平成29年度修士論文

有性生物の家系図ネットワークの粗視化

大阪府立大学大学院　工学研究科
電子・数物系専攻　数理工学分野
非線形力学研究グループ

学籍番号 2160103050
伏尾　佳悟

2018年2月

概要

生物の家系図は生物個体を頂点、親子関係を辺と考えることで，ネットワークとみなすことができる。特に、有性生物のそれは数多くのループを含み複雑な構造をとることが知られている。我々は、そのおおまかな特徴をとらえるため、粗視化の手法を提案する。我々の手法は家系図が有向非循環グラフ（DAG）であることを利用した方法であり、家系図からDAG特有の層構造を利用して、家系図の一部を取り出し連結成分に分解しつなぎ合わせる方法である。この手法を2集団に分かれているような家系図モデルに適用したところ、その様子を捉えることが出来た。また、特定の条件下では取り出した家系図内の連結成分のサイズ分布や巨大連結成分の存在を理論的に示すことが出来た。競走馬の家系図に対してもこの手法を提案し、家系図の内部構造の一端を明らかにした。

目次

第 1 章

はじめに

1.1　ネットワークとしての家系図

家系図ネットワークとは、ネットワークの頂点を生物個体とし辺を親子関係として、家系図をネットワークとしてみたものである。特に有性生物の家系図ネットワークはとても複雑な構造をしていることが知られている [1]。家系図ネットワークの複雑さは、ループ構造を持つ点にある。ループ構造を持つことは、容易に知ることが出来る。有性生物の祖先の数を考えてみよう。ある個体の祖先の数は、世代を遡ると、親は 2 人、祖父母は 4 人、と世代を遡るごとに 2 のべき乗で増大する。例えば 40 世代さかのぼった先祖の数は、およそ $2^{40} \approx 10^{12}$ となる。しかし、過去の総個体数は有限である。 1 世代を 25 年として 40 世代前の年代を計算すると、その世代は西暦 1000 年代であり、当時の人口を大きく超えてしまう。このようなパラドックスが生じる原因は、すべての祖先個体が違う個体であるという前提に基づいてこの計算がされていることに由来する。すなわち実際の家系図では違う個体の祖先に同じ個体が含まれており、ループ構造が数多くあるということがわかる。

また家系図ネットワークは有向非循環ネットワーク Directed Acyclic Graph のクラスに属する。子から親方向に辺が向きを持っていると定義しよう。ある個体に着目し、その個体の親、そのまた親と辿っていったとき、最初に着目した個体に戻ってくることはない。なぜなら、親は常に子供より先に生まれているからだ。したがって家系図ネットワークは有向非循環ネットワークである。

1.2　家系図ネットワークに対する研究

家系図のネットワーク構造に関する先行研究としては、まず Derrida らによるものが挙げられる [1]。これは、中立なモデルを用いて特定の個体の先祖をさかのぼったときの先祖個体の重複の程度や寄与の度合を求めたものである。本研究で用いるモデルもこの先行研究のモデルを参考にしており、第 4 章で詳しく紹介する。他に、性比や子数分布を一般化した堀内の研究 [2] や、家系図ネットワーク内の伝搬過程に着目した生田の研究 [3] なども挙げられる。また、本研究の直接の先行研究として、モデルに対して提案手法を適用し評価したもの [4] もある。

1.3　本研究の目的及び結果

生物には種分化という現象がある。これは、ある生物種が 2 つの生物種に分かれることを意味している。一例としてある時点まで一つだった生物集団が二つの小集団に分割されたとしよう。そしてやがてその二つの小集団が、生息環境、餌の種類、クチバシの形状、生殖器の形の外的な特徴—すなわち表現型—などの違いで区別できるようになったとする。このことを表現したのが分岐図である。その一例を図 1.1 に示した。この図は Darwin が「種の起源」の中で使った図である [5]。Darwin のそれに限らず、分岐図は数多くあるが、それらは個体の遺伝情報や地理的要因・歴史的要因にもとづいて描かれている場合が多い。しかし、種は生物個体の集団であり、種のつながりや分岐は個体同士の親子関係の総体としてとらえることもまた可能なはずである。すなわち、分岐図をその構成する個体が見えるくらいまで拡大するとそこには家系図が見えてこなければならない。そして、種分化は集団間の遺伝的な交流の途絶と密接に関連しているので、なんらかの形で家系図の構造に反映されていると考えられる。

ここで、種分化という巨視的な現象は家系図という微視的な立場からどう捉えられるのだろうかという問題が提起される。そして、この問題を解決すれば、たとえ生息域や表現型の違いをしらなくても、家系図の構造だけから種分化が起きていることを示すことができるのではないだろうか？

