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 $J \in \mathbb{N}$ demes, 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 $\omega \in \Omega$:

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

and

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

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

Migrations. Infective demes are interchangeable and we define the migration function from Deme i to j for any $i \neq j$ and $i, j = 1, \ldots, J$ as $M_{ij} : \mathbb{Z}^d \to \mathbb{N}$ to represent the number of infectives migrating to Deme j from Deme i. Thus the set of migration is $\mathbf{M}_{ij} \coloneqq M_{ij}^{-1}(\{1\})$ since we assume single migration at a single migration event. Similarly, we disallow coninciding migration events and other events.

Date: January 28, 2022.

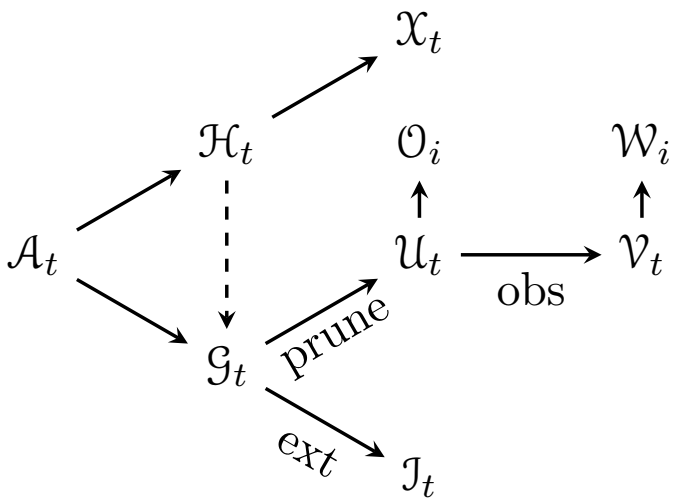

FIGURE 1. Relations among the various Markov processes discussed in the paper. Deterministic maps are indicated with solid arrows; random maps are shown as dashed arrows. All the maps shown commute. $\mathcal{X}_t$ is the *population process*, a model of the dynamics of some system, which we take as 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 $\mathbf{G} \coloneqq G^{-1}(\{1\})$. 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 demes. We can then define population size function and death functions in Deme j, respectively, as $I^j, D_j : \mathbb{Z}^d \to \mathbb{N}$ for any $j = 1, \cdots, J$.

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

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 Deme j as $\mathbf{D}_j \coloneqq D^{-1}(\{1\})$. We also insist, birth and death events don't coincide for any demes, i, j, that is, $\mathbf{B}_j \cap \mathbf{D}_i = \emptyset$. Eventually, we can have the function for total population size by letting $I(x) \coloneqq \sum_{j=1}^{J} I^j(x)$ and it's trivial that

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

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 demes: (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 ψ 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 demes of infectious individuals, the *per capita* transmissibility therefore in the first deme is $b_1(t)$ whilst in the second deme is $b_2(t)$. Migration rates between Deme i and j are also defined: Deme i infection turns into Deme 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 Deme 1 infectious with probability ρ while Deme 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 Deme 1, and those in Deme 2, respectively, and b_{11}, b_{12}, b_{21}, b_{22}, m_{12}, and m_{21} being the cumulative number of events of parents in Deme 1 giving to children in Deme 1 and 2, parents in Deme 2 giving birth to children in Deme 1 and 2, individuals migrating from Deme 1 to Deme 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 Deme 1 infected by individuals in Deme 1, infection into Deme 2 infected by individuals in Deme 1, infection in Deme 1 infected by individuals in Deme 2, infection into Deme 2 infected by individuals in Deme 1, migration from Deme 1 to Deme 2, migration from Deme 2 to Deme 1, recovery from Deme 1, recovery from Deme 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 τ and then stay and spread the disease. In this case, if this specific individual is defined as Deme 1 while other infectious individuals as Deme 2, the migration 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 demes of infectious individuals: one with extremely strong transmissibility and the other with weak transmissibility. That is, $b_1(t) \ggg b_2(t)$ if Deme 1 is the set of highly infective individuals and Deme 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.

