General theory for stochastic admixture graphs and F-statistics

Soraggi, Samuele; Wiuf, Carsten

General theory for stochastic admixture graphs and F-statistics

Samuele Soraggi [1], Carsten Wiuf [23]

July 18, 2018

Abstract: We provide a general mathematical framework based on the theory of graphical models to study admixture graphs. Admixture graphs are used to describe the ancestral relationships between past and present populations, allowing for population merges and migration events, by means of gene flow. We give various mathematical properties of admixture graphs with particular focus on properties of the so-called F-statistics. Also the Wright-Fisher model is studied and a general expression for the loss of heterozygosity is derived.

Keywords: introgression, genetic drift, Wright-Fisher model, heterozygosity, F-statistic, Markov graphical model. **MSC:** 92D15, 92D25.

---

1 Introduction

Inference on human demographic history from a genetic perspective has been a topic of wide interest in population genetics [2, 3, 16, 7, 13, 18, 17]. Methods for the assessment of migration between populations, and the identification of admixture and splitting events have recently been proposed based on the study of gene flow and introgression between populations [16, 13, 14]. The postulated demographic relationships between populations are described by a graph, generally referred to as an admixture graph, where each node represents a population, ancient or extant, and each directed edge represents an ancestral relationship between two populations. Allele frequencies or other genetic quantities characterising the populations are associated to the nodes, thereby creating, what we name, a stochastic admixture graph. The difference in allele frequencies between two nodes quantifies the gene flow between the two populations represented by the nodes.

The goal of this paper is to provide a stringent mathematical definition and treatment of stochastic admixture graphs and their properties. We use

---

[1]Department of Biology, University of Copenhagen
[2]Department of Mathematical Sciences, University of Copenhagen
[3]Corresponding author. Email `wiuf@math.ku.dk`

1

the theory of graphical models [6, 10, 19] to develop a general mathematical framework to describe stochastic admixture graphs and their associated variables (such as gene frequencies). In the literature stochastic admixture graphs and their properties are often studied with a particular population genetic model in mind. We abstract properties that generalise properties of particular models, using the theory of graphical models.

Many popular computational tools, building on admixture graphs, quantify similarities in genetic composition between populations by means of moment statistics, called F-statistics [13, 16, 14]. These tools rely on properties of the F-statistics that are not always motivated, but nonetheless have proven essential to disentangle complicated genetic ancestries of populations [1, 13, 15, 11]. We deduce general properties of the F-statistics, in particular the F_2-statistic, and show that the F-statistics can be decomposed in terms of the admixture paths between populations. Furthermore, we give conditions under which the F_2-statistic is additive (in a sense to be made precise later) and forms a metric on the nodes of the graph. In the final section of the paper, we consider the Wright-Fisher diffusion model and give a general formula for the decline of heterozygosity over time in a stochastic admixture graph.

We envisage that the general theory developed here and the vast amount of exiting literature on model selection and inference on graphical models will be useful to understand inferential properties of stochastic admixture graphs, what can be done and how. All proofs are given in the appendix.

2 Admixture graphs

We consider labeled graphs with *directed* and *undirected* edges, and use the notations $i \leftrightarrow j$ (equivalently $j \leftrightarrow i$) and $i \rightarrow j$ (equivalently $j \leftarrow i$) for an undirected edge between two nodes i and j, and a directed edge from i to j, respectively. An edge $i \rightarrow j$ is said to be ingoing to j and outgoing of i. The *parents* of j is the set $par(j) = \{i \mid i \rightarrow j \text{ is an edge}\}$. The node i is a *child* of j if j is a parent of i.

Definition 1. An *admixture graph* is an edge labeled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$ without directed cycles. The triplet consists respectively of the set of nodes, edges and labels. The set of nodes $\mathcal{V}$ is divided into:

(i) *roots* $\mathcal{R}$, nodes without ingoing edges. All pairs of roots, and only these, are connected by an undirected edge,

(ii) *admixed nodes* $\mathcal{A}$, nodes with ingoing directed edges,

(iii) *leaves* $\mathcal{A}_0 \subseteq \mathcal{A}$, admixed nodes without outgoing directed edges.

The label α_{ij} of an edge $i \to j$ is positive and fulfils

$$\sum_{i \in par(j)} \alpha_{ij} = 1. \tag{1}$$

An edge between two roots $r_1, r_2 \in \mathcal{R}$ has label $\alpha_{r_1 r_2} = 1$.

In the genetic context, an admixed node, say j, represents a population that is a mixture of several populations, such that $100\alpha_{ij}$ percent of the size of population comes from population i, $i \in par(j)$. The roots represents populations that are ancestral to the other populations in the graph, whereas the ancestral relationship between the root populations are assumed unknown and left unspecified.

By definition, an admixture graph is connected. We assume an admixture graph is not trivial, meaning that it does not consist of only roots and undirected edges. For convenience, we define $\alpha_{ji} = \alpha_{ij}$, $\alpha_{ii} = 1$, and $\alpha_e = \alpha_{ij}$, if e is an edge connecting the nodes i, j. See Figure 1 for examples.

In the following $\mathcal{G}$ denotes an admixture graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$.
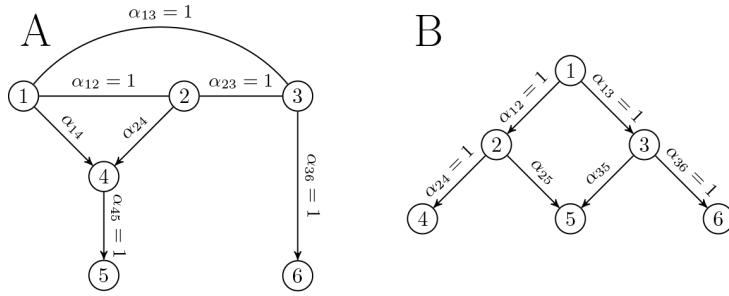

Figure 1: **Examples of admixture graphs.** (A) An admixture graph where nodes $1, 2, 3$ are roots, 4 an admixed node and $5, 6$ a leaves. (B) An admixture graph with three leaves. From 3 to 5 there is as an a-path of type (b), namely $(3, 5)$ with label α_{35}, and one of type (c), namely $(3, 1, 2, 5)$ with label $1 \cdot 1 \cdot \alpha_{36} = \alpha_{36}$.

Definition 2. Given an admixture graph $\mathcal{G}$ and two nodes $i, j \in \mathcal{V}$, an *admixture path* (or simply an *a-path*) γ from i to j is a sequence of edges with distinct nodes,

$$i_k \leftarrow i_{k-1}, \ldots, i_1 \leftarrow i_0, \qquad j_0 \to j_1, \ldots, j_{m-1} \to j_m,$$

where $i_k = i$, $j_m = j$, $k, m \geq 0$, and $i_0 \neq j_0$ only if i_0, j_0 are roots, in which case there is an edge $i_0 \leftrightarrow j_0$. The set of a-paths from i to j is denoted by Γ_{ij}.

The label p_γ of an a-path $\gamma \in \Gamma_{ij}$ is the product of the labels of its edges,

$$p_\gamma := \prod_{e \in \gamma} \alpha_e,$$

where $e \in \gamma$ means e is an edge in the sequence γ.

An a-path contains at most two roots. We might think of three types of a-paths between two nodes: (a) those where all directed edges have direction $\leftarrow$ ($k > 0, m = 0$), (b) those where all directed edges have direction $\rightarrow$ ($k = 0, m > 0$), and (c) those with both directions ($k, m > 0$). If both nodes are roots, then there is only one a-path consisting of an undirected edge ($k = m = 0$). There cannot be both type (a) and type (b) a-paths between two nodes as this would create a directed loop. Retuning to the genetic context, an a-path between two populations implies that one of the populations is ancestral to the other (a,b) or that they share a common ancestor (c). Due to the nature of the graph, there can be two types of a-paths between two populations, as in Figure 1(B).

The set Γ_{ii} contains only the empty sequence with label one. An a-path $\gamma \in \Gamma_{ij}$, $i \neq j$, is not symmetric, meaning that it is not considered the same as the a-path $\gamma' \in \Gamma_{ji}$ composed by the edges of γ in the opposite order. Therefore $\Gamma_{ij} \neq \Gamma_{ji}$. However, the labels of γ and γ' are identical.

For convenience, we often write an a-path as an ordered sequence of nodes $(i_k, \ldots, i_0, j_0, \ldots, j_m)$, leaving out j_0 if $i_0 = j_0$. If $\gamma = (i, \ldots, k, \ldots, j) \in \Gamma_{ij}$, then the subsequences $\gamma_1 = (i, \ldots, k) \in \Gamma_{ik}$ and $\gamma_2 = (k, \ldots, j) \in \Gamma_{kj}$ are a-paths as well.

An a-path γ is not a path in standard graph terminology [8].

Proposition 1. Consider two nodes $i, j \in \mathcal{V}$ of an admixture graph $\mathcal{G}$. Then $\Gamma_{ij} \neq \emptyset$. Further, the sum of the labels over Γ_{ij} is one, $\sum_{\gamma \in \Gamma_{ij}} p_\gamma = 1$.

A tree is a connected subgraph of an admixture graph with only directed edges and at most one incoming edge to each node. In what follows we characterize - in terms of a-paths and labels - when an admixture graph is a tree or a forest. Here a forest is a set of trees with roots connected by undirected edges.

Theorem 1. For an admixture graph $\mathcal{G}$, the following statements are equivalent:

(i) for any pair of nodes, there is only one a-path connecting them,

(ii) every a-path has probability one,

(iii) the admixture graph consists of a forest of R trees, where R is the number of roots.

4

Definition 3. Let $\ell \in \mathcal{A}$ be an admixed node and $r \in \mathcal{R}$ a root of an admixture graph $\mathcal{G}$. Let $\Omega_{\ell r} \subseteq \Gamma_{\ell r}$ be the set of a-paths from ℓ to r that do not contain another root. The *root weight* of r with respect to ℓ is the probability

$$q_{\ell r} = \sum\nolimits_{\gamma \in \Omega_{\ell r}} p_\gamma.$$

Proposition 2. Given an admixture graph, the root weights with respect to an admixed node form a probability distribution.

The probability $q_{\ell r}$ is the proportion of the ancestry of the node ℓ stemming from the root r. If the admixture graph is a forest, then $q_{\ell r}$ is one if the node ℓ is in the tree with root r, and otherwise $q_{\ell r} = 0$.

