A Mathematical Model for the Ancestor Paradox

by Sophie Pénisson

1. Introduction

Each individual has two parents, hence four grandparents, eight great-grandparents, and so on. It thus seems natural to expect that by going back n generations into the past, the individual has 2^n ancestors at this generation. This number grows very rapidly, however, and for a sufficiently large n might even exceed the number of individuals alive at that time. For instance, 2^{30} (which is more than one billion) exceeds the worldwide population 30 generations ago (a few hundred million). The flaw in this naive reasoning is that one ignores the fact that some ancestors might be counted several times.

When people marry relatives (with whom they share ancestors, by definition), their progeny will have fewer ancestors than this original reasoning suggests. For instance, Ferdinand I of Austria's parents were double first cousins, which implies that he only had 4 great-grandparents instead of 8, 8 great-great-grandparents instead of 16 and so on. Since everyone has some marriages among relatives in their ancestry, their cumulative effect reduces the expected number of ancestors at generation n from 2^n to an unknown number, unique to each individual. The overall shape of an ascendant family tree, if presented with the oldest generations at the top, should therefore not look like an inverted pyramid but rather like a diamond (it should get wider and wider but eventually start to narrow down to few ancestors).

The aim of this paper is to present a mathematical model that helps to study the influence of inbreeding on the number of ancestors in a family tree, hence on its shape. Considering that the reader might not be familiar with the mathematical jargon, this paper focuses on the main ideas and results and does not go into details. For example, we avoid writing any formulas as much as possible. For the more expert reader, however, the endotes provide the appropriate technical terms needed to fully understand the model and its properties. The model output is the ascendant family tree of one individual. Since our goal is to study the influence of inbreeding, the model parameters are linked to the proportions of cousin unions at each generation in the studied population or community. Family trees may vary greatly in size and complexity, even within a single community. This is why we propose a stochastic model. This means that the output is random; for the given parameters, the family tree may be different each time. A deterministic model on the other hand, would consistently provide the same tree. "Random" does not mean "any" family tree. The randomness is controlled by our model, in the sense that some family trees are more likely to be obtained than others. For example, if we choose as parameters large proportions of first cousin unions, the output very likely will be a rather narrow family tree (with few ancestors at each generation), and with small probability, be a very large tree.

2. Description of the model

Several family tree models can be found in the literature. Some are specifically tailored to best describe trees of a given community. For example, Pattison (2000) proposed a deterministic model obtained as an average of many thousands of pedigrees from Britain. Others are more theoretical, such as Wachter's deterministic model (1978), which computes the number of ancestors at one generation according to the surrounding population size at that time and to the number of ancestors in the previous generation. Finally, one can find some models which are mathematically

more challenging and interesting, such as Derrida's stochastic model of a family tree in a population of constant size (2000). However, none of these models allow to study precisely the influence of inbreeding on the shape of the family tree. This is why we present in this paper a model that is stochastic, allows unions among cousins, and is general enough such that it can be applied to different communities. Its parameters can nonetheless be fitted to the studied population, as described in Section 3.

2.1. Graphical representation

Since we are interested in the overall shape of the family tree, we assume for simplicity that there are neither polygamous marriages, remarriages, nor unions among individuals of different generations. It is thus more relevant in our model to consider couples rather than individuals. The following graphical representation of a family tree (Figure 1), which differs from the classical one, enables us to best visualize and describe our stochastic model. Each couple is represented by one node. The younger generations are at the bottom of the graph, and the oldest at the top. The single bottom node thus corresponds to the parents of the studied individual, the two nodes at the next generation correspond to its four grandparents (two couples) and so on. Each oriented edge then represents one individual, male or female. The right bottom edge in Figure 1 corresponds for instance to the individual's mother.