前節で述べたような状況を背景に、我々は次のような目的を立てた。すなわち、生物個体の親子関係のみに着目し、複雑な家系図ネットワークを粗視化

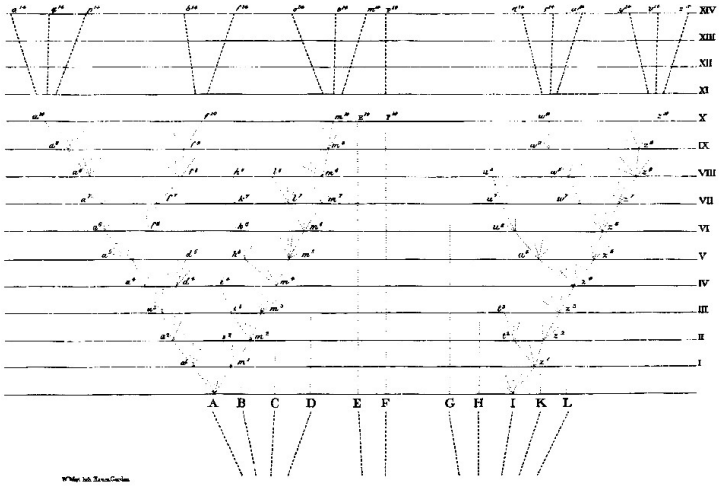


図 1.1. Darwin による生物種の模式的な分岐図 (文献 [5] より改変)。

し、家系図のマクロな構造である分岐図を抽出することである。具体的に行うのは以下の 3 点である。

(i) 粗視化を適用するための家系図を人工的に作成する。

(ii) 実際の生物集団として競走馬を採用し、家系図を再構成し、その特徴について議論する。

(iii) 粗視化の方法を提案し、実際に適用する。

(i) では、生物集団のモデルを用いて親子関係のネットワークを構築する。我々が採用したモデルは、Derrida らによる先行研究で導入された中立的なモデルを拡張したもの（以下、拡張 Derrida モデル）で、家系図ネットワークにある種の分岐構造を自由に埋め込むことができる。また、一夫一妻制や乱婚制など異なる交配条件を採用することができる。つまり、どのような構造を持っているかをあらかじめ把握した生物集団の家系図を作成することができるのである。また、(ii) では実際の生物集団のデータを取得し、家系図を再構成する。いくつかの統計量（子数分布や親子年齢差分布等）を計算し、競走馬の家系図の特徴について議論する。(iii) では、まず拡張 Derrida モデルに対しては、親子関係の総体だけから、その生物集団の交配条件や分岐構造を抽出することをめざす。我々が提案する粗視化の手続きは単純ながら、家系図の分岐構造や交配条件を抽出することができた。また、得られた結果に対する理論的な考察も行う。その次に競走馬の家系図に対しては、計算した統計量から粗視化パラメータを見積もり粗視化を行った。その結果として、時代の変化に伴う大規模な構造を抽出することができた。

本論文の構成は以下の通りである。まず次章、第 2 章で先行研究 [1] で導入されたモデルについて解説する。第 3 章では、我々の提案手法について解説する。第 4 章で Derrida らの提案した中立モデルを拡張し、提案手法を適用する。得られた結果を考察する。第 5 章では、実際の生物集団として競走馬に注目し、その統計的特徴に言及したあと実際に提案手法を適用する。最後に第 6 節で得られた結果及び今後の課題をまとめる。

第 2 章

Derrida らの先行研究

この章では、Derrida らの先行研究 [1] について解説する。本研究で用いた数理モデルは Derrida らが提案した中立モデル及びそれを拡張したものである。まずは、Derrida モデルの紹介をしよう。

Derrida モデルは以下の 5 つの特徴がある。

1. 有性生殖：生物個体には母親と父親がいる。
2. 閉じた生物集団：今考えている集団以外に父親や母親、または子供が存在することはない。
3. 雌雄対称：雄と雌は個体数がそれぞれ同数であり、子数分布も同じである。
4. 世代重複がない：時間に対して世代が離散化されている。すなわち、生物個体の親はその個体の 1 つ上の世代に存在し、子供はその個体の 1 つ下の世代にしか存在しないということである。
5. ランダム交配：子供の親はランダムな組み合わせである。

このモデルは次のような方法で実装された。まず、G は世代番号、N_G は G 世代における個体数、m は成長率をあらわす。G は 0 以上の整数、N_G は自然数、m は正の実数である。

a 最過去世代 $G = 0$ とその世代の個体数 N_0 体の生物個体を作成し、個体ごとに男か女をそれぞれ確率 0.5 で決める。この世代を親世代とする。

b 親世代雄から 1 個体、雌から 1 個体ランダムに選択し、子供を作成する。式 (2.1) に従う次の世代の人数に達するまでこの操作を繰り返す。

c 子世代の各個体の性別を確率 0.5 で決める。

d 子世代をあらたな親世代とみなし、世代をひとつ増やす。

e 親の世代が G_{max} になるまで b〜d を繰り返す。

$$N_{G+1} = \frac{m}{2} N_G. \tag{2.1}$$

このように実装した Derrida モデルでは、ランダムに父親と母親が選ばれるので、乱婚制と考えられる。乱婚制であれば仔数分布－すなわち、ある個体が k 人の子を持つ確率分布 $p(k)$ は個体数が多い極限で平均 m のポアソン分布

$$p(k) = \frac{m^k e^{-m}}{k!}, \tag{2.2}$$

になる。また、乱婚制では兄弟姉妹の両親が同じになる確率は低い。そして、考えている個体集団は小集団にわかれるようなこともなく、一様な構造をしている。

第 3 章

粗視化の手法

この章では、我々の提案する粗視化の手法について解説する。まず次の言葉及び変数を定義する。“窓”は粗視化する世代の範囲を示す。窓の幅は Δ、窓の番号は n、窓同士の間隔は b とする。Δ, n, b はいずれも自然数である。粗視化の手順を次のように定義する。

[手順 1] 窓によって区切られた世代 $n \times b \leq G < n \times b + \Delta$ までを n 番目の窓とし、その中の家系図を連結成分に分解する。それぞれの連結成分を粗視化された頂点 V_k とする。k は各頂点に与えられた頂点番号であり、以下のように決める。頂点を構成する個体の中で、親世代の個体の中の個体番号がもっとも小さいものをその連結成分の代表個体とする。窓の中の全ての頂点に対して、それぞれの頂点の代表個体の個体番号の小さい順に 1 から自然数 k をふる。最後に窓の中の全ての個体を、所属する頂点の頂点番号の小さい順に並べなおす。このときの個体の順番をメソ番号という。