Definition 4. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{L})$ be an admixture graph and let $C \subseteq \mathcal{V}$. We define the *admixture graph spanned by* C as the graph $\mathcal{G}_C = (\mathcal{V}_C, \mathcal{E}_C, \mathcal{L}_C)$, where

$$\mathcal{V}_C = \big\{ i \mid i \text{ is in an a-path of } \Gamma_{jk} \text{ for some } j, k \in C \big\},$$
$$\mathcal{E}_C = \big\{ e \mid e \in \mathcal{E} \text{ connects two nodes of } \mathcal{V}_C \big\},$$

and $\mathcal{L}_C$ is the set of labels inherited from $\mathcal{G}$. In particular, $\mathcal{G}_\mathcal{V} = \mathcal{G}$.

It is immediate to verify that the graph $\mathcal{G}_C$ is an admixture graph.

Proposition 3. Let $\mathcal{G}$ be an admixture graph, $\mathcal{A}_0 \subseteq \mathcal{V}$ the leaves and $\mathcal{G}_{\mathcal{A}_0}$ the admixture graph spanned by $\mathcal{A}_0$. Furthermore, assume every root has a child. One of the following two equivalent conditions holds

   (i) for each node $k \in \mathcal{V} \backslash \mathcal{A}_0$, there is a pair of nodes $i, j \in \mathcal{A}_0$ and two a-paths $\gamma \in \Gamma_{ik}$, $\delta \in \Gamma_{kj}$, such that γ and δ only have k in common,

   (ii) for each node $k \in \mathcal{V} \backslash \mathcal{A}_0$, there are two nodes $i, j \in \mathcal{A}_0$ and an a-path from i to j that includes k,

if and only if $\mathcal{G} = \mathcal{G}_{\mathcal{A}_0}$. Moreover $\mathcal{A}_0$ is the smallest set spanning the graph $\mathcal{G}$, in the sense that any other set that spans $\mathcal{G}$ contains $\mathcal{A}_0$.

An a-path of $\mathcal{G}$ between two nodes of C is also an a-path of $\mathcal{G}_C$ by definition.

3 Stochastic admixture graphs

We will assume an admixture graph expresses conditional independencies of a random vector. Specifically, an admixture graph is a chain graph (a graph with directed and undirected edges and no directed cycles), which gives rise to a special type of Markov graphical models, called chain graph models [10, 6, 19], see Appendix A. Many models in population genetics and

phylogenetics are assumed to fulfil the conditional independences expressed by a tree, which is a special type of chain graph. Here we make the natural extension to admixture graphs.

Given an admixture graph $\mathcal{G}$, we define the augmented graph $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ by

$$\mathcal{V}^* = \mathcal{V} \cup \big\{(i,j) \,|\, i \to j \in \mathcal{E}\big\},$$

$$\mathcal{E}^* = \{i \leftrightarrow j \,|\, i \leftrightarrow j \in \mathcal{E}\} \cup \{i \to (i,j) \,|\, i \to j \in \mathcal{E}\} \cup \{(i,j) \to j \,|\, i \to j \in \mathcal{E}\},$$

see Figure 2. Note that $\mathcal{G}^*$ is an admixture graph except for the labelling and that an a-path of $\mathcal{G}$ between two nodes $i,j \in \mathcal{V}$ corresponds to a unique a-path of $\mathcal{G}^*$ between the same two nodes $i,j \in \mathcal{V} \subseteq \mathcal{V}^*$, and vice versa.

Here and elsewhere, an equality between two random variables means equality almost surely with respect to the underlying probability measure. If X is a random variable with finite expectation, then $E(X|Y)$ denotes the conditional expectation of X given the random variable Y.

Definition 5. Let $\mathcal{G}$ be an admixture graph and $(V_i \,|\, i \in \mathcal{V})$ a random vector with finite expectation. Then, $(V_i \,|\, i \in \mathcal{V})$ is a *stochastic admixture graph* over $\mathcal{G}$ if there exists a random vector $(C_{ij} \,|\, i \to j \in \mathcal{E})$, defined on the same space as $(V_i \,|\, i \in \mathcal{V})$, with finite expectation and such that

  (i) $(V_j, C_{ij} \,|\, i \in par(j), j \in \mathcal{V})$ is a chain graph model over $\mathcal{G}^*$,

  (ii) $V_j = \sum_{i \in par(j)} \alpha_{ij} C_{ij}$, $j \in \mathcal{A}$,

  (iii) $E(C_{ij}|V_i) = V_i$ for $i \in par(j)$, $j \in \mathcal{A}$.

The variables C_{ij}, $(i,j) \in \mathcal{V}^*$, are called contribution variables.

The requirement that expectations are finite ensures the existence of conditional expectations [9]. In the genetic context, where the variables are allele frequencies, Definition 5(i) is a mathematical formalisation of the statement that evolution and gene flow happens independently along distinct lineages. Definition 5(ii) states that the allele frequency of a population is a weighted sum of the allele frequencies of the admixted populations. These frequencies are themselves the result of neutral evolution in the sense that their expectation is constant over time, Definition 5(iii).

Example 1. Assume $(V_i \,|\, i \in \mathcal{V})$ is a *Gaussian graphical model* over a directed acyclic graph (DAG) $\mathcal{G}$ [19]. The conditional distribution of V_j given the parent variables, $(V_i \,|\, i \in par(j))$, has a Gaussian distribution with form

$$\sum_{i \in par(j)} \alpha_{ij} V_i + \epsilon_j,$$

where $\alpha_{ij} \neq 0$ and the error terms ϵ_j are independent with expectation zero and variance $\sigma_j^2 > 0$. If $\alpha_{ij} > 0$ and $\sum_{i \in par(j)} \alpha_{ij} = 1$, then ϵ_j can be

realised as a weighted sum of independent terms $\epsilon_j = \sum_{i \in par(j)} \alpha_{ij}\epsilon_{ij}$, where ϵ_{ij} has expectation zero and variance σ_j^2/α_{ij}. By defining $C_{ij} = V_i + \epsilon_{ij}$, the three conditions of Definition 5 can be verified. If an arbitrary Gaussian distribution are specified for the variables of the roots of $\mathcal{G}$, then $(V_i \mid i \in \mathcal{V})$ is a stochastic admixture graph over $\mathcal{G}$ (assuming the roots are connected by undirected edges).

A similar remark could be said about *Structural equation models* (SEMs), where the distribution of $(V_i \mid i \in \mathcal{V})$ is allowed to take a general form while preserving (1).

Example 2. The Wright-Fisher model with two alleles, say A and B, assumes a population of constant size N. The frequency of the A allele evolves according to random sampling with replacement (without mutation)

$$P(X_{n+1} = \ell \mid X_n = k) = \binom{N}{\ell} \left(\frac{k}{N}\right)^{\ell} \left(1 - \frac{k}{N}\right)^{N-\ell},$$

where X_n denotes the number of A alleles in generation n [5]. The conditional expectation of X_{n+m} given X_n is $E(X_{n+m} \mid X_n) = X_n$.

Assume evolution occurs on the edges of an admixture graph according to the Wright-Fisher model and that the nodes i and (i, j) in the augmented graph represent a population at different time points. Let V_i and C_{ij} be the relative frequencies of the A allele at these time points, and assume the population at node j is the admixture of its parent populations with respective frequencies α_{ij}, $i \in par(j)$. Then Definition 5(ii)-5(iii) hold.

The same conclusion can be made if mutation between A and B is allowed, and in the diffusion limit as $N \to \infty$. We will return to this model in Section 5.
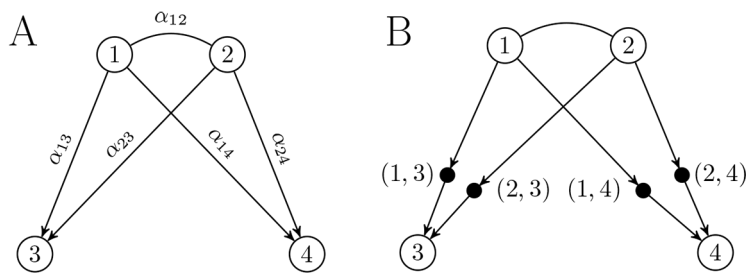

Figure 2: **Contribution nodes.** An example of an admixture graph (A) and its augmented graph (B). Black dots represent nodes associated with contribution variables.

Definition 5(iii) does not in general hold for variables associated to the roots.

Theorem 2. Let $\mathcal{G}$ be a stochastic admixture graph, and assume $V_1, \ldots, V_k$ are the variables associated with the roots. Then $E(V_i \,|\, V_j) = V_j$ holds for any pair of roots if and only if $V_1 = V_2 = \cdots = V_k$.

For reasons of exposition, we assume an order on the set of the roots, $(\mathcal{R}, \prec)$, of an admixture graph. This provides a natural way of writing undirected edges with the smallest node first, $k \leftrightarrow \ell$, if $k \prec \ell$ and $k, \ell \in \mathcal{R}$.

For two nodes $i, j \in \mathcal{V}$, the *drift* from i to j is defined as

$$D_{ij} := V_j - V_i.$$

Note that $D_{ji} = -D_{ij}$. Similarly, we define the *partial drift* of an edge $e = k \to \ell \in \mathcal{E}$ as

$$d_e = d_{k\ell} := C_{k\ell} - V_k,$$

and of an edge $e = k \leftrightarrow \ell \in \mathcal{E}$, $k \prec \ell$, between two roots as

$$d_e = d_{k\ell} := V_\ell - V_k.$$

If $k \leftrightarrow \ell$ is an undirected edge or if k is the only parent of ℓ, then the partial drift $d_{k\ell}$ coincides with the drift $D_{k\ell}$ from k and ℓ.

Let $e = k \to \ell$ or $k \leftrightarrow \ell$, $k \prec \ell$, be an edge in an a-path γ. The sign of e with respect to γ is defined as

$$\text{sign}_\gamma(e) = \begin{cases} \text{sign}_\gamma(e) = +1 & \text{if} \quad \gamma = (\ldots k, \ell \ldots), \\ \text{sign}_\gamma(e) = -1 & \text{if} \quad \gamma = (\ldots \ell, k \ldots), \\ \text{sign}_\gamma(e) = 0 & \text{if} \quad e \notin \gamma. \end{cases}$$

Theorem 3. Let $(V_i \,|\, i \in \mathcal{V})$ be a stochastic admixture graph over $\mathcal{G}$ and let $i, j \in \mathcal{V}$. The drift D_{ij} from i to j decomposes as

$$D_{ij} = \sum_{\gamma \in \Gamma_{ij}} \left(p_\gamma \sum_{e \in \gamma} d_e^\gamma \right), \tag{2}$$

where $d_e^\gamma = \text{sign}_\gamma(e) d_e$ is the signed partial drift with respect to γ.