Inventory process. First, we define a global set $\mathfrak{Q} := \mathbb{N}$, which is a collection of infinite unique names. For any Deme j of infective population, we define an inventory $\mathfrak{I}_t^j(\omega) = \text{inven}\,(\omega|_t, j)$, a list to document all the extant infected individuals in Deme 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 $\mathfrak{Q}$ and the inven procedure, we can then define six operators:

(1) $\text{next}(\mathfrak{Q})$ returns the minimum integer in $\mathfrak{Q}$, call it n, and update the global set $\mathfrak{Q} = \mathfrak{Q} \setminus \{n\}$;
(2) $\text{initialize}(I) := \bigcup_{i=1}^{I}\{\text{next}(\mathfrak{Q})\}$;
(3) $\text{add}(\mathfrak{I}) = \mathfrak{I} \cup \{\text{next}(\mathfrak{Q})\}$;
(4) $\text{drop}(\mathfrak{I}, n) = \mathfrak{I} \setminus \{n\}$; and
(5) $\text{movein}(\mathfrak{I}, n) = \mathfrak{I} \cup \{n\}$.

Now the deterministic and recursive procedure inven is described as follows, for ω is a finite jump sequence:

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

Eventually, we can define $\text{inven}(\omega) := \{\text{inven}(\omega, j)\}_{j=1}^{J}$ as a list of inventories, such that $\mathfrak{I}_t(\omega) := \{\mathfrak{I}_t^j(\omega)\}_{j=1}^{J}$.

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 *Migration 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 $\mathsf{F} := \{\text{green}, \text{black}, \text{blue}, \text{red}, \text{purple}\}$. 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 *deme*. For a node p, we will use $n(\mathsf{p})$, $t(\mathsf{p})$, $w(\mathsf{p})$, and $z(\mathsf{p})$ to denote the name, time, pocket, and deme of 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 migrations between demes, 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 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. We define $\mathsf{P}(\mathcal{G})$ as the node sequence of genealogy $\mathcal{G}$, where $\mathsf{p} \in \mathcal{G}$ when p is one node in $\mathsf{P}(\mathcal{G})$.

Effect of births, death, sampling and migration. Births, deaths, samples, and migrations 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).

A birth at time t in a structured genealogy is similar to that in a unstructured one (King et al., 2021): a black ball is selected as the parent and a new node introduced as the child, who is in the same deme as the parent. Let b be randomly selected black ball, named *n*, which is the parent, and there exists a node $\mathsf{p} \in \mathcal{G}$ holding $\mathsf{b} \in w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$ where b' is the other ball held by p. Here, we assume the parent is from Deme i (*i.e.,* $n \in \mathcal{I}_i$) and the child is born into Deme j. We then can produce a new node $\mathsf{p}' = (c, t, \{\mathsf{g}, \mathsf{b}''\}, i)$, where $c = \text{next}(\mathcal{Q})$, $\mathsf{g} = (\text{green}, c)$, $\mathsf{b}'' = (\text{black}, c)$. By swapping balls between p and 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 $\text{add}(\mathsf{P}(\mathcal{G}), n)$.

A death at time t is quite different from that defined in King et al. (2021), while we continue to use the notation $\text{drop}(\mathsf{P}(\mathcal{G}), n)$ to denote the resulting sequence of nodes for a death, where an individual with name *n* is selected randomly to be removed from the current population. When a death occurs, a black ball b with name *n* is selected arbitrarily, whose holder is node p, *i.e.,* $\mathsf{b} \in w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$. There exists an unique green ball and an unique node p' such that $\mathsf{g} = (\text{green}, n(p))$ and $\mathsf{g} \in w(\mathsf{p}') = \{\mathsf{g}, \mathsf{b}''\}$. The $\text{drop}(\mathsf{P}(\mathcal{G}), n)$ operation then proceeds differently, conditioned on the color and the deme of the other ball, b', held by p: (1) if b' is green, swap g and b' between p and p' then remove p from the sequence of nodes; (2) if b' is black, the above procedure is conducted when b' and p are in the same deme, otherwise, replace b by a new purple ball $\mathsf{c} = \{\text{purple}, q\}$ where $q = |\{\mathsf{b} \in w(\mathcal{G}) \mid \mathsf{b} \text{ is purple}\}|$; (3) if b' is blue, replace b by a new red ball with the name of b'; (4) if b' is purple, change the deme of b to that of p, swap g and b, and delete p. By doing aforementioned steps, we can keep track of the sampled individuals and the migration at a birth.

