Supplementary material

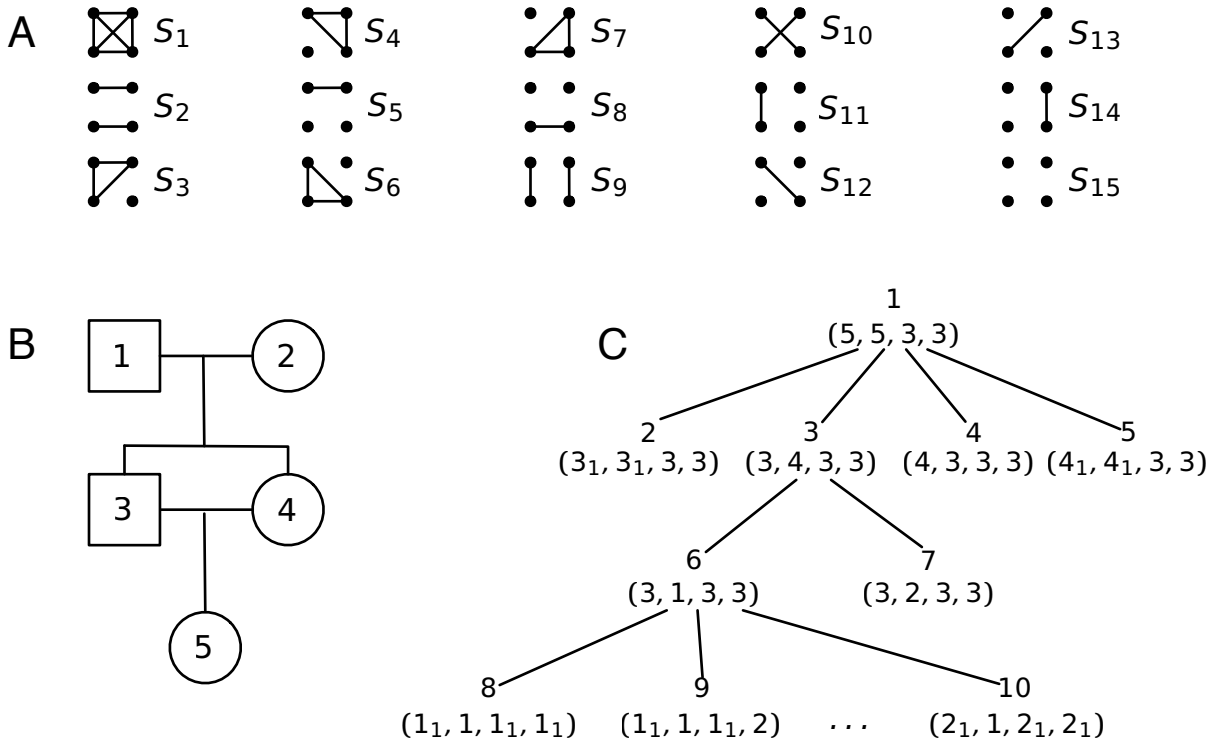

Figure 1: (A) The 15 IGs for four genes. (B) A small pedigree from which one can derive a portion of the KG as shown in (C).

Equivalence with Lange's Recurrence Rules In order to show the equivalence with the recurrence rules as set forth by Weeks and Lange (1988) (WL), we first describe the nodes of the KG. A node v is a set with two vector elements $\{G, D\}$ where $G = (g_1, g_2, \ldots, g_n)$ with each g_i a GID (i.e. it represents a randomly drawn gene from some individual in the pedigree where the draw is done with replacement) and $D = (d_1, d_2, \ldots, d_n)$ with each d_i a label for the corresponding g_i such that $d_i = d_j$ if and only if both g_i and g_j represent a single draw from an individual (which, in turn implies $g_i = g_j$). Note that if $d_i = d_j$ the g_i and g_j are necessarily IBD, but g_i and g_j may be IBD and have $d_i \neq d_j$. For example, a node v may have $G = (5,5,3,3), D = (1,2,3,4)$, meaning two genes are randomly drawn with replacement from individual 5 and two genes are drawn with replacement from individual 3. In this case there are four connected components, two with GID 5. Another node may have $G = (4,4,3,3), D = (1,1,2,3)$, meaning one gene is randomly drawn from individual 4 (occupying the first two elements of G) and two genes are drawn with replacement from individual 3. In this case there are three connected components, one with GID 4, and, because the first two genes are from a single draw, they are necessarily IBD and may or may not be IBD with the two other genes. Both in the remainder of the Supplementary Material and the main text, the elements of D are written as subscripts to the elements of G and non-repeated subscripts are dropped.