If $\gamma_1 \in \Gamma_{ij}$ is an a-path from i to j, and $\gamma_2 \in \Gamma_{ji}$ is the 'reversed' a-path from j to i, then $d_e^{\gamma_1} = -d_e^{\gamma_2}$. The signed partial drift of an undirected edge along an a-path is independent of the order defined on $\mathcal{R}$.

Lemma 1. Let $(V_i \,|\, i \in \mathcal{V})$ be a stochastic admixture graph over $\mathcal{G}$ and let $e_1, e_2 \in \mathcal{E}$ be two distinct edges, where at least one is directed. The product of their partial drifts is zero in expectation,

$$E(d_{e_1} d_{e_2}) = 0. \tag{3}$$

Furthermore, for any edge $e = i \to j \in \mathcal{E}$, we have $E(d_e^2) = \text{Var}(C_{ij}) - \text{Var}(V_i)$.

The second part of the lemma implies that $\text{Var}(C_{ij}) \geq \text{Var}(V_i)$, which was first shown for the Wright-Fisher model [14].

Example 3. Continuing with Example 1, for an edge $k \to \ell$, the expectation of the squared partial drift is $E(d_{k\ell}^2) = \text{Var}(V_k + \epsilon_{k\ell}) - \text{Var}(V_k) = \sigma_k^2 / \alpha_{k\ell}$.

4 F-statistics

In this section we discuss the F-statistics F_2, F_3 and F_4 [13, 16, 14], and give various results for these. We assume $(V_i \mid i \in \mathcal{V})$ has finite second moments.

Definition 6. Let $(V_i \mid i \in \mathcal{V})$ be a stochastic admixture graph over $\mathcal{G}$ and $i, j \in \mathcal{V}$. The F_2-*statistic* between i and j is

$$F_2(i, j) = E(D_{ij}^2). \tag{4}$$

The F_2-statistic is non-negative and symmetric by definition. Using Theorem 3 and Lemma 1, we might rewrite (4) in terms of squared partial drifts along the a-paths of Γ_{ij}, see below.

For two edges $e_1, e_2 \in \mathcal{E}$ and four nodes $i, j, k, \ell \in \mathcal{V}$ of an admixture graph, define the B-*coefficient* of (e_1, e_2) with respect to the pairs $(i, j), (k, \ell)$ as

$$B_{e_1,e_2}^{(i,j)(k,\ell)} = \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ij}^{e_1} \times \Gamma_{k\ell}^{e_2}} \text{sign}_{\gamma_1}(e_1)\,\text{sign}_{\gamma_2}(e_2) p_{\gamma_1} p_{\gamma_2}$$

$$= \left(\sum_{\gamma_1 \in \Gamma_{ij}^{e_1}} \text{sign}_{\gamma_1}(e) p_{\gamma_1} \right) \left(\sum_{\gamma_2 \in \Gamma_{k\ell}^{e_2}} \text{sign}_{\gamma_2}(e) p_{\gamma_2} \right). \tag{5}$$

For an edge $e \in \mathcal{E}$, the A-*coefficient* of e with respect to the pairs $(i, j), (k, \ell)$ is defined as

$$A_e^{(i,j)(k,\ell)} = B_{e,e}^{(i,j)(k,\ell)}. \tag{6}$$

For convenience, we write $A_e^{(i,j)}$ and $B_{e_1,e_2}^{(i,j)}$ if $(k, \ell) = (i, j)$. In that case the coefficients are symmetric in i, j.

Lemma 2. Let $\mathcal{G}$ be an admixture graph and $i, j, k, \ell \in \mathcal{V}$. For two edges $e_1, e_2 \in \mathcal{E}$, the following holds:

(i) $-1 \le B_{e_1,e_2}^{(i,j)(k,\ell)} \le 1$,

(ii) $B_{e_1,e_2}^{(i,j)(k,\ell)} = \pm 1$ if and only if $\Gamma_{ij}^{e_1} = \Gamma_{ij}$, $\Gamma_{k\ell}^{e_2} = \Gamma_{k\ell}$, $\text{sign}_{\gamma_1}(e_1)$ is independent of $\gamma_1 \in \Gamma_{ij}$ and $\text{sign}_{\gamma_2}(e_2)$ is independent of $\gamma_2 \in \Gamma_{k\ell}$,

(iii) $B_{e_1,e_2}^{(i,j)(k,\ell)} = 0$ for all positive values of α_e, $e \in \mathcal{E}$, such that (1) is fulfilled, if and only if $\Gamma_{ij}^{e_1} = \emptyset$ or $\Gamma_{k\ell}^{e_2} = \emptyset$.

The lemma implies that $0 \le A_e^{(i,j)} \le 1$ for any edge and $-1 < B_{e_1,e_2}^{(i,j)} \le 1$ for two undirected edges. The latter follows from (ii) and the fact that an a-path cannot pass through two distinct undirected edges.

Let $\mathcal{E}_{ij} \subseteq \mathcal{E}$ be the set of edges that appear in at least one path of Γ_{ij}. Further, let $\mathcal{E}_{ij}^u \subseteq \mathcal{E}_{ij}$ be the set of undirected edges and $\mathcal{E}_{ij}^d \subseteq \mathcal{E}_{ij}$ the set of directed edges.

Theorem 4. Let $(V_i \,|\, i \in \mathcal{V})$ be a stochastic admixture graph over $\mathcal{G}$ and $i, j \in \mathcal{V}$. The F_2-statistic $F_2(i, j)$ decomposes as

$$F_2(i, j) = \sum_{e \in \mathcal{E}_{ij}^d} A_e^{(i,j)} E(d_e^2) + \sum_{(e_1, e_2) \in \mathcal{E}_{ij}^u \times \mathcal{E}_{ij}^u} B_{e_1, e_2}^{(i,j)} E(d_{e_1} d_{e_2}) \qquad (7)$$

If $\mathcal{G}_{\{i,j\}}$ has at most two roots, (7) might be written compactly as

$$F_2(i, j) = \sum_{e \in \mathcal{E}_{ij}} A_e^{(i,j)} E(d_e^2).$$

If $\mathcal{G}$ is a Gaussian admixture graph, then $A_e^{(i,j)} E(d_e^2) = \sigma_k^2 A_e^{(i,j)} / \alpha_{k\ell}$ is a polynomial in the labels, that is, $\alpha_{k\ell}$ cancels out, see Example 3.

We next characterise the additivity of the F_2-statistic [12, 16, 13].

Proposition 4. Consider a stochastic admixture graph $(V_i \mid i \in \mathcal{V})$ over $\mathcal{G}$, and let $i, j, k \in \mathcal{V}$. If all a-paths of Γ_{ij} pass through k, then $F_2(i, j) = F_2(i, k) + F_2(k, j)$.

Assume $\mathcal{G}_{\{i,j\}}$ has at most two roots. If $F_2(i, j) = F_2(i, k) + F_2(k, j)$ for all positive values of $E(d_e^2)$ and α_e, $e \in \mathcal{E}$, such that (1) is fulfilled, then all a-paths of Γ_{ij} pass through k.

If additivity holds for any node on the a-paths between any two nodes, then and only then is the admixture graph a forest (Theorem 1).

Example 4. Consider the admixture graph in Figure 1B. The only a-path of Γ_{57} goes through node 2, so $F_2(5, 7) = F_2(5, 2) + F_2(2, 7)$. For node 6 we have $F_2(5, 7) = F_2(5, 6) + F_2(6, 7) - 2F_3(6; 5, 7)$, but $F_3(6; 5, 7)$ is only zero for certain values of the parameters. In fact,

$$E(D_{65} D_{67}) = \alpha_{36}^2 E(d_{36}^2) + \alpha_{46}^2 E(d_{46}^2) - 2\alpha_{36} \alpha_{46} E(d_{23}^2 + d_{24}^2),$$

which is zero for certain choices of parameters.

The F_3- and F_4-statistics are defined analogously to the F_2-statistic

$$F_3(i; j, k) = E(D_{ij} D_{ik}) \quad \text{and} \quad F_4(i, j; k, l) = E(D_{ij} D_{k\ell}),$$

for four nodes $i, j, k, \ell \in \mathcal{V}$. Similarly to the decomposition of the F_2-statistic, the F_3- and F_4-statistics decompose as

$$F_3(i; j, k) = \sum_{e \in \mathcal{E}_{ij}^d \cap \mathcal{E}_{ik}^d} A_e^{(i,j)(i,k)} E(d_e^2) + \sum_{(e_1, e_2) \in \mathcal{E}_{ij}^u \times \mathcal{E}_{ik}^u} B_{(e_1, e_2)}^{(i,j)(i,k)} E(d_{e_1} d_{e_2}) \qquad (8)$$

$$F_4(i, j; k, \ell) = \sum_{e \in \mathcal{E}_{ij}^d \cap \mathcal{E}_{k\ell}^d} A_e^{(i,j)(k,\ell)} E(d_e^2) + \sum_{(e_1, e_2) \in \mathcal{E}_{ij}^u \times \mathcal{E}_{k\ell}^u} B_{(e_1, e_2)}^{(i,j)(k,\ell)} E(d_{e_1} d_{e_2}). \qquad (9)$$

(the proofs are similar to the proof for the F_2-statistic).

A visual method to decompose the F-statistics is introduced in [13] and [16]. This is formally motivated here as a consequence of the decompositions. The steps to calculate the F_2-statistic between two nodes $i, j \in \mathcal{V}$ are the following, see Example 5:

1. Consider all pairs $(\gamma_1, \gamma_2) \in \Gamma_{ij} \times \Gamma_{ij}$, including coincident pairs,

2. For each pair (γ_1, γ_2), multiply the sum of the squared partial drifts of the directed edges that are in both a-paths by $p_{\gamma_1} p_{\gamma_2}$,

3. For each pair (γ_1, γ_2), sum the products of partial drifts $d_{e_1}^{\gamma_1} d_{e_2}^{\gamma_2}$ over the undirected edges $e_1 \in \gamma_1$ and $e_2 \in \gamma_2$, and multiply by $p_{\gamma_1} p_{\gamma_2}$,

4. Sum all terms above and take the expectation of the sum.

An analogous procedure is followed to calculate the F_3- and F_4-statistics, by considering the pairs of a-paths (γ_1, γ_2) between the pairs of nodes $(i, j), (i, k)$ and $(i, j), (k, \ell)$ for $F_3(i; j, k)$ and $F_4(i, j; k, \ell)$, respectively.