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: a black ball, named *n*, where $n \in \mathcal{I}_j$, and held by p (*i.e.,* $w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$), is selected at random. Then a new node $\mathsf{p}' = (c, t, \{\mathsf{g}, \mathsf{b}''\}, j)$ is generated, where $c = \text{next}(\mathcal{Q})$, $\mathsf{g} = (\text{green}, c)$, and $\mathsf{b}'' = (\text{blue}, q)$. At the end, p exchanges b for g with p'. Here, q is the name of the new blue ball, which is the ordinal number of the migration, $q = |\{\mathsf{b} \in w(\mathcal{G}) \mid \mathsf{b} \text{ is blue}\}|$. We use the notation $\text{sample}(\mathsf{P}(\mathcal{G}), n)$ as the resulting sequence of nodes.

A migration at time t leads to the inclusion of a new node with a new purple ball. Let b denote the black ball with name *n* and selected to indicate an extant in the population migrates from his/her current deme to a new one. Assuming this selected individual is from Deme i (*i.e.,* $i \in \mathcal{I}_i$) and migrating to Deme j ($j \neq i$), we then initialize a new node by taking $\mathsf{p}' = (c, t, \{\mathsf{g}, \mathsf{b}''\}, i)$, where $c = \text{next}(\mathcal{Q})$, $\mathsf{g} = (\text{green}, c)$, and $\mathsf{b}'' = (\text{purple}, q)$. Note that, we update $\mathcal{I}_i = \mathcal{I}_i \setminus \{n\}$ and $\mathcal{I}_j = \mathcal{I}_j \cup \{n\}$. Similar to other events, we therefore swap b for g between nodes p and p' and insert 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 migration, $q = |\{\mathsf{b} \in w(\mathcal{G}) \mid \mathsf{b} \text{ is purple}\}|$. We call this new node p' the *Migration Nodes* in the genealogy and denote the resulting sequence of nodes by $\text{migrate}(\mathsf{P}(\mathcal{G}), n)$.

Genealogical event times. the definition for sets of event times is problematic.

Given a genealogy $\mathcal{G}$, the set of *genealogical event times*, $\mathsf{E}(\mathcal{G})$, is the set of all node times. Its subsets, including those for different event types and those for different demes, are of interest. In particular, we

define

$$\begin{aligned}
\mathsf{E}^j(\mathcal{G}) &\coloneqq \{t(\mathsf{p}) \mid \mathsf{p} \in \mathcal{G}, z(\mathsf{p}) \text{ is } j\},\\
\mathsf{A}^j(\mathcal{G}) &\coloneqq,\\
\mathsf{C}^{j,j'}(\mathcal{G}) &\coloneqq,\\
\mathsf{L}^j(\mathcal{G}) &\coloneqq,\\
\mathsf{S}^j(\mathcal{G}) &\coloneqq,\\
\mathsf{Z}^{j,j'}(\mathcal{G}) &\coloneqq,\\
\mathsf{D}^j(\mathcal{G}) &\coloneqq \mathsf{A}^j(\mathcal{G}) \cap \mathsf{S}^j(\mathcal{G}).
\end{aligned}$$