By assumption, unions among relatives in our model can only be among cousins, of any degree. Since cousins share at least one common ancestor couple, the node representing their union is linked via two distinct paths to the common ancestor node. In Figure 1 for example, the individual's maternal grandparents, depicted with a pink node, are second cousins. By definition, they share great-grandparents. This pink node, therefore, is linked via two different paths to a node three generations back. We thus have an easy way to visualize inbreeding in a family tree; a union among cousins of degree k corresponds to a loop of height k+1 rooted in the node. It allows multiple consanguinity degrees as well (double first cousins, cousins of degree two and seven and so on), the various degrees corresponding to the different loops rooted in the node.

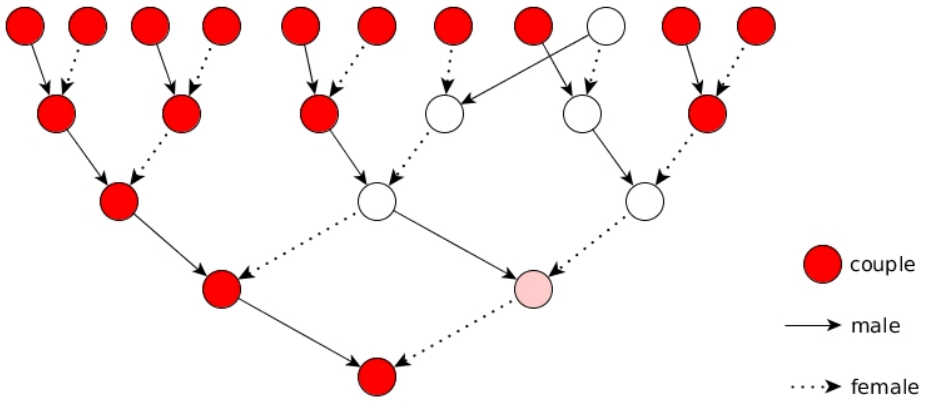


Figure 1. Family tree over five generations, with inbreeding. The pink node represents a second cousin union, and the white nodes the paths to their common great-grandparents.

2.2. Stepwise description of the stochastic model

The generations are enumerated backward in time, *i.e.* the parents are at generation one, the grandparents at generation two, etc. The number of ancestor couples at generation n is denoted C_n, therefore C_1=1, C_2=2, C_3=2, 3 or 4 etc. We call n-node any node from the n^{th} generation (hence there are C_n n-nodes in the tree). We define the random family tree iteratively, meaning that we describe how to build generation n+1 if the tree is known up to generation n. The procedure illustrated in Figure 2 goes as follows.

- **Step 1.** Since there are C_n ancestor couples in the tree at generation n, there can be at most $2C_n$ ancestor couples at generation n+1. However, because of inbreeding, some of these $2C_n$ ancestor couples might be shared by several couples of the n^{th} generation. We model this phenomenon by allowing some of the $2C_n$ n+1-nodes to randomly merge with each other, as described in the following steps.
  In Figure 2, C_3=4. Hence, at most there can be eight couples at generation four ($C_4 \leq 8$).
- **Step 2.** Since unions among siblings are not allowed in our model, two couples from the n+1th generation linked to the same n-node should not merge with each other. For this purpose, we randomly assign two colors (blue and green) to each such pair of n+1-nodes. As a consequence, the $2C_n$ nodes at generation n+1 are separated into two subgroups: one group of C_n blue nodes and one group of C_n green nodes. Only nodes of the same color are allowed to merge with each other, as described in the following step.
  In Figure 2, there are four blue nodes and four green nodes at the fourth generation. In particular, no nodes of the same color are linked to the same node of the third generation, such that no union among siblings is possible.
- **Step 3.** Within each colored subgroup, each pair of nodes has probability p_n to merge (and probability 1-p_n not to merge)[1].
  In Figure 2, each of the six pairs of blue nodes (resp. green nodes) has probability p_3 to merge. As depicted, one blue pair and two green pairs have merged.
- **Step 4.** All the nodes which have been connected through the previous procedure then merge into one single node, which keeps all the corresponding outgoing edges. This implies that the resulting couple has more than one offspring appearing in the family tree. Note that the size C_{n+1} of the n+1th generation and its genealogical connections with the n^{th} generation are purely random[2] since they depend on the random procedure at Step 3.