Example 5. Consider $F_2(5, 6)$ in the admixture graph of Figure 1A. There are two a-paths in Γ_{56}, namely $\gamma_1 = (5, 4, 1, 3, 6)$ and $\gamma_2 = (5, 4, 2, 3, 6)$, resulting in four pairs of a-paths, see Figure 3A. Applying the visual method yields:

$$
\begin{aligned}
F_2(5, 6) = E\Big(& p_{\gamma_1}^2 (d_{45}^2 + d_{14}^2 + d_{13}^2 + d_{36}^2) + p_{\gamma_2}^2 (d_{45}^2 + d_{24}^2 + d_{23}^2 + d_{36}^2) \\
& + 2 p_{\gamma_1} p_{\gamma_2} (d_{45}^2 + d_{36}^2 + d_{13} d_{23}) \Big).
\end{aligned}
$$

By collecting terms with the same partial drifts, we obtain

$$
\begin{aligned}
F_2(5, 6) = & E(d_{45}^2) + E(d_{36}^2) + p_{\gamma_1}^2 E(d_{14}^2) + p_{\gamma_1}^2 E(d_{13}^2) + p_{\gamma_2}^2 E(d_{24}^2) \\
& + p_{\gamma_2}^2 E(d_{23}^2) + 2 p_{\gamma_1} p_{\gamma_2} E(d_{13} d_{23}).
\end{aligned}
$$

We recognize the A- and B-coefficients,

$$
\begin{aligned}
& A_{4 \to 5}^{(5,6)} = A_{3 \to 6}^{(5,6)} = 1, \quad A_{1 \to 4}^{(5,6)} = A_{2 \to 4}^{(5,6)} = p_{\gamma_1}^2, \\
& B_{2 \leftrightarrow 3, 2 \leftrightarrow 3}^{(5,6)} = p_{\gamma_2}^2, \qquad B_{1 \leftrightarrow 3, 1 \leftrightarrow 3}^{(5,6)} = p_{\gamma_1}^2, \qquad B_{1 \leftrightarrow 3, 2 \leftrightarrow 3}^{(5,6)} = B_{2 \leftrightarrow 3, 1 \leftrightarrow 3} = p_{\gamma_1} p_{\gamma_2}.
\end{aligned}
$$

The F_3-statistics can be given in terms of the F_2-statistics,

$$
F_2(i, j) = F_2(i, k) + F_2(k, j) - 2 F_3(k; i, j)
$$

[16]. This shows that the F_2-statistic is a metric on $\mathcal{V}$ (that is, it fulfils the triangular inequality) if and only if $F_3(k; i, j) \geq 0$, for all $i, j, k \in \mathcal{V}$.
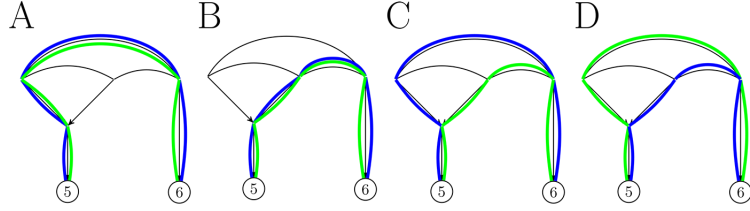
Figure 3: **Visual method for the F-statistic.** Illustration of the visual method to calculate $F_2(5,6)$ in the admixture graphs of Figure 1A. A pair of paths is represented by a blue and a green lines connecting adjacent nodes. (A-B) Pairs of coincident admixture paths sharing all their edges. (C-D) A pair of different paths where the shared directed edges are $4 \to 5$, $3 \to 6$. The product of partial drifts of the two undirected edges on the blue and green paths appears in $F_2(5,6)$ (see step 3 of the visual method).

If there are at most two roots then there is at most one undirected edge and $F_3(i;j,k)$ can be written as a sum involving only A-coefficients, see (6) and (8). Express $F_3(i;j,k)$ as the sum $F_{3+}(i;j,k) + F_{3-}(i;j,k)$, where

$$F_{3+}(i;j,k) = \sum_{e \in \mathcal{E}_{ij} \cap \mathcal{E}_{ik}} A_{e+}^{(i,j)(i,k)} E(d_e^2), \quad F_{3-}(i;j,k) = \sum_{e \in \mathcal{E}_{ij} \cap \mathcal{E}_{ik}} A_{e-}^{(i,j)(i,k)} E(d_e^2),$$

and $A_{e+}^{(i,j)(i,k)}$ and $A_{e-}^{(i,j)(i,k)}$ collect the terms with same and opposite sign of e in $(\gamma_1, \gamma_2) \in \Gamma_{ij}^e \times \Gamma_{ik}^e$, respectively.

The next statement relates the sign of the F_3-statistic to the topology of the admixture graph.

Proposition 5. Let $\mathcal{G}$ be an admixture graph and $C = \{i, j, k\} \subseteq \mathcal{V}$ such that $\mathcal{G}_C$ has at most two roots. Then $F_{3-}(i;j,k) \neq 0$ if and only if there exist $e \in \mathcal{E}_{ij} \cap \mathcal{E}_{ik}$ and a pair of a-paths $(\gamma_1, \gamma_2) \in \Gamma_{ij}^e \times \Gamma_{ik}^e$ such that

$$\gamma_1 = i \leftarrow \cdots \to j \quad \text{or} \quad \gamma_1 = i \leftarrow \ldots \leftrightarrow j,$$
$$\gamma_2 = i \leftarrow \cdots \to k \quad \text{or} \quad \gamma_2 = i \leftarrow \ldots \leftrightarrow k,$$

and $\mathrm{sign}_{\gamma_1}(e) = -\mathrm{sign}_{\gamma_2}(e)$.

5 The Wright-Fisher model

In this section we consider the Wright-Fisher model in the diffusion limit as the population size becomes infinite [5], see also Example 2.

Definition 7. A *Wright-Fisher admixture graph* $(V_i | i \in \mathcal{V})$ over $\mathcal{G}$ is a stochastic admixture graph such that V_i, $i \in V$, are the frequencies of a particular allele in the populations represented by the nodes of the graph,

and such that C_{ij} conditional on $V_i = v \in [0,1]$, $i \in par(j)$, $j \in \mathcal{V}$, is distributed as the frequency in a Wright-Fisher diffusion process at time τ_{ij}, given it starts at frequency v, and τ_{ij} is the population scaled time between the nodes i and (i,j).

[12, chapter 13] showed that the expected heterozygosity of an allele declines exponentially over time,

$$E(V_1(1 - V_1)) = (1 - e^{-\tau_{12}})E(V_2(1 - V_2)),$$

where V_1 is the frequency at time zero and V_2 the frequency τ_{12} population scaled time units later. We generalise this formula to a Wright-Fisher admixture graph.

Proposition 6. Consider a Wright-Fisher admixture graph $(V_i \mid i \in \mathcal{V})$ over $\mathcal{G}$. Then the expected heterozygosities fulfil

$$E(V_j(1 - V_j)) = \sum_{(r_1,r_2) \in \mathcal{R} \times \mathcal{R}} D^j_{r_1,r_2} E(V_{r_1}(1 - V_{r_2})), \quad j \in \mathcal{V},$$

where $D^j_{r_1,r_2}$ are non-negative constants depending on the edge labels and the population scaled times. Furthermore,

$$\sum_{r_1,r_2} D^j_{r_1,r_2} \leq 1, \quad j \in \mathcal{V}.$$

If all population scaled times τ_{ij}, $i \to j \in \mathcal{E}$, between two nodes are set to zero, then the sum of $D^j_{r_1,r_2}$ over r_1, r_2 is precisely one (see the proof of the. proposition). In that case, $E(V_j(1 - V_j))$ is the probability of drawing two different alleles from the root populations. With probability $D^j_{r_1,r_2}$, alleles from population r_1 and r_2 are drawn. In general, when $D^j_{r_1,r_2}$ does not sum to one, heterozygosity is lost.

If there is only one root, then the sum consists of just one term, hence all expected heterozygosities can be expressed in terms of the expected heterozygosity of the root variable. [12] showed that the partial drift of an edge $e = i \to j$ fulfils

$$E(d_e^2) = (1 - e^{-\tau_{ij}})E(V_i(1 - V_i)), \tag{10}$$

so that the F_2-statistics can be expressed in terms of the expected heterozygosities of the root variables (Theorem 4).

To end we consider a few examples. The linear Wright-Fisher admixture graph with edges $1 \to 2$, $2 \to 3$ was already studied in [12, chapter 13] and it was found that

$$F_2(2,3) = e^{-\tau_{12}}(1 - e^{-\tau_{23}})E(V_1(1 - V_1)).$$

It is also a consequence of (10) and additivity of the F_2-statistic.

In a Wright-Fisher admixture graph with edges $1 \rightarrow 2$, $2 \rightarrow 3$, $2 \rightarrow 4$, $3 \rightarrow 5$, $4 \rightarrow 5$, such that there are two overlapping a-paths from 1 to 5, we have

$$F_2(1,5) = \left(1 - \alpha_{35}^2 e^{-\hat{\tau}_1} - \alpha_{45}^2 e^{-\hat{\tau}_2} - 2\alpha_{35}\alpha_{45}e^{-\tau_{12}}\right)E(V_1(1 - V_1)),$$

where $\hat{\tau}_1 = \tau_{12} + \tau_{23} + \tau_{35}$ and $\hat{\tau}_2 = \tau_{12} + \tau_{24} + \tau_{45}$.

Acknowledgement

References

[1] Castelo, R., Roverato, A., 2006. A robust procedure for gaussian graphical model search from microarray data with p larger than n. Journal of Machine Learning Research 7, 2621–2650.

[2] Cavalli-Sforza, L.L., 1966. Population structure and human evolution. Proceedings of the Royal Society of London. Series B, Biological sciences 164, 362–79. doi:10.1098/rspb.1966.0038.

[3] Cavalli-Sforza, L.L., Edwards, A.W., 1967. Phylogenetic analysis. models and estimation procedures. American Journal of Human Genetics 19, 233–57. doi:10.1111/j.1558-5646.1967.tb03411.x.

[4] Cox, D.A., Little, J.B., O'Shea, D., 2007. Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra. Springer.

[5] Ewens, W.J., 1979. Mathematical population genetics. Springer-Verlag.

[6] Frydenberg, M., 1990. The chain graph markov property. Scandinavian Journal of Statistics 17, 333–353.

[7] Green, R.E., Krause, J.e.a., 2010. A draft sequence of the neandertal genome. Science 328, 710–22. doi:10.1126/science.1188021.

[8] Harary, F., 1969. Graph theory. Addison Wesley.

[9] Hoffmann-Jørgensen, J., 1994. Probability with a view toward statistics. Chapman & Hall.

[10] Lauritzen, S., Richardson, T., 2002. Chain graph models and their causal interpretations. J. R. Statist. Soc. 64, 321–361. doi:10.1111/1467-9868.00340.