With these definitions, $\mathsf{A}^j(\mathcal{G})$ comprises the internal node times of nodes that are derived from nodes in Deme j, $\mathsf{C}^{j,j'}(\mathcal{G})$ is the set of coalescent times of nodes in Deme j and containing a ball in Deme j', $\mathsf{L}^j(\mathcal{G})$ contains the leaf times where the tips are in Deme j, $\mathsf{S}^j(\mathcal{G})$ holds the sample times for samples in Deme j, $\mathsf{Z}^{j,j'}(\mathcal{G})$ holds the deme-transition times derived from nodes in Deme j and containing a ball in Deme j', and $\mathsf{D}^j(\mathcal{G})$ is the set of *direct-descent times* derived from samples in Deme j.

Structured genealogy process. The structured genealogy evolves over time, so we can proceed to define the *structured genealogy process* $\mathcal{G}_t$ now. We define $\mathcal{G}_t(\omega) = (t, \mathrm{geneal}(\omega|_t))$, for the end time $t \in \mathbb{R}_+$ and the sequence of jumps $\omega \in \Omega$. We first initialize the global set of names by letting $\mathcal{Q} = \mathbb{N} \setminus \{1, 2, \ldots, I(\mathcal{X}_0(\omega))\}$, then the operation geneal is defined recursively as follows for $\omega \in \mathring{\Omega}$:

$$\mathrm{geneal}(\omega) \coloneqq \begin{cases}
(k, 0, \{(\text{green}, k), (\text{black}, k)\}, 0)_{k=0}^{I(\mathcal{X}_0(\omega))}, & \text{if } K(\omega) = 0;\\
\mathrm{add}(\mathrm{geneal}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_{K(\omega)} \in \mathbf{B};\\
\mathrm{drop}(\mathrm{geneal}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_{K(\omega)} \in \mathbf{D};\\
\mathrm{sample}(\mathrm{geneal}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_{K(\omega)} \in \mathbf{G};\\
\mathrm{migrate}(\mathrm{geneal}(\omega\text{-}), N_{K(\omega)}(\omega)), & \text{if } K(\omega) > 0 \text{ and } U_{K(\omega)} \in \mathbf{M};\\
\mathrm{geneal}(\omega\text{-}), & \text{otherwise;}
\end{cases} \tag{6}$$

where $\mathbf{B} \coloneqq \bigcup_{i,j} \mathbf{B}_{ij}$, $\mathbf{D} \coloneqq \bigcup_j \mathbf{D}_j$, and $\mathbf{M} \coloneqq \bigcup_{i,j} \mathbf{M}_{ij}$. A deterministic map, ext, can be defined, such that $\mathcal{I}_t(\omega) = \mathrm{ext}(\mathcal{G}_t(\omega))$ for all $\omega \in \Omega$.

Visible genealogy process. The sampled genealogy contains only relations among genomic samples. In other words, information in a full genealogy about unsampled individuals and deme-transitions are unobserved in a sampled genealogy, which we call it a visible genealogy. We divide the procedure of deriving a visible genealogy from a full genealogy into two separate steps: *pruning* and *obscuring*. The former drops all unsampled individuals (*i.e.,* black balls) in a genealogy and remains the relationship between samples and deme-transition records of samples, and the latter further erases all deme-transition records of samples and obtains an obscured genealogy.

Given a genealogy $\mathcal{G}$, we define $\mathcal{I} = \mathrm{ext}(\mathcal{G})$. Let $\mathrm{prune}(\mathcal{G})$ be the pruned genealogy derived by applying iterative drop operation to $\mathcal{G}$ for each name $c \in \mathcal{I}$. Specifically, suppose $\mathcal{I}^j = \{n_0^j, \ldots, n_{m_j-1}^j\}$ is the inventory for Deme j and $\mathcal{I} = \{\mathcal{I}^j\}_{j=1}^J$, where J is the total number of demes. Then we can define the set of all names in $\mathcal{I}$ as $\mathcal{N} \coloneqq \bigcup_{j=1}^J \mathcal{I}^j = \{c_0, \ldots, c_{m-1}\}$, where $m = \sum_j m_j$. Let $\mathsf{P}_0 = \mathsf{P}(\mathcal{G})$ and $\mathsf{P}_k = \mathrm{drop}(\mathsf{P}_{k-1}, c_{k-1})$, for $k = 1, \ldots, m$. Then $\mathrm{prune}(\mathcal{G}) \coloneqq (t(\mathcal{G}), \mathsf{P}_m)$. **??** illustrates the pruning procedure.