  In Figure 2, the connected pair of blue nodes merge into one node and only three blue nodes remain. Similarly, the two pairs of connected green nodes merge into one node, and only two green nodes remain. Consequently, five (instead of eight) ancestor couples remain, namely C_4=5. Note that due to the new genealogical connections subsequently created between the third and the fourth generations, the individual's parents are second cousins, and both his maternal and paternal grandparents are first cousins.

1. This random procedure corresponds to an Erdős-Rényi random graph $G(C_n, p_n)$ with C_n nodes and edge probability p_n.
2. Conditionally on C_n, C_{n+1} has the distribution of the sum of two independent random variables identically distributed as the random number of connected components in $G(C_n, p_n)$. The number of outgoing edges for each of the C_{n+1} nodes then correspond to the size of the corresponding connected component.

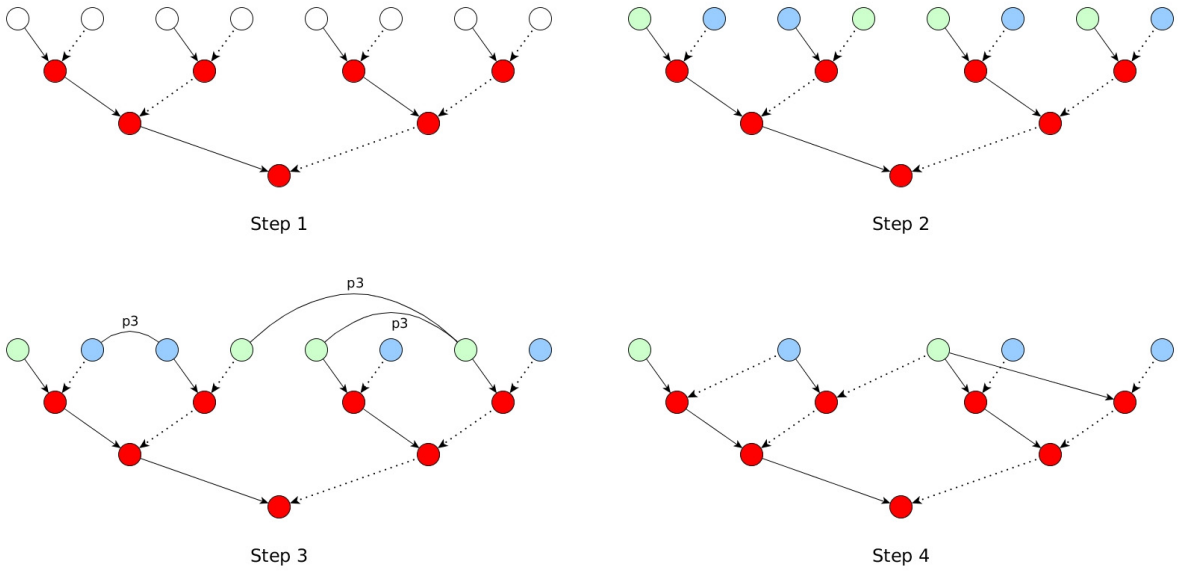
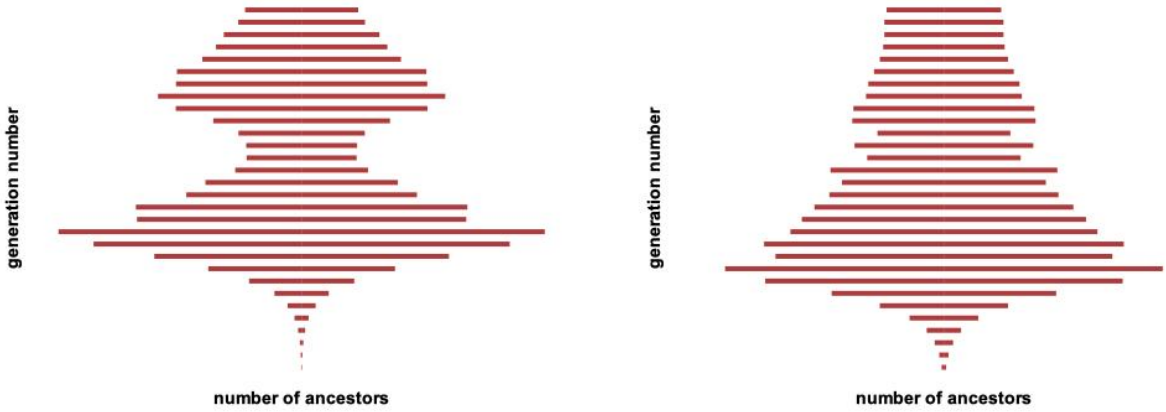