[手順 2] 窓の範囲を b だけずらし同様の手続きを行う。

[手順 3] 異なる窓で粗視化された頂点 V 同士をつなぐ。

[手順 3] でつなぎ方は二通りある。

A. 窓の範囲をずらす b を窓の幅 Δ より小さく設定することで、異なる窓で、違う連結成分でありながら、同じ個体が含まれているようにする。異なる連結成分同士で同じ個体が含まれているとき、その連結成分がつながっているとみなす。

B. 親子関係をもとにしてつなぐ。ある窓のある連結成分に所属する個体には親が存在し、その親が所属している連結成分とつながっているとみなす。

以上2通りである。

方法Aでは、bの大きさはΔ未満という制約が出てくる。また、このつなぎ方では親子関係は無視されており、我々の本来の目的—親子関係のみの情報をもとにして粗視化する—という点から外れている。さらに、同じ個体であるにも関わらず、違う連結成分に同時に存在するという状況は、同じものを分割して二重に表示しているといってよく、好ましくないと考えられる。方法Bでは本来の目的にも合致し、かつbに制約はほとんどないといえる。本研究では方法Bを用いることにした。粗視化によって得られた新しいネットワークを及び頂点、辺と元のネットワークの頂点及び辺との関係を表3.1に示した。粗視化する前は祖先方向の度数が2と決まっており、子孫方向の度数は子数であるが、粗視化されたネットワークでは自明ではなく、祖先方向の度数は2以上も可能である。辺は親子関係である点は変わっていないが、親子関係の集合となる。

ネットワーク	頂点	辺
元の家系図 ネットワーク	表すもの：生物個体 祖先方向の度数：2 子孫方向の度数：子数	表すもの：親子関係 有向
粗視化された ネットワーク	名前：メソ頂点 表すもの：生物個体群 祖先方向の度数：1以上 子孫方向の度数：メソ頂点内の個体の子数の合計以下	名前：メソ辺 表すもの：血縁関係 有向

表 3.1: 粗視化する前の家系図ネットワークと粗視化されたネットワーク間の頂点と辺の関係

粗視化という手法は第1章で説明した外的な特徴で二つの小集団に分かれている場合だけでなく変化がもっと緩やかな状態、すなわち外的な違いから分割プロセスを見るのは難しい時であっても、親子関係から抽出することが出来ると考えられる。さらに我々は、この粗視化に対して、辺の太さwと連結成分の絶滅を導入した。wは異なる連結成分間に親子関係によってつながれている個体数である。また、窓内のある連結成分が絶滅しているとは、その連結成分に含まれる全ての個体に子孫方向に窓外にでる親子関係の辺が存在していないことを意味している。

次章では、Derridaモデルを出発点として、一つの生物集団が世代を経るにつれて人工的に小集団に分割されるようなモデルへと拡張する。そして、本章で導入した粗視化の手法を用いて、実際に親子関係だけから小集団への分

割プロセスが抽出できることを示す。

第 4 章

モデルによる解析

この章では、前章で説明した粗視化の手法を Derrida らによって導入された Derrida モデルに適用した結果を紹介する。また粗視化した後の連結成分のサイズ分布についても考察する。

4.1 Derrida らによるモデル

Derrida らのモデルの家系図を図 4.1 に示す。この図では、グラフ上に生物個体をプロットし親子関係を持つ個体同士を辺で結んだ。辺は複雑に絡み合い、構造を把握することは困難である。

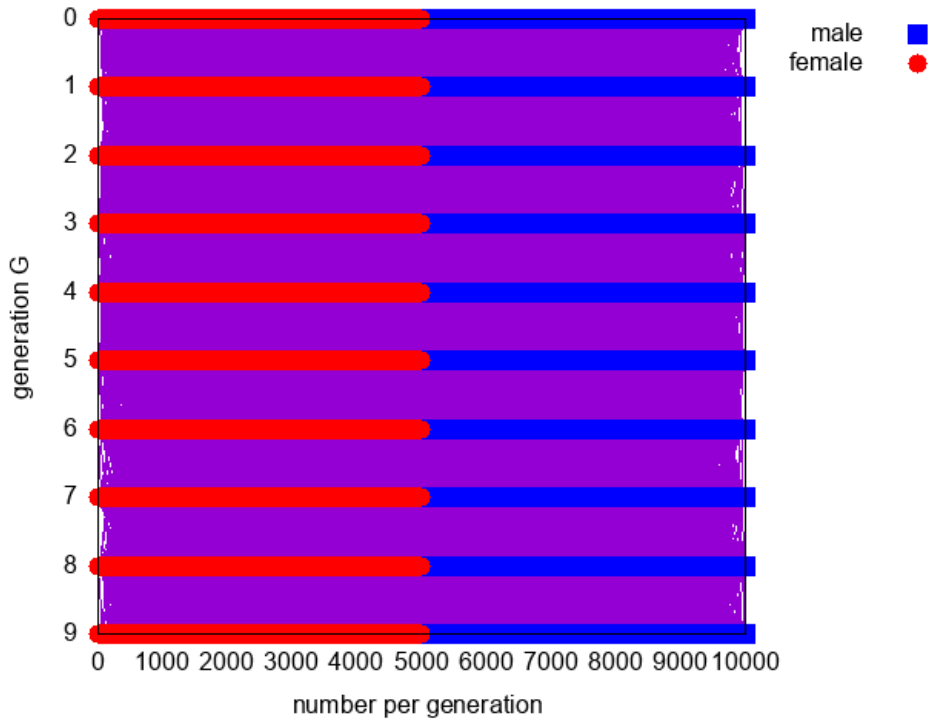


図 4.1. Derrida モデルによる家系図。横軸が世代あたりの個体数で、縦軸が世代を示す。●が雌個体を、■が雄個体を示している。線が親子関係を示している。各世代の個体数は 10,000 であり、各世代の辺の本数は 20,000 となる。辺は複雑に絡み合い構造を把握することは困難である。