[11] Lipson, M., Loh, P.R., Levin, A., Reich, D., Patterson, N., Berger, B., 2013. Efficient moment-based inference of admixture parameters and sources of gene flow. Molecular Biology and Evolution 30, 1788–1802. doi:`10.1093/molbev/mst099`.

[12] Nei, M., 1987. Molecular evolutionary genetics. Columbia University Press.

[13] Patterson, N.J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D., 2012. Ancient admixture in human history. Genetics 192, 1065–1093. doi:`10.1534/genetics.112.145037`.

[14] Peter, B.M., 2016. Admixture, population structure and f-statistics. Genetics 202, 1485–1501. doi:`10.1534/genetics.115.183913`.

[15] Pickrell, J.K., Pritchard, J.K., 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8, 1–17. doi:`10.1371/journal.pgen.1002967`.

[16] Reich, D., Thangaraj, K., Patterson, N., Price, A., Singh, L., 2009. Reconstructing indian population history. Nature 461, 489–94. doi:`10.1038/nature08365`.

[17] Skoglund, P., Mallick, S., Bortolini, M.C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M.L., Salzano, F.M., Patterson, N., Reich, D., 2015. Genetic evidence for two founding populations of the americas. Nature 525, 104–108. doi:`10.1038/nature14895`.

[18] Wall, J.D., Yang, M.A., Jay, F., Kim, S.K., Durand, E.Y., Stevison, L.S., Gignoux, C., Woerner, A., Hammer, M.F., Slatkin, M., 2013. Higher levels of neanderthal ancestry in east asians than in europeans. Genetics 194, 199–209. doi:`10.1534/genetics.112.148213`.

[19] Whittaker, J., 1990. Graphical models in applied multivariate statistics. Wiley.

Appendix A

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}, \mathcal{E}$, are the nodes and the edges, respectively, and edges are directed or undirected. A directed edge from i to j and an undirected edge between i and j are written as $i \rightarrow j$ and $i \leftrightarrow j$, respectively. A *path* from node i to node j is a sequence of adjacent nodes $i_0, i_1, \ldots, i_K$, where $i_0 = i, i_K = j$, and either $i_k \rightarrow i_{k+1}$ or $i_k \leftrightarrow i_{k+1}$, for $k = 0, \ldots, K-1$. A *cycle* is a path with $i = j$. A cycle is directed if $i_k \rightarrow i_{k+1}$ for $k = 0, \ldots, K-1$. (A path is *not* the same as an a-path.)

A graph where the nodes are connected by either directed or undirected edges and without directed cycles is called a *chain graph*.

The node set $\mathcal{V}$ of a chain graph $\mathcal{G}$ can be partitioned into unique *blocks* $B_1, \ldots, B_N$ such that $B_1 = \mathcal{R}$, and if two nodes $i, j \in \mathcal{V}$ are connected by a sequence of directed edges $i \to i_1, i_1 \to i_2, \ldots, i_k \to j$ in $\mathcal{E}$, then $i \in B_{n_i}$, $j \in B_{n_j}$ and $n_i < n_j$. Two nodes i, j are said to be in the same *component* if they are connected by an undirected or directed path from i to j and from j to i. Note that a component does not coincide with a block. In Figure 1B each node is a distinct component, but the blocks are $B_1 = \{1\}$, $B_2 = \{2, 8\}$, $B_3 = \{3, 4\}$, $B_4 = \{5, 6, 7\}$.

A subset $A \subseteq \mathcal{V}$ of a chain graph $\mathcal{G}$ induces a graph $\mathcal{G}(A) = (A, \mathcal{E}(A))$, where $\mathcal{E}(A)$ contains the edges of $\mathcal{G}$ whose nodes are in A

The *border* of a subset $\mathcal{A} \subseteq \mathcal{V}$ is defined as

$$bd(\mathcal{A}) := \big\{ i \in \mathcal{V} \ : \ i \to j \text{ or } i \leftrightarrow j \text{ for some } j \in A \big\},$$

and the moral graph of $\mathcal{G}$ is $\mathcal{G}^m = (\mathcal{V}, \mathcal{E}^m)$, where $\mathcal{E}^m$ consists of the union of

- the set $\mathcal{E}^u$ consisting of the edges of $\mathcal{E}$ made undirected,

- the set of undirected edges of $\mathcal{E}$ that connect pairs of nodes that are in the border of a component of $\mathcal{G}$.

In Figure 1B the moral graph is obtained by making all edges undirected, and by adding an undirected edge between 3 and 4.

Let A, B, C be three disjoint subsets of $\mathcal{V}$. The sets A and B are *separated* by C when every path from $i \in A$ to $j \in B$ (or vice versa) includes at least one node of C. Given $A \subseteq V$, the set of ancestors of A, $an_G(A)$, is the subset of nodes of $\mathcal{V}$ having at least one path to some node of A.

A random vector $(U_i | i \in \mathcal{V})$ is a chain graph model over $\mathcal{G}$ if it fulfils the *global $\mathcal{G}$-Markovian* (GM) property,

$$(U_i | i \in A) \ \perp \ (U_i | i \in B) \ \ | \ \ (U_i | i \in C),$$

whenever C separates A and B in $\mathcal{G}^m_{an(A \cup B \cup C)}$ and $X \perp Y | Z$ denotes that the random vectors X and Y are conditionally independent given Z.

Let $(V_i | i \in \mathcal{V})$ be a stochastic admixture graph over $\mathcal{G}$. The Markov structure of the chain graph implies in particular that two contribution variables $C_{ij}, C_{k\ell}$, where i, k are not necessarily distinct, are conditionally independent given the variables V_i, V_k of their parents, that is,

$$C_{ij} \perp C_{k\ell} \mid (V_i, V_k). \tag{11}$$

Further, for a contribution variable C_{ij}, let B_{n_i} be the block in which node i is located. Then

$$C_{ij} \perp \left(V_k \mid k \in \bigcup\nolimits_{n=1}^{n_i} B_n \right) \mid V_i. \tag{12}$$

16

As a consequence of (12), if a node k is in $\bigcup_{n=1}^{n_i} B_n$, it follows that

$$C_{ij} \perp V_k \mid V_i. \tag{13}$$

Appendix B

Proofs of statements in the main text are presented here. We start by listing in Table 1 the possible forms of an a-path between two distinct nodes $i, j \in \mathcal{V}$. In a,e) there might be only one edge, in the others at least two. For two a-paths γ, δ, the concatenation of γ with δ is the sequence of edges in γ followed by those of δ. It might or might not be an a-path.

$$
\begin{array}{llllllll}
\text{a)} & i & \leftarrow & \dots & \leftarrow & j & \text{d)} & i & \leftrightarrow & \dots & \rightarrow & j \\
\text{b)} & i & \leftarrow & \dots & \leftrightarrow & j & \text{e)} & i & \rightarrow & \dots & \rightarrow & j \\
\text{c)} & i & \leftarrow & \dots & \rightarrow & j
\end{array}
$$

Table 1: Possible admixture paths according to Definition 2.

Lemma 3. Let $i, j \in \mathcal{V}$ be two distinct nodes of an admixture graph $\mathcal{G}$, such that not both of them are roots. Then either all a-paths $\gamma \in \Gamma_{ij}$ end with a directed edge $k \rightarrow j$, for some $k \in \mathcal{V}$, or all a-paths $\gamma \in \Gamma_{ij}$ start with a directed edge $i \leftarrow k$, for some $k \in \mathcal{V}$, where k might depend on γ. $\quad\square$

Proof of Lemma 3. It follows by inspecting the possibilities for a-paths.

Lemma 4. Let $i, j \in \mathcal{V}$ be two distinct nodes of an admixture graph $\mathcal{G}$, such that $i \notin \mathcal{R}$ and assume that all a-paths $\gamma \in \Gamma_{ij}$ start with an edge $i \leftarrow k$, for some $k \in \mathcal{V}$ (as in Lemma 3). Then

$$\Gamma_{ij} = \cup_{\ell \in par(i)} \{(i \leftarrow \ell, \gamma) \mid \gamma \in \Gamma_{\ell j}\},$$

where $(i \leftarrow \ell, \gamma)$ denotes the concatenation of the a-path $i \leftarrow \ell$ with γ. $\quad\square$

Proof of Lemma 4. It follows trivially that Γ_{ij} is included in the right hand side of the equality. To prove the converse we need to show that $(i \leftarrow i', \gamma)$, $\gamma \in \Gamma_{i'j}$ is an a-path from i to j, that is, the edges are directed according to Definition 2 (see Table 1) and i does not belong to γ. By inspection of the possibilities for γ, it shows that the edges have the correct directions (including undirected). Assume i is in γ. It cannot be a root by assumption. Then by inspection, either there is a subsequence $i' \leftarrow \dots \leftarrow i$ of γ, or a subsequence $i \rightarrow \dots \rightarrow j$. In the former case, this yields a directed cycle from i to i via i' by adding the edge $i' \rightarrow i$. In the latter case, it yields an a-path from i to j, violating the assumption of the starting edge. Hence the proof is completed. $\quad\square$

Lemma 5. Let $i, j, k \in \mathcal{V}$ be distinct nodes of an admixture graph $\mathcal{G}$, and assume any a-path of Γ_{ij} passes through k. Then it holds that

$$\Gamma_{ij} = \Gamma_{ik} \times \Gamma_{kj},$$

where $(\gamma, \delta) \in \Gamma_{ik} \times \Gamma_{kj}$ denotes concatenation of γ and δ. $\qquad\square$

Proof of Lemma 5. We will prove the two sets are included into each other. It trivially holds that $\Gamma_{ij} \subseteq \Gamma_{ik} \times \Gamma_{kj}$. Next we prove the converse inclusion. According to Lemma 3, we might assume all a-paths $\gamma \in \Gamma_{ik}$ start with a directed edge $i \leftarrow i'$, $i' \in par(i)$, hence the same is true for the a-paths of Γ_{ij}. According to Table 1 this yields $3 \cdot 5 = 15$ possibilities for the forms of (γ, δ). In the cases (a,a), (a,b), (a,c), (a,d), (a,e), (b,e), (c,e) the concatenation of γ with δ is an a-path from i to j. The cases (b,a), (b,b), (b,c), (c,d) give impossible constraints on k, hence they cannot occur. The case (b,d) gives $i \leftarrow \ldots \leftarrow k_1 \leftrightarrow k \leftrightarrow k_2 \rightarrow \ldots \rightarrow j$ for two roots $k_1, k_2 \in \mathcal{R}$. Hence there is an a-path $i \leftarrow \ldots \leftarrow k_1 \leftrightarrow k_2 \rightarrow \ldots \rightarrow j$, not going though k, which is impossible. The remaining three cases are