Figure 2. Family tree stochastic model: step by step construction of the fourth generation assuming that the tree is known up to its third generation. The different steps are described in Section 2.2.

3. Pedigree collapse

A question of interest are the conditions on the model parameters $p_2, p_3 \ldots$ to have a pedigree collapse[3], that is, for the family tree to have a diamond shape. We can prove that if there exists some positive $\underline{p}$ such that all the parameters satisfy $p_n \geq \underline{p}$, then each family tree in the stochastic model has a diamond shape[4]. Intuitively, it says that if the merging probabilities are not too small, it prevents the family tree to indefinitely get wider and wider at its top.

Figure 3. Family trees simulated with the stochastic model (see Section 5), satisfying the condition presented in Section 3 to have a pedigree collapse.



3. We say that an infinite family tree collapses if $\limsup_n C_n < +\infty$.
4. If $\liminf_n p_n > 0$, then the family tree almost surely collapses.

4. Fitting of the parameters

The parameters of our stochastic model are p_2, p_3 … where p_n corresponds to the pairwise merging probability at generation n+1, as described in Step 3, Section 2.2. Intuitively, the larger these probabilities are, the more likely it is that many ancestors are shared by several individuals in the family tree, hence the stronger the inbreeding. Nonetheless, we are willing to compare these theoretical parameters to more tractable quantities. In view of the data available in the literature, we choose to link these parameters with the proportions of first and second cousin unions in the studied population or community. Ideally one would need the proportions of unions among cousins of any degree over several generations, but such data is not commonly available.

Let us assume that we study a population or community for which the proportions p_n^{1st} and p_n^{2nd} of first and second cousin unions at each generation n are known. We shall choose the model parameters p_2, p_3 ... such that the probability for a union at generation n to be among first cousins in our model is (approximately) equal to p_n^{1st}. We require the same condition for second cousin unions. The computation of these probabilities is not trivial[5] and will not be detailed here. We shall simply say that the probability for a union at generation n to be among first cousins depends on the probability that some of the grandparental couples at generation n+2 merge, hence depends on everything which happens until generation n+2. Therefore, this probability is a function of the parameters $p_1,\ldots, p_{n+1}$, which we denote $f_n^{1st}(p_2,\ldots, p_{n+1})$. The same reasoning leads to the conclusion that the probability for a union at generation n to be among second cousins is some function $f_n^{2nd}(p_2,\ldots, p_{n+2})$.

If the proportions p_n^{1st} and p_n^{2nd} are known up to some generation N, then we shall choose $p_2,\ldots, p_{N+2}$ such that the cousin union probabilities are close[6] to these known proportions, namely

$$p_n^{1st} \approx f_n^{1st}(p_2,\ldots, p_{n+1})$$
$$p_n^{2nd} \approx f_n^{2nd}(p_2,\ldots, p_{n+2})$$

for each n from 1 to N. Note that if we are solely interested in fitting the model parameters to the first cousin union proportions (or if these are the only proportions available), then only the first condition is considered and it becomes an equality. Namely, there exists a unique sequence $p_2,\ldots, p_{N+1}$ such that for each n from 1 to N,

$$p_n^{1st} = f_n^{1st}(p_2,\ldots, p_{n+1}).$$

5. Simulation of the stochastic model

This stochastic model can be simulated[7], with appropriate software. It means that once the

---

5. For instance, the probability for a union at generation n to be among first cousins is $1-E[(1-R(C_{n+1}, p_{n+1}))^2]$, where $R(c, p)$ is the probability for two nodes in $G(c, p)$ to be connected (for which there is no explicit formula).