Derrida らのモデルに対して、$\Delta = 2$、$b = 2$ で粗視化を行った。その結果を図 4.2 に示した。メソ頂点はそのサイズに比例した幅の線分で表しており、全てのメソ辺は同じ太さで描いている。粗視化をすることにより、各窓内に窓内の個体数に対して 8 割程度の大きさを持つ巨大連結成分が現れた。また、巨大連結成分以外の連結成分はサイズが 1 や 3 程度の小さいものであった。巨大連結成分の存在については、4.3 節にて解説する。

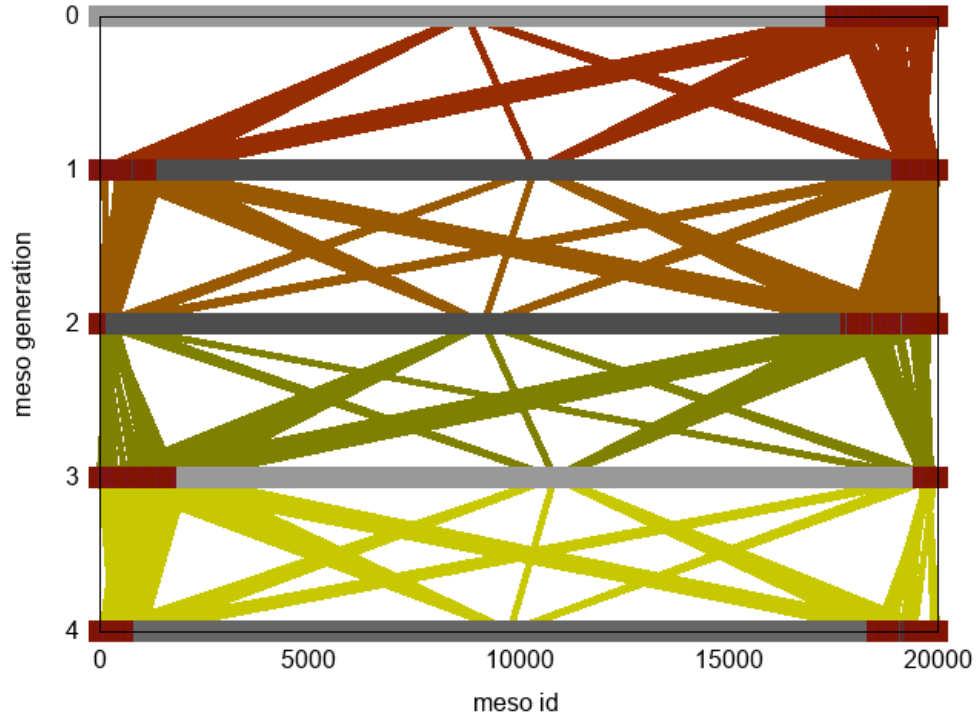


図 4.2. Derrida モデルによる家系図を粗視化した図。図 4.1 のデータを粗視化したもの。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。各世代の個体数は $10,000$ であるので、各メソ世代は $20,000$ 個体が含まれている。全体の 8 割程度の巨大連結成分ができていることが確認できるが、辺同士は複雑に絡み合っている。

図 4.2 でも辺が複雑に絡み合い、親子関係を把握することは困難である。そこで、窓内の親世代で子孫を残していない、すなわちサイズが 1 となる連結成分は、微細構造であるとみなして表示しないことにした。さらに、窓内の個体数の 1 割以上が含まれる連結成分間の辺は太くして強調し、それ以外の辺は点線でしめした。それが図 4.3 である。これを見ると巨大連結成分が太い幹で繋がっている様子が確認できる。また、小さい連結成分にも分かれていることがわかる。

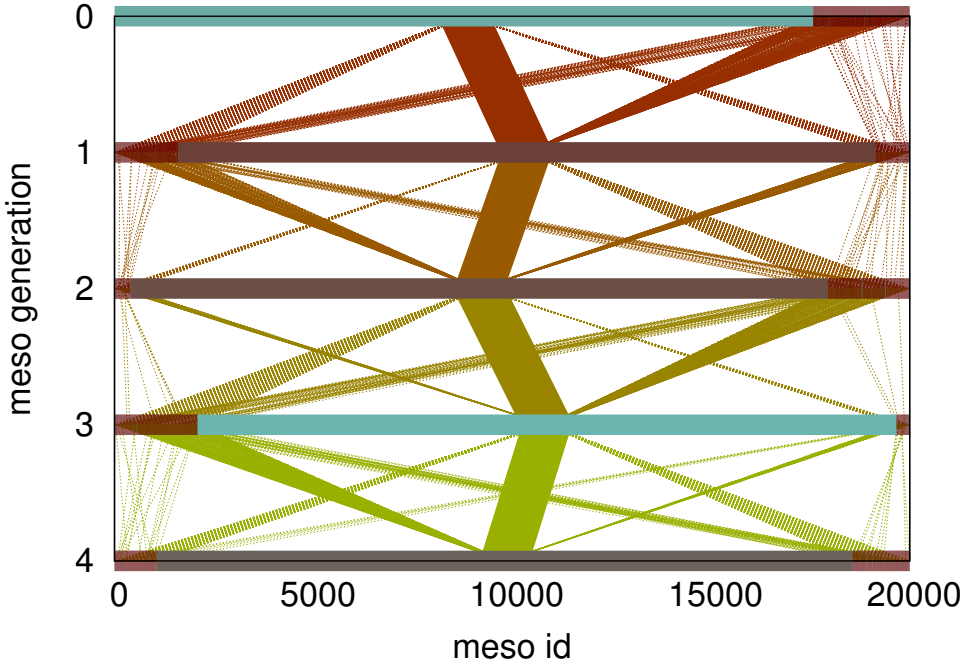


図 4.3. Derrida モデルによる家系図を粗視化した図の改良版。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。各世代の個体数は $10,000$ であるので、各メソ世代は $20,000$ 個体が含まれている。連結成分のサイズ 1 とつながる辺を書かず、かつ窓内の個体数の 1 割以上が含まれる連結成分間の辺は太くし、それ以外の辺は点線で示している。