				γ				δ		
(c,a)	i	$\leftarrow$	$\ldots$	$\rightarrow$	k	$\leftarrow$	$\ldots$	$\leftarrow$	j
(c,b)	i	$\leftarrow$	$\ldots$	$\rightarrow$	k	$\leftarrow$	$\ldots$	$\leftrightarrow$	j
(c,c)	i	$\leftarrow$	$\ldots$	$\rightarrow$	k	$\leftarrow$	$\ldots$	$\rightarrow$	j

The node k cannot be a root, $k \notin \mathcal{R}$. We will construct an a-path from i to j, based on γ and δ, that does not go through k, hence (c,a), (c,b), (c,c) cannot occur. Specifically, there are three possibilities for the a-path γ,

i	$\leftarrow$	$\ldots$	$\leftarrow$		i'		$\rightarrow$	$\ldots$	$\rightarrow$	k
i	$\leftarrow$	$\ldots$	$\leftarrow$		r		$\rightarrow$	$\ldots$	$\rightarrow$	k
i	$\leftarrow$	$\ldots$	$\leftarrow$	r	$\leftrightarrow$	r'	$\rightarrow$	$\ldots$	$\rightarrow$	k

In the first case, there is $r \in \mathcal{R}$ and an a-path $r \rightarrow \ldots \rightarrow i'$ by definition of an admixture graph. It does not contain k as this would create a directed cycle. Consequently, in all three cases, there is $r \in \mathcal{R}$ and an a-path $r \rightarrow \ldots \rightarrow i$, not containing k.

For δ we can do similarly. In the first case above, there exists $r' \in \mathcal{R}$ and an a-path $r' \rightarrow \ldots \rightarrow j$, not containing k as this would create a directed cycle. In the second case, $j \in \mathcal{R}$ and we define $r' = j$. In the third case, we proceed as for γ and conclude there is $r' \in \mathcal{R}$ and an a-path $r' \rightarrow \ldots \rightarrow j$, not containing k.

Now consider the a-path η from r to i, and the a-path η' from r' to j, constructed above ($r = r'$ or $r \neq r'$). If η and η' do not share any nodes, except perhaps for r, r', then their concatenation is an a-path $(\eta, \eta') \in \Gamma_{ij}$, not containing k. If they do share nodes, chose sub-a-paths $\zeta \colon \ell \rightarrow \ldots \rightarrow i$ and $\zeta' \colon \ell \rightarrow \ldots \rightarrow j$, sharing only the node ℓ. Then their concatenation

is an a-path $(\zeta, \zeta') \in \Gamma_{ij}$, not containing k. Hence, none of the cases (c,a), (c,b), (c,c) are valid, and the lemma holds.

Proof of Proposition 1. If $i = j$ then the statements are true by definition. Given distinct $i, j \in \mathcal{A}$, there exists $r_1, r_2 \in \mathcal{R}$ and two a-paths $r_1 \to i_1 \to \ldots \to i_{k-1} \to i$ and $r_2 \to j_1 \to \ldots \to j_{k'-1} \to j$ for some nodes $i_1, \ldots, i_{k-1}, j_1, \ldots,$
$j_{k'-1} \in \mathcal{A}$, and $k, k' \geq 1$ (because the admixture graph is connected and there are no directed cycles). Either all these nodes are distinct (except perhaps for r_1, r_2) in which case they form an a-path from i to j by concatenation, or there is $i_\ell = i'_{\ell'}$ for some ℓ, ℓ'. Choose ℓ, ℓ' such that $\ell + \ell' \leq k + k'$ is as large as possible. Then $i \leftarrow i_{k-1} \leftarrow \ldots \leftarrow i_{\ell+1} \leftarrow i_\ell \to i'_{\ell'+1} \to \ldots \to i'_{k'-1} \to j$ is an a-path from i to j by definition of an a-path. There cannot be any repeated nodes, otherwise $\ell + \ell'$ is not as large as possible. If $i, j \in \mathcal{R}$, then they are trivially connected by an a-path. If $i \in \mathcal{R}$ and $j \in \mathcal{A}$, then there is an a-path $j \leftarrow i_{k-1}, \ldots \leftarrow i_1 \leftarrow i$ (potentially with $i_1 \leftrightarrow i$ and i_1 a root) for some $i_1, \ldots, i_{k-1} \in \mathcal{A}$ (as before). Hence $\Gamma_{ij} \neq \emptyset$.

To prove the second part of the proposition, we proceed by induction in the length of the a-paths. For $i, j \in \mathcal{V}$, consider Γ_{ij} and let l_γ denote the number of edges in an a-path $\gamma \in \Gamma_{ij}$. Assume $l_\gamma \leq 0$ for all $\gamma \in \Gamma_{ij}$. Hence $i = j$ and $\sum_{\gamma \in \Gamma_{ij}} p_\gamma = 1$ by definition.

Assume now the statement holds if $\sum_{\gamma \in \Gamma_{ij}} p_\gamma = 1$ and $l_\gamma \leq k, \gamma \in \Gamma_{ij}$, for some $k \geq 1$. Consider two nodes $i, j \in \mathcal{V}$ such that all a-paths between them fulfil $l_\gamma \leq k + 1$ and start with an edge $i \leftarrow k$ for some $k \in \mathcal{V}$ (otherwise exchange the roles of i and j). Then, using Lemma 4 to decompose the a-paths according to the parents of i, we obtain

$$\sum_{\gamma \in \Gamma_{ij}} p_\gamma = \sum_{\ell \in par(i)} \sum_{\gamma \in \Gamma_{\ell j}} \alpha_{i\ell} p_\gamma = \sum_{\ell \in par(i)} \alpha_{i\ell} \sum_{\gamma \in \Gamma_{\ell j}} p_\gamma = \sum_{\ell \in par(i)} \alpha_{i\ell} = 1,$$

as all a-paths in $\Gamma_{\ell j}$ must have length at most k. $\square$

Proof of Theorem 1. We will prove $(ii) \Rightarrow (i) \Rightarrow (iii) \Rightarrow (ii)$. Assume (ii). If the label p_γ of an a-path between i and j is one, then $\Gamma_{ij} = \{\gamma\}$, according to Proposition 1. It proves (i). Let $\mathcal{A}_r$ be the nodes in $\mathcal{A}$ for which there is an a-path from $r \in \mathcal{R}$ to a node in $\mathcal{A}$, not involving any other root. Any $i \in \mathcal{A}$ is in at least one $\mathcal{A}_r$, and cannot be in two such sets $\mathcal{A}_{r_1}, \mathcal{A}_{r_2}$, because then there would be a-paths $(r_1, \ldots, i)$ and $(r_1, r_2, \ldots, i)$ from r_1 to i, contradicting uniqueness of a-paths (i). The set $\{r\} \cup \mathcal{A}_r$ with the inherited edges form a tree. It proves (iii). The last implication is straightforward using the definition of a forest and taking into account the undirected edges between the roots, so (ii) is proven. $\square$

Proof of Proposition 2. Consider $r_1 \in \mathcal{R}$. Any a-path $\gamma_1 \in \Gamma_{\ell r_1}$ contains one root or two roots. In the latter case, the a-path is the concatenation of an a-path $\gamma' \in \Omega_{\ell r_2}$, $r_2 \neq r_1$, with the edge $r_2 \leftrightarrow r_1$ and $p_\gamma = p_{\gamma'}$. It follows

that

$$\sum_{r \in \mathcal{R}} q_{\ell r} = \sum_{r \in \mathcal{R}} \sum_{\gamma \in \Omega_{\ell r}} p_\gamma = \sum_{\gamma \in \Gamma_{\ell r_1}} p_\gamma = 1.$$

$\square$

Proof of Proposition 3. For the if and only if, assume (ii). By Definition 4, we have $\mathcal{V}_{\mathcal{A}_0} = \mathcal{V}$ and $\mathcal{E}_{\mathcal{A}_0} = \mathcal{E}$. Oppositely, if $\mathcal{G}_{\mathcal{A}_0} = \mathcal{G}$, then again by Definition 4, (ii) holds.

We next prove the equivalence between (i) and (ii). (ii) trivially implies (i). To prove the converse, let $\gamma \in \Gamma_{ik}$, $\delta \in \Gamma_{kj}$ be as stated, sharing only the node $k \in \mathcal{V} \setminus \mathcal{A}_0$ and $i, j \in \mathcal{A}_0$. There are five possibilities for a-paths, see Table 1. As $i, j \in \mathcal{A}_0$, then d,e) are not an option for γ, and a,b) are not an option for δ, yielding nine possible combinations. By inspection, (a,c), (a,d), (a,e), (b,e), (c,e) yield valid a-paths from i and j, hence (ii) holds in these case. The cases (b,c), (c,d) yield impossible constraint on k, and thus cannot occur. The two remaining cases (b,d), (c,c) give

$$
\begin{array}{ccccccccc}
 & & & \gamma & & & & \delta & \\
(\text{b,d}) & i & \leftarrow & \ldots & \leftrightarrow & k & \leftrightarrow & \ldots & \rightarrow & j \\
(\text{c,c}) & i & \leftarrow & \ldots & \rightarrow & k & \leftarrow & \ldots & \rightarrow & j
\end{array}
$$

Since $k \notin \mathcal{A}_0$ then there is an a-path, say η, $k \rightarrow \ldots \rightarrow k'$, with $k' \in \mathcal{A}_0$ (in the first case we use that every root has a child). If this a-path does not share a node with γ (similarly, with δ) except from k, then γ concatenated with η is an a-path from i to k' through k, as required. Otherwise, if η shares nodes with both a-paths, let $k'' \neq k$ (say in δ, similar if in γ) be the first such node counting from k. Let η' be the a-path from k to j given by $k \rightarrow \ldots \rightarrow k'' \rightarrow \ldots \rightarrow j$ by concatenating the part of η ending at k'' with the part of δ beginning at k''. The edge from k'' towards j cannot be $\leftrightarrow$ nor $\leftarrow$ for the following reasons. In the former case, $\rightarrow k'' \leftrightarrow$ which is impossible. In the latter case, δ takes the form $k \leftarrow \ldots \leftarrow k'' \leftarrow \ldots \leftarrow j$, creating a directed loop from k to k'' and back to k via η'. Finally, γ concatenated with η' is an a-path from i to j through k, as required.