For $\omega \in \Omega$, we define the *pruned genealogy process*,

$$\mathcal{U}_t(\omega) \coloneqq \text{prune}(\mathcal{G}_t(\omega)). \tag{7}$$

Note that $\mathcal{U}_t$ so defined is a itself a genealogy. **??**B depicts one example of $\mathcal{U}_t$ both graphically (as a tree) and diagrammatically (as a node-sequence).

Before we proceed to produce the visible genealogy, we first define an operation $\mathrm{reverse}$ on the sequence of nodes of a genealogy as follows: (1) a random purple ball b with name q is selected such that there exists a node p, where $\mathsf{b} \in w(\mathsf{p}) = \{\mathsf{b}, \mathsf{b}'\}$ and b' is the other ball held by p; (2) there also exists a node p' such that $\mathsf{g} = (\mathrm{green}, n(\mathsf{p})) \in w(\mathsf{p}') = \{\mathsf{g}, \mathsf{b}''\}$ where b'' is the other ball held by p'; (3) swapping balls between p and p' results in $w(\mathsf{p}') = \{\mathsf{b}', \mathsf{b}''\}$ and $w(\mathsf{p}) = \{\mathsf{g}, \mathsf{b}\}$; (4) we delete p from the node sequence. The resulting sequence of nodes is denoted as $\mathrm{reverse}(\mathsf{P}(\mathcal{G}), q)$.

Suppose for a pruned genealogy $\mathcal{G}'$, the set of the serial number of the purple balls is $\{0, 1, \ldots, q\}$ and the list of inventories is $\mathcal{I} = \{\mathcal{I}^j\}_{j=1}^J$. $\mathrm{obs}(\mathcal{G}')$ applies the following steps on $\mathcal{G}'$: (a) let $\mathsf{P}_0 = \mathsf{P}(\mathcal{G}')$ and $\mathsf{P}_k = \mathrm{reverse}(\mathsf{P}_{k-1}, k)$, for $k = 1, \ldots, q$; and (b) for P_q, we update $\mathcal{I} = \bigcup_j \mathcal{I}^j$ and for each node $\mathsf{p} \in \mathsf{P}(\mathcal{G})$, $z(\mathsf{p}) = 0$.

For $\omega \in \Omega$, we define the *visible genealogy process*,

$$\mathcal{V}_t(\omega) \coloneqq \text{obs}(\text{prune}(\mathcal{G}_t(\omega))). \tag{8}$$

Note that $\mathcal{V}_t$ so defined is a itself a genealogy. **??**B depicts one example of $\mathcal{V}_t$ both graphically (as a tree) and diagrammatically (as a node-sequence).

Node color. In a visible genealogy, we can define four kinds of nodes based on the colors of balls in the pockets: (a) *green nodes*, which have more than one green balls in their pockets; (b) *blue nodes*, which have one blue; (c) *red nodes*, which have one red; (d) *purple nodes*, which have one purple ball. In a *compact* visible genealogy, green, blue, red, and purple nodes correspond to *coalescent points*, leaves, direct-descent events, and deme-transition events, respectively. Thus, for a visible genealogy $\mathcal{V}$ and any deme j:

Branch linetype.

Deme proportion and lineage count. We can further define the deme proportion function $\mathcal{D}(t, \omega) \coloneqq \{\mathcal{D}^j(t, \omega)\}_{j=1}^J$, where $\mathcal{D}^j(t, \omega) \coloneqq \frac{|\mathcal{I}_t^j(\omega)|}{|\mathcal{I}_t(\omega)|}$.