4.2　拡張した Derrida モデル

本節では、Derrida モデルを出発点とし、一つの生物集団が世代を経るに従って二つ以上の小集団に人工的に分割されるように拡張する。そして拡張されたモデルによって形作られた家系図に提案手法を適用することで、二つの集団への分割過程を抽出する。また、雌雄の交配条件として一夫一妻制と乱婚制を組み入れる。

拡張 1 交配条件：親の交配条件として以下の 2 種類を考える。

(i) 乱婚制： ランダムに選択した親から子供を産む。

(ii) 一夫一妻制： 全ての兄弟姉妹は同じ両親から生まれる。

拡張 2 個性と相性関数：各個体に内部変数“個性”$0 < p < 1$ を与え、親同士の個性の関数として相性関数 $A(p_{\mathrm{man}}, p_{\mathrm{woman}})$ を導入し、子供の生成確率を決める。子供の個性はどちらかの親の値を引き継ぐ。

拡張 1 の (i) は Derrida らが提案した中立モデルと変わらない。ランダムに選ばれた雌雄の組み合わせに対して子供が作られる。一夫一妻制の場合、全

ての兄弟姉妹は同じ両親から生まれる。ヒトに近い交配条件である。

拡張2は任意の世代で集団を複数の集団に分けることができる。例えば、図4.4の網掛け同士または塗りつぶし同士の部分でしか子供が作られないようにすると、同図（左）は$A(p_{\mathrm{man}}, p_{\mathrm{woman}}) = 1$であり、全ての組み合わせで子供が作れるが、同図（右）のように、$A(p_{\mathrm{man}}, p_{\mathrm{woman}}) = C(p_{\mathrm{man}}, p_{\mathrm{woman}})$とすると、個性が0.5より大きいもの同士、または0.5より小さいもの同士でしか子供を作ることが出来ない。ここでCは式(4.1)で表される。同図（右）にしたがって家系図を生成すれば、その家系図は2集団にわかれる。

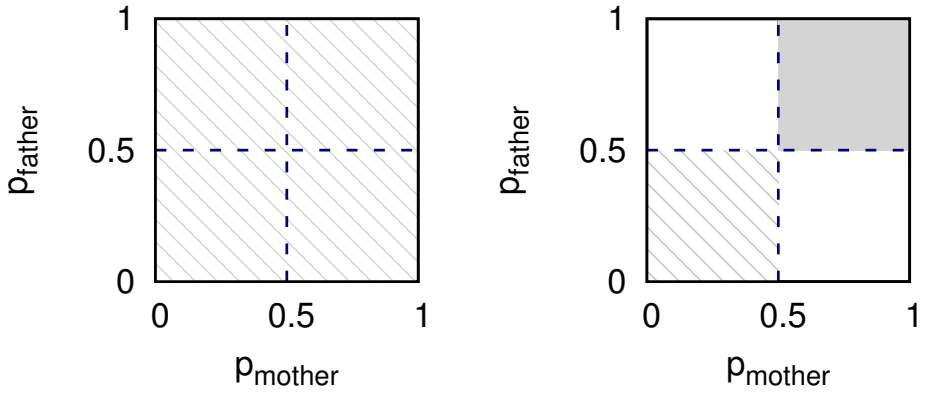


図 4.4. 相性関数 $A(p_{\mathrm{man}}, p_{\mathrm{woman}})$ の例。網掛け同士または塗りつぶし同士の部分で成功する。両図とも、横軸は母親の個性。縦軸は父親の個性。（左）$A(p_{\mathrm{man}}, p_{\mathrm{woman}}) = 1$、（右）$A(p_{\mathrm{man}}, p_{\mathrm{woman}}) = C(p_{\mathrm{man}}, p_{\mathrm{woman}})$。

$$C(p_{\mathrm{man}}, p_{\mathrm{woman}}) = \begin{cases} 1 & \text{if } (p_{\mathrm{man}} - 0.5)(p_{\mathrm{woman}} - 0.5) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

4.2.1 乱婚制

まず個体の交配制度が乱婚制の場合を考えよう。乱婚制は以下のように実装した。世代番号はG、その世代に属する個体数はN_Gとする。

A 最過去世代番号は$G = 0$とし、N_0個体を作成し、個体ごとに雌か雄をそれぞれ確率0.5で決める。個体ごとに個性pを$0 \leq p < 1$の乱数で定める。この世代を親世代P、世代番号をG_Pとする。

B 子供世代 C は世代番号 G_C を $G_C = G_P + 1$ とし、個体数 $N_{G_C} = (m_{G_C}/2)N_{G_p}$ とし、その世代における成長率 $m_{G_P} = 2N_{G_C}/N_{G_P}$ とする。（成長率は平均出生率と言い換えてもよい。）