Proof of Theorem 2. A general reference is [9]. Let V_i, V_j be two root variables, corresponding to distinct roots in $\mathcal{R}$. By definition of conditional expectation (first equality) and by assumption (second equality)

$$\int_A V_i \, dP = \int_A E(V_i \mid V_j) \, dP = \int_A V_j \, dP$$

for all R_j-measurable subsets A. Similarly, by exchanging the role of V_i and V_j, the same holds for all V_i-measurable subsets. Hence,

$$\int_A V_i \, dP = \int_A V_j \, dP$$

20

for all (V_i, V_j)-measurable subsets A. Consequently, using that V_i, V_j have finite expectation by assumption, it follows that $V_i = V_j$. $\qquad\square$

Proof of Theorem 3. We prove the statement by induction on the maximum number of edges n of the a-paths $\gamma \in \Gamma_{ij}$. If $n = 0$, then $i = j$ and there is nothing to prove. Assume now (2) for is true for $n = \ell$ and consider $n = \ell + 1$ for some $\ell \geq 1$. Further, assume that i is such that any a-path of Γ_{ij} involves a parent of i as the first node (Lemma 3). Then using the inductive hypothesis, Proposition 1, Definition 5(ii), Lemma 4 and (1), it holds that

$$
\begin{aligned}
\sum_{\gamma \in \Gamma_{ij}} p_\gamma \sum_{e \in \gamma} d_e^\gamma &= \sum_{i' \in par(i)} \sum_{\gamma' \in \Gamma_{i'j}} \alpha_{i'i} p_{\gamma'} \left(d_{i'i}^{(i,i')} + \sum_{e \in \gamma'} d_e^{\gamma'} \right) \\
&= \sum_{i' \in par(i)} \alpha_{i'i} \left(d_{i'i}^{(i,i')} + \sum_{\gamma' \in \Gamma_{i'j}} p_{\gamma'} \sum_{e \in \gamma'} d_e^{\gamma'} \right) \\
&= \sum_{i' \in par(i)} \alpha_{i'i} \left(C_{i'i} - V_{i'} + V_{i'} - V_j \right) \\
&= \sum_{i' \in par(i)} \alpha_{i'i} C_{i'i} - \sum_{i' \in par(i)} \alpha_{i'i} V_j = V_i - V_j.
\end{aligned}
$$

If any a-path of Γ_{ij} does not involve a parent of i as the first node after i, then it holds that any a-path involves a parent of j as second last node (Lemma 4). Note that $D_{ij} = -D_{ji}$, and that the decomposition holds for D_{ji}. Thus

$$
D_{ij} = - \sum_{\gamma \in \Gamma_{ji}} p_\gamma \sum_{e \in \gamma} d_e^\gamma = \sum_{\gamma \in \Gamma_{ij}} p_\gamma \sum_{e \in \gamma} d_e^\gamma,
$$

using that the partial drifts change sign when an a-path is reversed. $\qquad\square$

Proof of Lemma 1. Let $e_1 = i \to j$ and $e_2 = k \to \ell$ (with one of the two potentially being undirected). Further, assume that the blocks B_{n_i}, B_{n_k}, associated to the nodes i, k are such that $n_i \geq n_k$. The expected value in (3) can be rewritten as

$$
E(d_{ij} d_{k\ell}) = E(C_{ij} C_{k\ell}) - E(C_{ij} V_k) - E(C_{k\ell} V_i) + E(V_i V_k).
$$

Consider the first expectation,

$$
E(C_{ij} C_{k\ell}) = E(E(C_{ij} C_{k\ell} | V_i, V_k)) = E(E(C_{ij} | V_i, V_k) E(C_{k\ell} | V_i, V_k)) \quad (14)
$$
$$
= E(E(C_{ij} | V_i) E(C_{k\ell} | V_i, V_k)) = E(V_i E(C_{k\ell} | V_i, V_k)). \quad (15)
$$

Here we used (11), (13) to derive (14), and Definition 5 to derive (15). With similar considerations, it follows that

$$
E(C_{k\ell} V_i) = E(E(C_{k\ell} V_i | V_i, V_k)) = E(V_i E(C_{k\ell} | V_i, V_k)).
$$

Finally,

$$E(C_{ij}V_k) = E(E(C_{ij}V_k|V_i)) \tag{16}$$

$$= E(E(C_{ij}|V_i)E(V_k|V_i)) \tag{17}$$

$$= E(V_iE(V_k|V_i)) = E(E(V_iV_k|V_i)) = E(V_iV_k), \tag{18}$$

using (13) and Definition 5(ii) to derive (16) and (17), respectively, and properties of the conditional expectation to derive (18). Combining it all yields $E(d_{ij}d_{k\ell}) = 0$. Finally consider $E(d_{ij}^2) = E(C_{ij}^2 + V_i^2 - 2C_{ij}V_i)$. Using the definition of variance and Definition 5(ii), it follows that

$$E(C_{ij}^2) = Var(C_{ij}) + E(V_i)^2 \quad \text{and} \quad E(V_i^2) = Var(V_i) + E(V_i)^2.$$

By conditioning on V_i and using Definition 5(ii) again, we have $E(C_{ij}V_i) = -Var(V_i) - E(V_i)^2$. Using the linearity of the expectation it can be concluded that $E(d_{ij}^2) = Var(C_{ij}) - Var(V_i)$. $\square$

Proof of Theorem 4. Rewrite the definition of $F_2(i,j)$ using the decomposition of drift in (2),

$$F_2(i,j) = E\left[\left(\sum_{\gamma_1 \in \Gamma_{ij}} \left(p_{\gamma_1} \sum_{e_1 \in \gamma_1} d_{e_1}^{\gamma_1} \right) \right) \left(\sum_{\gamma_2 \in \Gamma_{ij}} \left(p_{\gamma_2} \sum_{e_2 \in \gamma_2} d_{e_2}^{\gamma_2} \right) \right) \right]. \tag{19}$$

Distributing the products and exploiting the linearity of the expectation, (19) is equivalent to

$$F_2(i,j) = \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ij} \times \Gamma_{ij}} \sum_{e_1 \in \gamma_1} \sum_{e_2 \in \gamma_2} p_{\gamma_1} p_{\gamma_2} E(d_{e_1}^{\gamma_1} d_{e_2}^{\gamma_2}),$$

$$= \sum_{(e_1,e_2) \in \mathcal{E}_{ij} \times \mathcal{E}_{ij}} \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ij}^{e_1} \times \Gamma_{ij}^{e_2}} p_{\gamma_1} p_{\gamma_2} E(d_{e_1}^{\gamma_1} d_{e_2}^{\gamma_2}).$$

Observe that $\mathcal{E}_{ij}$ is the disjoint union of $\mathcal{E}_{ij}^d$ and $\mathcal{E}_{ij}^u$. Moreover, the product of two distinct edges, where at least one is directed, has expectation zero, see Lemma 1. Thus

$$F_2(i,j) = \sum_{e \in \mathcal{E}_{ij}^d} \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ij}^e \times \Gamma_{ij}^e} \text{sign}_{\gamma_1}(e)\, \text{sign}_{\gamma_2}(e) p_{\gamma_1} p_{\gamma_2}\, E(d_e^2)$$

$$+ \sum_{(e_1,e_2) \in \mathcal{E}_{ij}^u \times \mathcal{E}_{ij}^u} \sum_{(\gamma_1,\gamma_2) \in \Gamma_{ij}^{e_1} \times \Gamma_{ij}^{e_2}} \text{sign}_{\gamma_1}(e_1)\, \text{sign}_{\gamma_2}(e_2)\, p_{\gamma_1} p_{\gamma_2}\, E(d_{e_1} d_{e_2})$$

$$= \sum_{e \in \mathcal{E}_{ij}^d} A_e^{(i,j)} E(d_e^2) + \sum_{(e_1,e_2) \in \mathcal{E}_{ij}^u \times \mathcal{E}_{ij}^u} B_{e_1,e_2}^{(i,j)} E(d_{e_1} d_{e_2}).$$

$\square$

Proof of Lemma 2. (i) Let $i, j \in \mathcal{V}$ and $e \in \mathcal{E}$. Using Proposition 1, $\Gamma_{ij}^{e_1} \subseteq \Gamma_{ij}$ and the definition of the sign function, it follows that

$$1 = \sum_{\gamma_1 \in \Gamma_{ij}} p_{\gamma_1} \geq \sum_{\gamma_1 \in \Gamma_{ij}^{e_1}} p_{\gamma_1} \geq \sum_{\gamma_1 \in \Gamma_{ij}^{e_1}} \text{sign}_{\gamma_1}(e_1) p_{\gamma_1}. \qquad (20)$$

An analogous inequality holds by changing the sign of each term. The same results apply to the pair of nodes $k, \ell \in \mathcal{V}$. Therefore

$$-1 \leq \left(\sum_{\gamma_1 \in \Gamma_{ij}^{e_1}} \text{sign}_{\gamma_1}(e_1) p_{\gamma_1} \right) \left(\sum_{\gamma_2 \in \Gamma_{k\ell}^{e_2}} \text{sign}_{\gamma_2}(e_2) p_{\gamma_2} \right) \leq +1,$$

hence the statement is proved. (ii) The value of $|B_{e_1,e_2}^{(i,j)(k,\ell)}|$ is 1 if and only if each of the two factors of the B-coefficient assumes the value ± 1. Equivalently, (20) and the analogous with changed sign, has an equality for both pairs of nodes (i, j) and (k, ℓ). This is verified only under the conditions contemplated in statement (ii), hence (ii) is proved. (iii) Observe that, if two distinct a-paths of Γ_{ij} share some edges, they must go through two different outgoing edges of a shared node, thus their path labels differ for at least one edge label in their factorization. An analogous statement holds for a-paths of $\Gamma_{k\ell}$. Hence there are no terms resulting from the product in (5) that can sum to zero. Therefore it holds that $B_{e_1,e_2}^{(i,j)(k,\ell)} = 0$ for all $\alpha_e, e \in \mathcal{E}$, if and only if one of the two factors in (5) is equal to zero for all $\alpha_e, e \in \mathcal{E}$ (see [4, p. 4]). Thus at least one of these factors is equal to zero [4, p. 510]. This condition is verified if and only if one or both sets $\Gamma_{ij}^{e_1}$ and $\Gamma_{k\ell}^{e_2}$ do not contain any a-path. $\qquad \square$