Given a pruned genealogy, $\mathcal{U}$, at each time $t \in \mathbb{R}_+$, there are a finite number, $\ell^j(t, \mathcal{U})$, of lineages in Deme j present in the genealogy at that time. Evidently, one has

$$\ell^j(t, \mathcal{U}) \coloneqq \tag{9}$$

Embedded chains. Due to discrete sample times, defined as S_i, we now proceed to consider the embedded chain for pruned genealogy process $\mathcal{U}_t$ and then for visible genealogy process $\mathcal{V}_t$. For $\omega \in \Omega$, let $\mathcal{O}_i(\omega) \coloneqq \mathcal{U}_{S_i(\omega)}(\omega)$ be the embedded chain of the pruned genealogy process, and $\mathcal{W}_i(\omega) \coloneqq \mathcal{V}_{S_i(\omega)}(\omega)$ be that of the visible genealogy process. At each sample time S_i, one extra lineage (*i.e.,* the i-th sample) is to be attached to the previous existing genealogy (*i.e.,* $\mathcal{O}_i$ or $\mathcal{W}_i$) at a random attachment time, denoted by A_i. Both $\mathcal{O}_i$ and $\mathcal{W}_i$ are Markovian.

5. Likelihood, filtering equation, and algorithms.

6. Results.

The derivation of the probability density of $\mathcal{W}_i$ given $\mathcal{H}_{S_i}$ can be divided into two steps:

$$P_{\mathcal{W}_i|\mathcal{H}} = \sum_{\mathcal{O}_i} P_{\mathcal{W}_i|\mathcal{O}_i} P_{\mathcal{O}_i|\mathcal{H}} \tag{10}$$

$$= \sum_{\mathcal{O}_i} P_{\mathcal{W}_i|\mathcal{O}_i} \left(P_{\mathcal{O}_1|\mathcal{H}} \prod_{q=2}^{i} P_{\mathcal{O}_q|\mathcal{O}_{q-1}} \right) \tag{11}$$

We first compute $P_{\mathcal{O}_q|\mathcal{O}_{q-1},\mathcal{H}}$ by fixing $A_q = a_q$, $\mathcal{O}_q = o_q$, $S_q = s_q$, and $\mathcal{H}_{s_q} = h = \left(s_q, (t_k, u_k)_{k=0}^K\right)$. For convenience, we write $x_k = \mathcal{X}(t_k)$, and define the sets $\Omega_q \coloneqq \{\omega \in \Omega \mid \mathcal{W}_q(\omega) = w_q, \mathcal{H}_{s_i}(\omega) = h\} \subset \Omega$ and $\Omega_{qk} \coloneqq N_k(\Omega_q) \subset \mathbb{N}$, for $q \leq i$ and $k \leq K$.

$$P_{\mathcal{O}_q|\mathcal{O}_{q-1},\mathcal{H}}(o_q|o_{q-1}, h) = \prod_{k=1}^{K} \sum_{\omega\in\Omega_q} \beta_{u_k, x_k - u_k}(N_k(\omega)) = \prod_{k=1}^{K} \sum_{\omega\in\Omega_q} \left(\sum_{j=1}^{J} \frac{\mathbb{1}_{\mathcal{I}^j_{t_{k-1}}(\omega)}(N_k(\omega))}{I^j(x_k - u_k)} \right),$$

due to that β is uniform within any certain deme.

We define $\phi_{qk} \coloneqq \sum_{\omega\in\Omega_q} \beta_{u_k, x_k-u_k}(N_k(\omega)) = \prod_{k=1}^{K} \sum_{\omega\in\Omega_q} \left(\sum_{j=1}^{J} \frac{\mathbb{1}_{\mathcal{I}^j_{t_{k-1}}(\omega)}(N_k(\omega))}{I^j(x_k - u_k)} \right)$.