6. We use a least squares method.

7. For this purpose one needs to simulate (twice for each generation) the number of connected components in an Erdős-Rényi random graph, say $G(c, p)$. Such a graph can be simulated via its random adjacency matrix A (a symmetric matrix of order c with a null diagonal and upper-triangular entries following a Bernoulli distribution with parameter p), from which we can deduce its degree matrix D (diagonal matrix of order c with $d_{ii}=a_{i1}+...+a_{ic}$), and its Laplacian matrix $L=D-A$. The number of connected components of the graph is then the algebraic multiplicity of the 0 eigenvalue of L. If one wishes to simulate not only the number of ancestors but also the genealogical structure, one needs to consider the Laplacian matrices themselves at each generation.

model parameters $p_2,\ldots, p_n$ are fixed (either arbitrarily, or based on genealogical data as described in Section 4), one can simulate as many family trees as needed over n+1 generations. Since the model is stochastic, each simulated family tree is different. Some trees or types of trees are however more likely to appear, depending on the chosen parameters. Some examples of simulations are presented in Figure 3, Figure 4 and Figure 5.

6. Numerical applications

As mentioned above, the proportions of first and second cousin unions in a given community are rarely available. At most they have been studied for a given period of time, but their evolution over several generations often remains unknown. Nevertheless we found some literature providing these proportions over three generations, namely for the Utah Mormon population (Jorde, 1989) and for a midwestern United States Amish isolate (Hammond and Jackson, 1958). For each of the community we infer from the cousin union proportions the model parameters p_2, p_3, p_4, p_5 according to the procedure described in Section 4. It then enables us to simulate family trees over 6 generations, as explained in Section 5. For each community we shall present one of these simulated trees (Figure 4 and Figure 5).

6.1. Utah Mormon population

The proportions of first and second cousin unions in this community are given in Table 1. From this data we infer the model parameters and obtain

$$(p_2, p_3, p_4, p_5) = (0.0003, 0.0006, 0.0007, 0.0003).$$

This enables us to simulate family trees over six generations of individuals of the Utah Mormon community born in 1915–45. One simulated tree is illustrated in Figure 4.

n	Period	p_n^{1st}	p_n^{2nd}
1	1825–1855	0.17	0.02
2	1855–1885	0.13	0.14
3	1885–1915	0.06	0.26

Table 1. Proportions (in percent) of first and second cousin unions over three generations in the Utah Mormon community (Jorde, 1989).

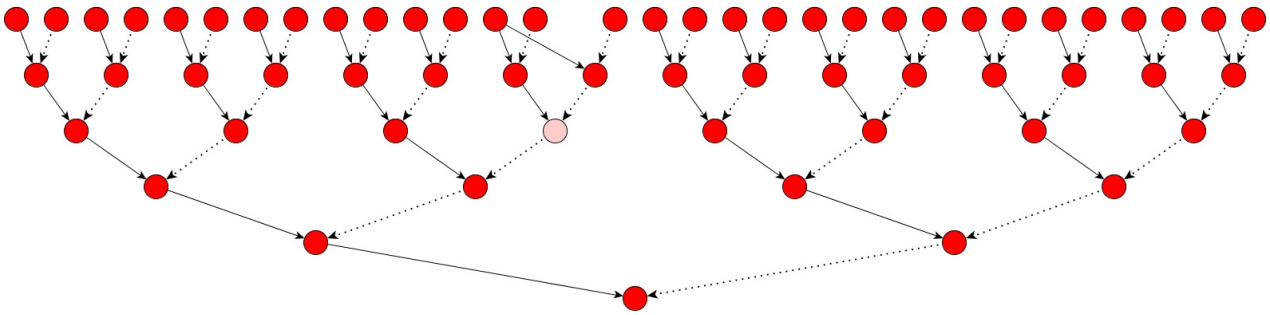


Figure 4. Simulated family tree over six generations of an individual of the Utah Mormon community born in 1915–45. The only union among cousins is depicted by a pink node.