Proof of Proposition 4. Assume that all a-paths of Γ_{ij} pass through k, and that i, j, k are all distinct, otherwise there is nothing to prove. By Lemma 5,

$$\Gamma_{ij} = \Gamma_{ik} \times \Gamma_{kj},$$

where $(\gamma, \delta) \in \Gamma_{ik} \times \Gamma_{kj}$ denotes concatenation of γ and δ. By Lemma 1, we have

$$\begin{aligned} D_{ij} &= \sum_{\eta \in \Gamma_{ij}} \left(p_\eta \sum_{e \in \eta} d_e^\eta \right) \\ &= \sum_{\gamma \in \Gamma_{ik}} \sum_{\delta \in \Gamma_{kj}} \left(p_\gamma p_\delta \left(\sum_{e \in \gamma} d_e^\gamma + \sum_{e \in \delta} d_e^\delta \right) \right) \\ &= D_{ik} + D_{kj}, \end{aligned}$$

using Lemma 4, the definition of path label and drift. Hence $D_{ij}^2 = D_{ik}^2 + D_{kj}^2 + 2D_{ik}D_{kj}$. Note that no edge can be shared between $\gamma \in \Gamma_{ik}$ and $\delta \in \Gamma_{kj}$, as $(\gamma, \delta) \in \Gamma_{ij}$ is an a-path. Similarly, there can at most be one undirected edge in (γ, δ), hence by Lemma 1, $E(D_{ik}D_{kj}) = 0$, and

$$F_2(i, j) = E(D_{ij}^2) = E(D_{ik}^2) + E(D_{kj}^2) = F_2(i, k) + F_2(k, j).$$

For the reverse statement, assume there exists an a-path $\gamma \in \Gamma_{ij}$ that does not pass through k, and assume that all a-paths of Γ_{ij} start with an edge of type $i \leftarrow i'$, for some $i' \in par(i)$, i.e. they are of type (a,b,c,d) (see Table 1). Then there is at least one pair of a-paths $\gamma_1 \in \Gamma_{ik}$, $\gamma_2 \in \Gamma_{kj}$ sharing edges. In fact, denote by δ_1, δ_2 the two subpaths of γ from i to k' and from k' to j, respectively. Here, k' is the first node not having parents in γ, so that δ_1 is fixed as an a-path of type (a), whereas δ_2 can be of type (d) or (e). From Proposition 1, there is an a-path $\delta_1' \in \Gamma_{k'k}$ that can be of any type. Let $\delta_2' \in \Gamma_{kk'}$ be the reverse a-path of δ_1'. It is possible to verify by inspection that the concatenations $\gamma_1 = (\delta_1, \delta_1')$, $\gamma_2 = (\delta_2', \delta_2)$, are two a-paths of Γ_{ik}, Γ_{kj}, respectively, and share the edges of δ_1' and δ_2'. With an analogous construction, the same holds for the other possible types of a-paths of Γ_{ij}.

Since $F_2(i,j) = F_2(i,k) + F_2(k,j)$, then $E(D_{ik}D_{kj}) = 0$. Note that

$$E(D_{ik}D_{kj}) = \sum_{\substack{\gamma_1 \in \Gamma_{ik} \\ \gamma_2 \in \Gamma_{kj}}} p_{\gamma_1} p_{\gamma_2} \sum_{e \in \mathcal{E}_{ik} \cap \mathcal{E}_{kj}} \text{sign}_{\gamma_1}(e)\text{sign}_{\gamma_2}(e)E(d_e^2)$$

$$= \sum_{e \in \mathcal{E}} \left(\sum_{\gamma_1 \in \Gamma_{ik}} \text{sign}_{\gamma_1}(e)p_{\gamma_1} \right) \left(\sum_{\gamma_2 \in \Gamma_{kj}} \text{sign}_{\gamma_2}(e)p_{\gamma_2} \right) E(d_e^2)$$

where we have used the definition of sign and the linearity of expectation. The expression is linear in $E(d_e^2)$, $e \in \mathcal{E}$, hence if $E(D_{ik}D_{kj}) = 0$ for arbitrary (positive) values of $E(d_e^2)$, then the coefficient of $E(d_e^2)$, $e \in \mathcal{E}$, is zero. The argument in the proof of Lemma 2(iii) also applies here to conclude that either of the two factors in the coefficient is zero. Since $\Gamma_{ik} \neq \emptyset$ and $\Gamma_{kj} \neq \emptyset$ (assuming i, j, k are different), then $\text{sign}_{\gamma_1}(e) = \text{sign}_{\gamma_2}(e) = 0$, contradicting the existence of a pair of overlapping a-paths. Therefore all a-paths of Γ_{ij} go through k. $\qquad\square$

Proof of Proposition 5. Assume $F_{3-}(i; j, k) \neq 0$. Then there is at least one edge $e \in \mathcal{E}$ for which $A_{e-}^{(i,j)(i,k)} \neq 0$. It follows that there is a pair of a-paths $(\gamma_1, \gamma_2) \in \Gamma_{ij}^e \times \Gamma_{ik}^e$ such that $\text{sign}_{\gamma_1}(e) \neq \text{sign}_{\gamma_2}(e)$. Let us consider the possible combinations of two a-paths $(\gamma_1, \gamma_2) \in \Gamma_{ij} \times \Gamma_{ik}$ from Table 1, and verify for which pairs it is possible to have $\text{sign}_{\gamma_1}(e) \neq \text{sign}_{\gamma_2}(e)$. By inspection, it is immediate to verify that e has the same sign in the combinations (a,a), (a,b), (a,c), (b,d), (d,d) (e,e). Combination (a,d) is not possible, because i would be a root and an admixed node at the same time. Combinations (a,e), (b,e), (c,e) and (d,e) imply that $e \notin \Gamma_{ij}^e \cap \Gamma_{ik}^e$. There remains combinations (b,b), (b,c) and (c,c), that admit one or more shared edges of opposite sign in the two paths. The opposite implication is immediate by definition of $F_{3-}(i; j, k)$ and $A_{e-}^{(i,j)(i,k)}$. $\qquad\square$

Proof of Proposition 6. We will prove a slightly stronger statement, namely

$$E(V_j(1 - V_{j'})) = \sum_{(r,r')\in\mathcal{R}\times\mathcal{R}} D^{j,j'}_{r,r'} E(V_r(1 - V_{r'})), \quad j, j' \in \mathcal{V},$$

by induction on the block index (Appendix A), and where $D^{j,j'}_{r,r'}$ are non-negative and sum to at most one over r, r' for any j, j'.

First consider nodes $j, j' \in B_1 = \mathcal{R}$. By defining $D^{j,j'}_{r,r'} = 1$ if $r = j$, $r' = j'$, and zero otherwise, we have

$$E(V_j(1 - V_{j'})) = \sum_{(r,r')\in\mathcal{R}\times\mathcal{R}} D^{j,j'}_{r,r'} E(V_r(1 - V_{r'})),$$

and the claim holds for $j, j' \in B_1$. Now assume the statement holds for nodes in $\cup_{k=1}^m B_k$, $m \geq 1$, and consider $j, j' \in \cup_{k=1}^{m+1} B_k$ such that at least one of j, j' is in B_{m+1}, say $j \in B_{m+1}$ (similar if $j' \in B_{m+1}$). If $j \notin B_1$ then

$$E(V_j(1 - V_{j'})) = E\left(\sum_{i\in par(j)} \alpha_{ij} C_{ij} \left(1 - \sum_{i'\in par(j')} \alpha_{i'j'} C_{i'j'}\right)\right) \qquad (21)$$

$$= \sum_{(i,i')\in par(j)\times par(j')} \alpha_{ij}\alpha_{i'j'} E(C_{ij}(1 - C_{i'j'})),$$

using that $1 = \sum_{i'\in par(j')} \alpha_{i'j'}$. If $(i, j) \neq (i', j')$ then $E(C_{ij}(1 - C_{i'j'})) = E(V_i(1 - V_{i'}))$ by conditioning on $(V_i, V_{i'})$ and using Definition 5(iii). If $(i, j) = (i', j')$ then $E(C_{ij}(1 - C_{ij}) = e^{-\tau_{ij}} E(V_i(1 - V_i))$, where τ_{ij} is the population scaled time from i and j [12]. Define $P(j, j') = par(j) \times par(j')$. Then

$$E(V_j(1 - V_{j'})) = \sum_{(i,i')\in P(j,j')} e^{-\tau_{ij}\delta_{ii'}} \alpha_{ij}\alpha_{i'j'} E(V_i(1 - V_{i'}))$$

$$= \sum_{(i,i')\in P(j,j')} e^{-\tau_{ij}\delta_{ii'}} \alpha_{ij}\alpha_{i'j'} \sum_{(r,r')\in\mathcal{R}\times\mathcal{R}} D^{i,i'}_{r,r'} E(V_r(1 - V_{r'}))$$

$$= \sum_{(r,r')\in\mathcal{R}\times\mathcal{R}} \left[\sum_{(i,i')\in P(j,j')} e^{-\tau_{ij}\delta_{ii'}} \alpha_{ij}\alpha_{i'j'} D^{i,i'}_{r,r'} E(V_r(1 - V_{r'}))\right]$$

$$= \sum_{(r,r')\in\mathcal{R}\times\mathcal{R}} D^{j,j'}_{r,r'} E(V_r(1 - V_{r'}))$$

where the second line holds by induction hypothesis since $i, i' \in \cup_{k=1}^m B_k$, $\delta_{ii'} = 1$ if $i = i'$ and zero otherwise, and

$$D^{j,j'}_{r,r'} = \sum_{(i,i')\in P(j,j')} e^{-\tau_{ij}\delta_{ii'}} \alpha_{ij}\alpha_{i'j'} D^{i,i'}_{r,r'}.$$

It is easily seen that the constants above have the desired properties. If $j \in B_1$ (similar if $j' \in B_1$) then we do the same as above except only $V_{j'}$ is replaced by a sum in (21). We obtain similarly to (21),

$$
\begin{aligned}
E(V_j(1 - V_{j'})) &= \sum_{i' \in par(j')} \alpha_{i'j'} E(V_j(1 - C_{i'j'})) \\
&= \sum_{i' \in par(j')} \alpha_{i'j'} E(V_j(1 - V_{i'})) \\
&= \sum_{i' \in par(j')} \alpha_{i'j'} \sum_{(r,r') \in \mathcal{R} \times \mathcal{R}} D_{r,r'}^{j,i'} E(V_r(1 - V_{r'})) \\
&= \sum_{(r,r') \in \mathcal{R} \times \mathcal{R}} D_{r,r'}^{j,j'} E(V_r(1 - V_{r'})),
\end{aligned}
$$

where

$$
D_{r,r'}^{j,j'} = \sum_{i' \in par(j')} \alpha_{i'j'} D_{r,r'}^{j,i'}.
$$

If $j' \in B_1$ then the role of j and j' is reversed, yielding a similar answer. In both cases the constants fulfil the desired requirements.

Letting $j = j'$ yields the desired result. $\qquad \square$