Note that $\ell_{1k} = 0$ for all k.

All cases can be categorized into four groups based on the occurrence of population event within time $[a_q, s_q)$: no population event or simultaneous events, a birth from Deme j to j', a sample of Deme j', and a migration from Deme j to j' where $j \neq j'$. Within each group, we investigate $\phi_{qk}^{j''}$ by considering whether the corresponding population event was presented in the pruned genealogy and involving the q-th lineage sample, which was in Deme j''. For $q > 1$, we then have:

(a) If $t_k \notin [a_q, s_q)$, or if $u_k \notin \mathbf{B} \cup \mathbf{G} \cup \mathbf{M}$, so we have $\phi_{qk}^{j''} = 1$.

(b) If $t_k \in \mathsf{C}(\mathcal{O}_{q-1}) \cup \mathsf{D}(\mathcal{O}_{q-1})$, then again, $\phi_{qk}^{j''} = 1$.

(c) If $j'' \notin \{j, j'\}$, then the new lineage sample was not involved in any population events involving Demes j and j' at time t_k and $\phi_{qk}^{j''} = 1$.

(d) If $t_k \in (a_q, s_q) \setminus \mathsf{A}(\mathcal{O}_{q-1})$, $t_k \in \mathsf{Z}^{j,j'}(\mathcal{O}_q)$, and $u_k \in \mathbf{M}^{j,j'}$, then there was a migration event at time t_k and the j-th sample lineage was involved. Whether this migration event was presented as a deme-transition event in $\mathcal{O}_q$ depends on whether $j'' = j'$. The population size of Deme j right before time t_k was I^j_{k-1}, while that of Deme j' right after t_k was $I^{j'}_k$. Lineages of $\mathcal{O}_{q-1}$ were not involved in this migration event, so they were excluded. Thus, at time t_k, one of the $I^j_{k-1} - \ell^j_{q,k-1}$ individuals was chosen to migrate and right after, among $I^{j'}_k - \ell^{j'}_{qk}$ unobserved individuals, the j-th sample was the exact one just migrated. Therefore, $\phi_{qk}^{j''} = \delta_{j',j''} / \left((I^j_{k-1} - \ell^j_{q,k-1})(I^{j'}_k - \ell^{j'}_{qk}) \right)$.

(e) If $t_k \in (a_q, s_q),\ t_k \notin \mathsf{Z}^{j,j'}(\mathcal{O}_q),$ and $u_k \in \mathbf{M}^{j,j'}$, the logic is the same as that in case d, but the migration event was not presented in $\mathcal{O}_q$. One can avoid the exact genealogical deme-transition induced by a migration at time t_k, with $\phi_{qk}^{j''} = 1 - \delta_{j',j''} / \left((I_{k-1}^{j} - \ell_{q,k-1}^{j})(I_k^{j'} - \ell_{qk}^{j'}) \right)$.

(f) If $t_k = a_q \in \mathsf{L}^{j}(\mathcal{O}_{q-1})$, and obviously, $u_k \in \mathbf{G}_j$. A individual in Deme j was sampled at t_k and the q-th lineage sample was exactly the direct-descend of it, *i.e.,* a direct-descend event caused by the q-th lineage sample was presented in $\mathcal{O}_q$ at time t_k.

(g)

(h) If $t_k \in (a_q, s_q) \setminus \mathsf{A}(\mathcal{O}_{q-1})$, $t_k \in \mathsf{Z}^{j,j'}(\mathcal{O}_q)$, and $u_k \in \mathbf{B}^{j,j'}$, then there was a birth event at time t_k and the j-th sample lineage was involved. However, this birth event was observed as a deme-transition event (*i.e.,* $j \neq j'$) in $\mathcal{O}_q$, instead of a coalescent event. This birth event resulted in a newly added node holding two black balls, which were separately from Deme j and Deme j'. Thus, the population size of Deme j and j', right after t_k, were I_k^j and $I_k^{j'}$, respectively. There were in total $(I_k^j - \ell_{qk}^j)(I_k^{j'} - \ell_{qk}^{j'})$ combinations of the black balls, because none of the lineages in $\mathcal{O}_{q-1}$ was involved. Among all these combinations, the j-th sample lineage was exactly the black ball newly added to Deme j' after the birth event. Therefore, $\phi_{qk}^{j''} = \delta_{j',j''} / \left((I_k^j - \ell_{qk}^j)(I_k^{j'} - \ell_{qk}^{j'}) \right)$.