C 親世代の雄から 1 個体、親世代の雌から 1 個体それぞれランダムに取り出すことにより親の組み合わせを決める。親同士の個性から、交配が成功するかどうかをその世代の相性関数で判別する。交配が成功したら子供を作成する。失敗すればまたランダムに親を選択する。成長率によって得られた子供の数だけ子供を作成する。子供の個性 p は雌親か雄親のどちらかの個性がランダムに選ばれる。

D 子世代の各個体の性別を確率 0.5 で決める。

E 子世代をあらたな親世代とみなし、世代をひとつ増やす。子の世代が G_{max} になるまで B〜E を繰り返す。

この手順により、乱婚制の拡張 Derrida モデルを作成した。$G_{max} = 9, N_0 = 10000, m = 2$ であり、$0 \leq G < 5$ で図 4.4（左）の交配条件、$5 \leq G < 10$ で同図（右）の交配条件にした乱婚制の拡張 Derrida モデルの家系図を作成した。それを図 4.5 に示した。2 集団に分かれることを強調するため、横軸は個性にしている。$G \geq 5$ で親子関係を表す辺が横軸 0.5 をまたいでいないので、2 集団に分かれていることがわかる。

次に、得られた家系図を $\Delta = 2, b = 2$ で粗視化した。その結果を図 4.6 に示した。図 4.3 と同様に、サイズが 1 の連結成分と結ばれる辺は書かず、かつ窓内の個体数の 1 割以上が含まれる連結成分間の辺は太くし、それ以外の辺は点線で示している。これにより、メソ世代が 2 以上、元の家系図での世代 4 以上では、2 集団に分かれていることが確認できる。個性という内部自由度を知ることなしに、提案手法で 2 集団を抽出できた。

4.2.2 一夫一妻制

本小節では、交配制度が一夫一妻制の場合について述べる。一夫一妻制は以下のように実装した。乱婚制の時と同様に世代番号は G、その世代に属する個体数は N_G とする。

Aa 乱婚制 A と同様である。

Bb 乱婚制 B と同様である。

Cc 雌のパートナーをランダムに選び、相性関数から交配が成功するか調べ

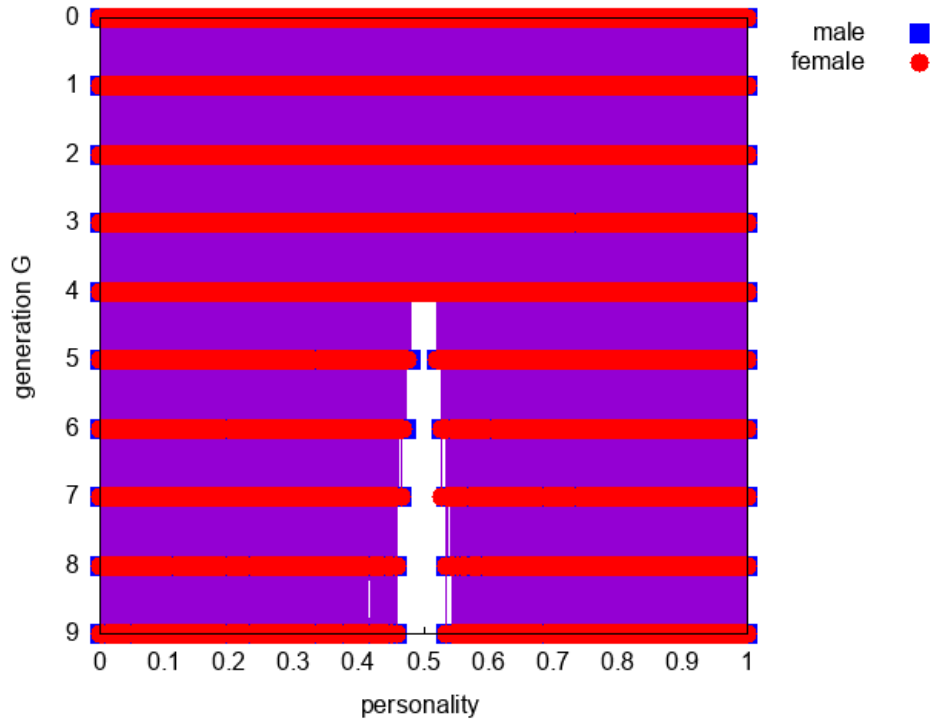


図 4.5. 乱婚制の拡張 Derrida モデルによる家系図。横軸が個性で、縦軸は世代を示す。辺は親子関係を示している。●が雌個体を、■が雄個体を示している。辺が親子関係を示している。$0 \leq G < 5$ で $C(p_{mother}, p_{father}) = 1$、$5 \leq G < 10$ で式（4.1）に従っている。

る。交配が成功する場合、パートナー成立とする。ただし、パートナーをランダムに選ぶ回数がその窓のサイズを超えた場合は、その雌個体にはパートナーがいないとする。全ての雌のパートナーを決める。

Dd パートナーとなった一組の雌雄とランダムに選ぶ。（雌個体をランダムに選ぶと、雄個体は手順 Cc で決めたものになる。）子供を 1 個体作成し、性別を確率 0.5 で決める。子供世代 C の個体数 N_{G_C} になるまで、この手順を繰り返す。子供の個性 p は雌親か雄親のどちらかが選ばれる。

Ee 子世代の各個体の性別を確率 0.5 で決める。

Ff 子世代をあらたな親世代とみなし、世代をひとつ増やす。

Gg 親の世代が G_{max} になるまで Bb〜Ff を繰り返す。

この手順により、一夫一妻制の拡張 Derrida モデルを作成した。手順 Cc と Dd であるが、相性関数ありきで雌雄の組を決めている。図 4.4 の右図のような相性関数で、個性が 0.5 未満の個体数が全体の個体数の半分の時に、相性関数を考えず雌雄の組を決めてしまうと、実に 1/2 の組が子供を持てない組となる。そうなると、子供を作るために残りの組が多くの子供を持つことにな

