, at the end $\mathsf{p}$ exchanges $\mathsf{b}$ for $\mathsf{g}$ with $\mathsf{p}'$. Here, $q$ is the name of the new blue ball, which is the ordinal number of the transition, $q = |\{\mathsf{b} \in w(\mathcal{G}) \mid \mathsf{b} \text{ is blue}\}|$. Note that, same as that in a birth, $z(\mathsf{p}') = z(\mathsf{p})$. We also use the same notation $\mathrm{sample}(\mathcal{G}, n)$ as the resulting sequence of nodes.

A transition at time $t$ leads to the inclusion of a new node with a new purple ball. Let $\mathsf{b}$ denotes the $n$-th black ball selected to indicate an extant in the population transits from his/her/their current class to a new one. Assuming that the unique node $\mathsf{p} \in \mathcal{G}$ baring this selected black ball (*i.e.,* $\mathsf{b} \in w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$) is in Class $i$ (*i.e.,* $z(\mathsf{p}) = i$) and it's transiting to Class $j$ where $j \neq i$. Now we initialize a new node by taking $\mathsf{p}' = (c, t, \{\mathsf{g}, \mathsf{b}''\}, j)$, where $c = \mathrm{Select}(\mathcal{Q})$, $\mathsf{g} = (\mathrm{green}, c)$, and $\mathsf{b}'' = (\mathrm{purple}, q)$. Similar to other events, we swap $\mathsf{b}$ for $\mathsf{g}$ between nodes $\mathsf{p}$ and $\mathsf{p}'$ and insert $\mathsf{p}'$ at the last position of the node-sequence. Here, $q$ is the name of the new purple ball, which is the ordinal number of the transition, $q = |\{\mathsf{b} \in w(\mathcal{G}) \mid \mathsf{b} \text{ is purple}\}|$. We call this new node $\mathsf{p}'$ the *Transition Nodes* in the genealogy and denote the resulting sequence of nodes by $\mathrm{transition}(\mathsf{P}(\mathcal{G}), n, j)$.

**Event times and structured genealogy process.** Given a genealogy $\mathcal{G}$, the set of *genealogical event times*, $\mathsf{E}(\mathcal{G})$, is the set of all node times. Several of its subsets are of interest. In particular, we define

$$\begin{aligned}
\mathsf{E}(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}\},\\
\mathsf{A}(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}, w(\mathsf{p}) \text{ contains a green ball}\},\\
\mathsf{C}(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}, w(\mathsf{p}) \text{ contains two green balls}\},\\
\mathsf{L}(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}, w(\mathsf{p}) \text{ contains no green balls}\},\\
\mathsf{S}(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}, w(\mathsf{p}) \text{ contains a blue ball}\},\\
\mathsf{Z}(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}, w(\mathsf{p}) \text{ contains a purple ball}\},\\
\mathsf{D}(\mathcal{G}) &\coloneqq \mathsf{A}(\mathcal{G}) \cap \mathsf{S}(\mathcal{G}).
\end{aligned}$$

With these definitions, $\mathsf{A}(\mathcal{G})$ comprises the internal node times, $\mathsf{C}(\mathcal{G})$ is the set of branch times, $\mathsf{L}(\mathcal{G})$ contains the leaf times, $\mathsf{S}(\mathcal{G})$ holds the sample times, $\mathsf{Z}(\mathcal{G})$ holds the transition times, and $\mathsf{D}(\mathcal{G})$ is the set of *direct-descent times*, i.e., the times of samples that are themselves directly ancestral to other samples.

The structured genealogy evolves over time, so we can proceed to define the *structured genealogy process* $\mathcal{G}_t$ now.

## 5. Likelihood, filtering equation, and algorithms.

## 6. Results.

## 7. Discussion.

## References.

King, A. A., Lin, Q., & Ionides, E. L. (2021) Markov genealogy processes. *Theoretical Population Biology* .