To each node v there is an associated random vector X_v whose state space is the set of all possible partitions of the elements of G where genes are in the same partition if and only if they are IBD. Each partition is represented by an IG and the event $X_v = i$ occurs when the genes represented by G in node v have an IBD sharing state represented by the IG S_i. Each node v that is not a terminal node has 2^s child nodes where s is the number of connected components in G that have GID $g^* = \max(g_1, \ldots, g_n)$. As described in the main text, the PMF of X_v

can be written in terms of the PMFs of v's child nodes,

$$p(X_v) = \frac{1}{2^s} \sum_{i=1}^{2^s} p(X_{c_i(v)}), \tag{1}$$

where $c_i(v)$ is the ith child node of v.

To demonstrate the validity of (1) for computing generalized kinship coefficients, we show its equivalence to the WL recurrence rules. In WL the notation $L_{i_1}, \ldots, L_{i_s}$ represents s genes each drawn with replacement from individual i and $(L_{i_1}, L_{i_2}, L_j)(L_k)$ is the event that three genes drawn from individuals i and j, with two genes drawn with replacement from individual i, are IBD and are not IBD with a fourth gene drawn from individual k. This IBD sharing state is represented by IG S_3 in figure 1A. In the notation introduced here this corresponds to a node v with $G = (i,i,j,k), D = (1,2,3,4)$ or in our abbreviated notation $v = (i,i,j,k)$. The probability that the genes are in IBD state S_3 is $p(X_v = 3)$.

The first recurrence rule in WL covers the case where only a single gene is drawn from i,

$$\Pr[(L_i, \ldots)()\ldots()] = \frac{1}{2}\Pr[(L_m, \ldots)()\ldots()] + \frac{1}{2}\Pr[(L_f, \ldots)()\ldots()], \tag{2}$$

where m and f are the mother and father of i, respectively. If we assume that i is the largest GID in the node, then in the language of the kinship graph equation (2) is

$$p(X_{(i,\ldots)} = x_1) = \frac{1}{2}p(X_{(m,\ldots)} = x_1) + \frac{1}{2}p(X_{(f,\ldots)} = x_1), \tag{3}$$

where x_1 is the IG that represents the IBD sharing given by $(L_i, \ldots)()\ldots()$. Equation (3) follows from (1) because the number of connected components with GID i is one.

The second of the WL recurrence relations is

$$\Pr[(L_{i_1}, \ldots, L_{i_s}, \ldots)()\ldots()] = \frac{1}{2^s}\Pr[(L_m, \ldots)()\ldots()] + \frac{1}{2^s}\Pr[(L_f, \ldots)()\ldots()] + [1 - \frac{1}{2^{s-1}}]\Pr[(L_m, L_f, \ldots)()\ldots()]. \tag{4}$$

To write this in the KG framework first note that $(L_{i_1}, \ldots, L_{i_s}, \ldots)()\ldots()$ corresponds to a node $v = (G, D)$ where the first s elements of G have GID i and are all disconnected. The number of child nodes of v is then 2^s. Also, note that the partitioning of $(L_{i_1}, \ldots, L_{i_s}, \ldots)()\ldots()$ corresponds to IGs where genes $L_{i_1}, \ldots, L_{i_s}$ are IBD. In this case, let $x_2 \in \mathscr{I} = \{i \mid S_i \text{ is an IG with vertices } 1, \ldots, s \text{ IBD}\}$. Equation (4) now becomes

$$\begin{aligned} p(X_{(\underbrace{i, \ldots, i}_{s \text{ times}}, \ldots)} = x_2) = \frac{1}{2^s} p(X_{(\underbrace{m_1, \ldots, m_1}_{s \text{ times}}, \ldots)} = x_2) + \frac{1}{2^s} p(X_{(\underbrace{f_1, \ldots, f_1}_{s \text{ times}}, \ldots)} = x_2) + \\ \frac{1}{2^s}\left[p(X_{(\underbrace{m_1, \ldots, m_1}_{s-1 \text{ times}}, f, \ldots)} = x_2) + \cdots + p(X_{(m, \underbrace{f_1, \ldots, f_1}_{s-1 \text{ times}}, \ldots)} = x_2) \right]. \end{aligned} \tag{5}$$

Note that there are $(2^s - 2)$ terms in the square brackets in equation (5), each with at least one m and one f. Because rule 2 in the main text describing the KG construction specifies that every occurrence of m in these terms be connected (i.e. they all correspond to a single draw from m), and similarly for f, symmetry dictates that all these terms are equal (i.e. in each term there is only one draw from m and one draw from f), and, hence, we recover the WL recurrence rule (4).