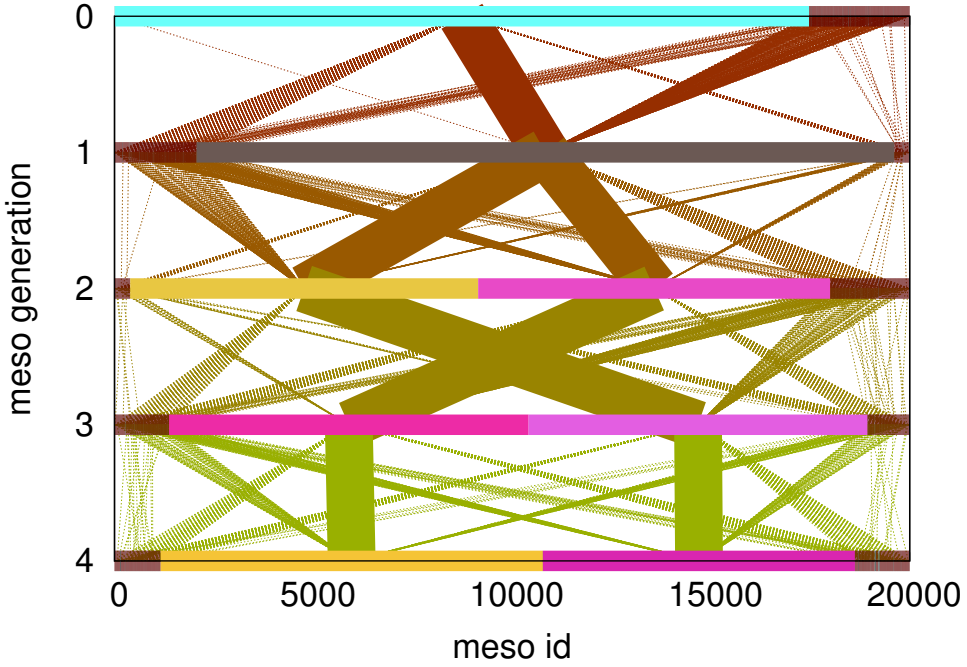


図 4.6. 乱婚制の拡張 Derrida モデルによる家系図を粗視化した図。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。各世代の個体数は $10,000$ であるので、各メソ世代は $20,000$ 個体が含まれている。サイズが 1 の連結成分と結ばれる辺は書かず、かつ窓内の個体数の 1 割以上が含まれる連結成分間の辺は太くし、それ以外の辺は点線で示している。

る。これでは、子供を持つ組に属する個体の子数分布は $\lambda = 4$ のポアソン分布になる。個体の子数分布をできるだけ変えない為に、本研究ではこのような手順にしている。

$G_{max} = 18, N_0 = 10000, m = 2$ であり、$0 \leq G < 9$ で図 4.4（左）の交配条件、$9 \leq G < 18$ で同図（右）の交配条件にした一夫一妻制の拡張 Derrida モデルの家系図を作成した。それを図 4.7 に示した。2 集団に分かれることを強調するため、横軸は個性にしている。$G \geq 9$ で親子関係を表す辺が横軸 0.5 をまたいでいないので、2 集団に分かれていることがわかる。

次に、得られた家系図を $\Delta = 2, b = 2$ で粗視化した。その結果を図 4.8 に示した。交配条件の性質上、常に同じ親が選択されるので $\Delta = 2$ の場合では、粗視化された頂点 V は子供とその親だけの大きさにとどまる。乱婚の場合と異なり、巨大連結成分は存在しない。

しかし、 $\Delta = 3, b = 3$ とすると、状況は一変する。それを図 4.9 に示した。図 4.3 のように、サイズが 4 の連結成分と結ばれる辺は書かず、かつ窓内の個体数の 1 割以上が含まれる連結成分間の辺は太くし、それ以外の辺は点線で