(i)

(j)

To summarize, we have

$$\phi_{qk}^{j''} = \begin{cases} 1, & \text{if } t_k \notin [a_q, s_q) \text{ or } u_k \notin \mathbf{B} \cup \mathbf{G} \cup \mathbf{M}, \\ 1, & \text{if } t_k \in \mathsf{C}(\mathcal{O}_{q-1}) \cup \mathsf{D}(\mathcal{O}_{q-1}), \\ 1, & \text{if } j'' \notin \{j, j'\} \\ \frac{\delta_{j',j''}}{(I_{k-1}^j - \ell_{q,k-1}^j)(I_k^{j'} - \ell_{qk}^{j'})}, & \text{if } t_k \in (a_q, s_q),\ t_k \in \mathsf{Z}^{j,j'}(\mathcal{O}_q), \text{ and } u_k \in \mathbf{M}^{j,j'}, \\ 1 - \frac{\delta_{j',j''}}{(I_{k-1}^j - \ell_{q,k-1}^j)(I_k^{j'} - \ell_{qk}^{j'})}, & \text{if } t_k \in (a_q, s_q),\ t_k \notin \mathsf{Z}^{j,j'}(\mathcal{O}_q), \text{ and } u_k \in \mathbf{M}^{j,j'}, \\ \frac{\delta_{j',j''}}{I_k^{j'} - \ell_{qk}^{j'}}, & \text{if } t_k = a_q \in \mathsf{L}^{j'}(\mathcal{O}_{q-1}), \\ 1 - \frac{\delta_{j',j''}}{I_k^{j'} - \ell_{qk}^{j'}}, & \text{if } t_k \in (a_q, s_q) \setminus \mathsf{D}(\mathcal{O}_{q-1}) \text{ and } u_k \in \mathbf{G}^{j'}, \\ \frac{\delta_{j',j''}}{(I_k^{j} - \ell_{qk}^{j})(I_k^{j'} - \ell_{qk}^{j'})}, & \text{if } t_k \in (a_q, s_q),\ t_k \in \mathsf{Z}^{j,j'}(\mathcal{O}_q), \text{ and } u_k \in \mathbf{B}^{j,j'}, \\ \frac{1+\delta_{j,j'}}{I_k^j (I_k^{j'} - \delta_{j,j'}) - \ell_{qk}^j (\ell_{qk}^{j'} - \delta_{j,j'})}, & \text{if } t_k = a_q \in \mathsf{C}^{j,j'}(\mathcal{O}_q), \\ 1 - \frac{(1+\delta_{j,j'})\ell_{qk}^j}{I_k^j (I_k^{j'} - \delta_{j,j'}) - \ell_{qk}^j (\ell_{qk}^{j'} - \delta_{j,j'})} - \frac{1}{I_k^{j'} - \ell_{qk}^{j'}}, & \text{if } t_k \in (a_q, s_q),\ t_k \notin \mathsf{Z}^{j,j'}(\mathcal{O}_q), \text{ and } u_k \in \mathbf{B}^{j,j'}, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

We then can establish our first main result:

Theorem 1. *With the definitions as above,* $P_{\mathcal{O}_i|\mathcal{H}}(w|h) = \prod_{q=1}^{i} \prod_{k=1}^{K} \phi_{qk}$.

7. Discussion.

References.

King, A. A., Lin, Q., & Ionides, E. L. (2021) Markov genealogy processes. *arXiv* **2105.12730**.