6.2. Midwestern United States Amish isolate

The proportions of first and second cousin unions in this community are given in Table 2. From this data we infer the model parameters and obtain

$$(p_2, p_3, p_4, p_5) = (0.014, 0.015, 0.063, 0.070).$$

Note that these parameters are significantly larger than those for the Mormon community presented in Section 6.1. This is not surprising since inbreeding is much stronger in the Amish isolate (see Table 2). These parameters enable us to simulate family trees over six generations of individuals of the Amish isolate born in 1950–75. One simulated tree is illustrated in Figure 5. The number of cousin unions in this tree (six) is as expected larger than in the tree simulated for a Mormon individual (one).

n	Period	p_n^{1st}	p_n^{2nd}
1	1875–1900	19	7.1
2	1900–1925	3.1	26.6
3	1925–1950	2.9	16.1

Table 2. Proportions (in percent) of first and second cousin unions over three generations in a Midwestern United States Amish isolate (Hammond and Jackson, 1958).

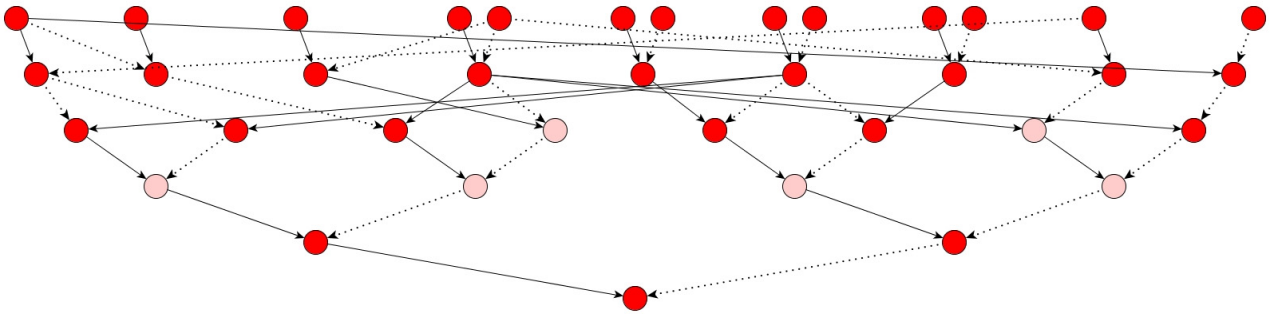


Figure 5. Simulated family tree over six generations of an individual of a Midwestern United States Amish isolate born in 1950–75. The unions among cousins are depicted by pink nodes.

7. Conclusion

The stochastic model presented in this paper is a convenient tool to study the influence of inbreeding (quantified via the proportions of cousin unions) on the overall shape of a family tree. Some more technical results can be obtained, which ought to be published in a mathematics journal. The limit of this model is that it requires specific data which is rarely available, and if so, only over very few generations. However, it provides simulations of family trees, which is a valuable result given that family trees are almost never known in their entirety.

8. References

Derrida B., Manrubia S.C. and Zanette D.H. (2000) On the Genealogy of a Population of Biparental Individuals. *J. Theor. Biol.*

Hammond D.T. and Jackson C.E. (1958) Consanguinity in a midwestern United States isolate. *Am. J. Hum. Genet.*

Jorde L.B. (1989) Inbreeding in the Utah Mormons: an evaluation of estimates based on pedigrees, isonymy, and migration matrices. *Ann. Hum. Genet.*

Pattison J.E. (2000) New method of estimating inbreeding in large semi-isolated populations with application to historic Britain. *HOMO*.

Wachter K.W. (1978) Ancestors at the Norman conquest. *Statistical Studies of Historical Social Structure.*