The WL recurrence rule (4) only applies when the partitioning of the genes is one in which the first s vertices are IBD, that is, when $X_v = x_2$. To cover the cases where the first s vertices are not all IBD, they give a third

recurrence relation,

$$\Pr[(L_{i_1},\ldots,L_{i_r},\ldots)(L_{i_{r+1}},\ldots,L_{i_{r+t}},\ldots)\ldots] = \frac{1}{2^{r+t}}\Pr[(L_m,\ldots)(L_f,\ldots)\ldots] + \frac{1}{2^{r+t}}\Pr[(L_f,\ldots)(L_m,\ldots)\ldots]. \tag{6}$$

Again, translating this into the KG framework, we note that the partitioning $(L_{i_1},\ldots,L_{i_r},\ldots)(L_{i_{r+1}},\ldots,L_{i_{r+t}},\ldots)\ldots$ corresponds to IGs indexed by $\mathscr{J}$ where $\mathscr{J} = \{j \mid S_j$ is an IG where the first r genes with GID i are IBD and the other t genes with GID i are IBD, and the two groups are not IBD with each other$\}$. Let $x_3 \in \mathscr{J}$. By noting that the $r+t$ genes with GID i are disconnected, we obtain

$$\begin{aligned} p(X_{(\underbrace{i,\ldots,i}_{r\text{ times}},\underbrace{i,\ldots,i}_{t\text{ times}},\ldots)} = x_3) \\ &= \frac{1}{2^{r+t}}p(X_{(\underbrace{m_1,\ldots,m_1}_{r\text{ times}},\underbrace{f_2,\ldots,f_2}_{t\text{ times}},\ldots)} = x_3) + \frac{1}{2^{r+t}}p(X_{(\underbrace{f_1,\ldots,f_1}_{r\text{ times}},\underbrace{m_2,\ldots,m_2}_{t\text{ times}},\ldots)} = x_3) + \ldots \\ &= \frac{1}{2^{r+t}}p(X_{(\underbrace{m_1,\ldots,m_1}_{r\text{ times}},\underbrace{f_2,\ldots,f_2}_{t\text{ times}},\ldots)} = x_3) + \frac{1}{2^{r+t}}p(X_{(\underbrace{f_1,\ldots,f_1}_{r\text{ times}},\underbrace{m_2,\ldots,m_2}_{t\text{ times}},\ldots)} = x_3) \end{aligned} \tag{7}$$

Only two terms are retained in (7) because all other nodes on the right hand side are not consistent with the event $X_v = x_3$ because genes with GIDs that are connected are necessarily IBD. In practice, this is enforced by boundary condition 1 where IGs inconsistent with the connectedness of the GIDs are assigned probability zero.

The equivalence of the WL recurrence rules and the KG formulation presented is evident because given a node v all child nodes and the associated PMFs of X can be reduced to those cases above, up to a permutation of the GIDs. This is so because the recursion relations (3), (5), (7) hold for when the number of connected components in G with GID i is s where $1 \le s \le n$. For instance, if i occurs s times and all i are connected, then equation (3) holds. Furthermore, when $s > 1$, X_v must take on values in one of $\mathscr{I}$ or $\mathscr{J}$, up to a permutation of the GIDs. This is because multiple draws from an individual cannot result in those genes falling into more than two distinct IBD groups (assuming diploid individuals).

Step 2 example: Letting 3 and 4 be the father and mother of 5, the child nodes of $(5,5,3,3)$ are $(3_1,3_1,3,3)$, $(3,4,3,3)$, $(4,3,3,3)$, $(4_1,4_1,3,3)$, with the constraints on the first of these child nodes being $p(S_j) = 0$ for $j = 6,\ldots,15$ (see figure 1A for the four-gene IGs S_j). Similarly, the child nodes of $(5_1,5_1,5_2,5_2)$ are $(3_1,3_1,3_1,3_1)$, $(3_1,3_1,4_2,4_2)$, $(4_1,4_1,3_2,3_2)$, $(4_1,4_1,4_1,4_1)$, with the constraints on the second and third of these child nodes being $p(S_j) = 0$ for $j = 3,\ldots,15$. identical subscripts within the same node indicate IBD.

Boundary condition example: As an example of the boundary conditions, consider the case where individuals 1 and 2 are founders. The terminal node t with GIDs $(1_1,1_1,2_2,2_2)$ corresponds to IG S_2, hence $p(X_t = 2) = 1$ by boundary condition 1. On the other hand, boundary condition 2 states that a terminal node with GIDs $(1_1,1_1,1,2)$ has two compatible IGs, S_3 and S_5 and $p(X_t = x) = 0.5$ for $x = 3$ and $x = 5$.

KG construction example: A partial example is shown in figure 1C for the pedigree in figure 1B. In order to find the condensed identity coefficients for the pair of individuals 5 and 3 we need all generalized kinship coefficients for two genes from 5 and two genes from 3 (node 1 in the figure). There are four child nodes (2–5) with the child nodes of node 3 shown. Node 6 has eight child nodes, only three of which are shown. These nodes (8–10) are terminal nodes with boundary condition 2 applying to nodes 8 and 9 and boundary condition 1 applying to node 10. The PMFs at these nodes are: node 8, $p(S_6) = .5$, $p(S_1) = .5$; node 9, $p(S_3) = .5$, $p(S_{11}) = .5$; node 10, $p(S_6) = 1.0$. Summing the probabilities for all terminal nodes according to step 3 above gives the PMF at node 6.