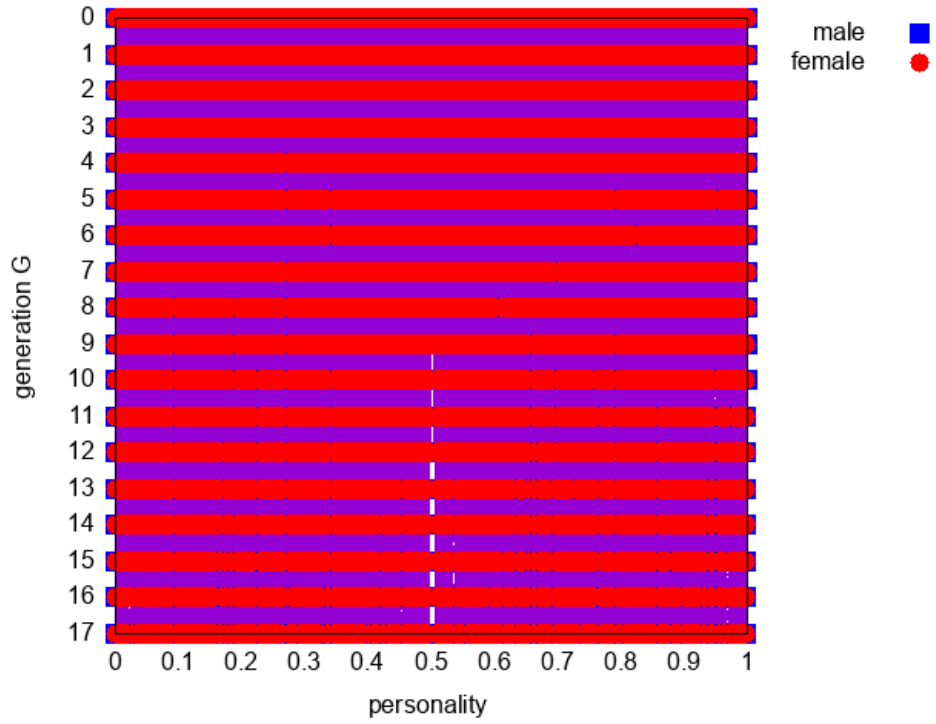


図 4.7. 一夫一妻制の拡張 Derrida モデルによる家系図。横軸が個性で、縦軸は世代を示す。辺は親子関係を示している。●が雌個体を、■が雄個体を示している。辺が親子関係を示している。$0 \leq G < 9$ で $C(p_{mother}, p_{father}) = 1$、$9 \leq G < 18$ で式（4.1）に従っている。

示している。書かないサイズの大きさを 4 以下にしたのは、$\Delta = 3$ の窓内にあるサイズが 4 以下の連結成分は孫世代（子世代に対する子世代）に個体がいないことを示している。絶滅がはやく、家系図内の微細構造であると判断した。これにより、メソ世代が 3 以上、元の家系図での世代 9 以上では、2 集団に分かれていることが確認できる。交配条件が乱婚制でなくても粗視化パラメータを変えることにより個性という内部自由度を知ることなしに、提案手法で 2 集団を抽出できた。

4.3　連結成分のサイズ分布

本節では、連結成分のサイズ分布について、理論とモデルの両方から述べる [6]。ランダムネットワークについては Newman のレビュー [7] が詳しい。連結成分のサイズとは、その連結成分に含まれる個体数を意味する。まず結論から述べると、我々は以下の二点を理論的に明らかにした。乱婚制の Derrida モデルにおいて、$\Delta = 2$ で 1 世代の個体数 N が無限に大きい時、巨大連結成分が存在することと、連結成分のサイズ分布を子数分布から導くことが出来

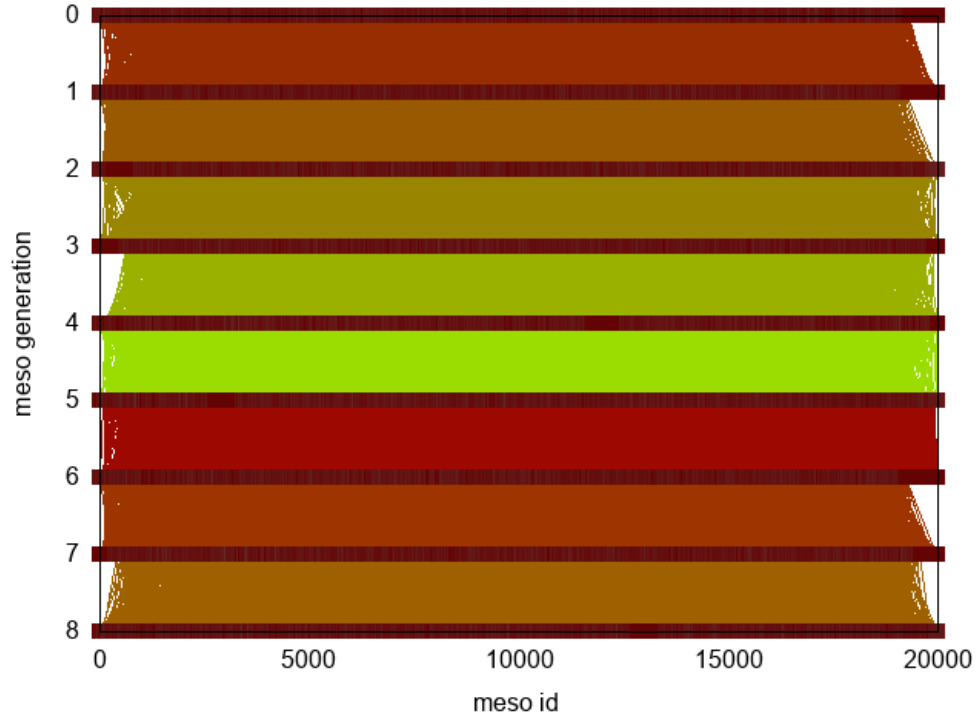


図 4.8. 一夫一妻制の拡張 Derrida モデルによる家系図を粗視化した図。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。各世代の個体数は $10,000$ であるので、各メソ世代は $20,000$ 個体が含まれている。

ることである。なお本節内では Δ は常に 2 である。

4.3.1 小さなサイズの連結成分

本小節では、親世代の個体数に対する $\Delta = 2$ で窓に切り出したときに得られる小さな連結成分に所属する親世代の個体数の割合を理論的に計算していく。ここで、子数分布 p_k すなわち子どもを k 個体もつ個体数の、その世代の個体数 N に対する割合を p_k であるとする。図 4.10 に $\Delta = 2$ で Derrida モデルから窓を切り出したときに、得られる連結成分の例を小さい順に列挙した。

p_k がわかっている時、小さなサイズの連結成分を理論的に求めることができる。そして、それを用いて巨大連結成分のサイズも近似的に求めることができる。ここでの割合とは、親世代の中でサイズ s の連結成分に含まれる個体数の割合 $q(s)$ を意味する。実際に図 4.10 で示した 4 つの例に対して、その割合を求めていこう。

(1) この連結成分の割合は容易に導出することができる。窓内の親世代における、子供がいない個体の割合がそのまま、この連結成分の割合になるからだ。すなわち $q(1) = p_0$。

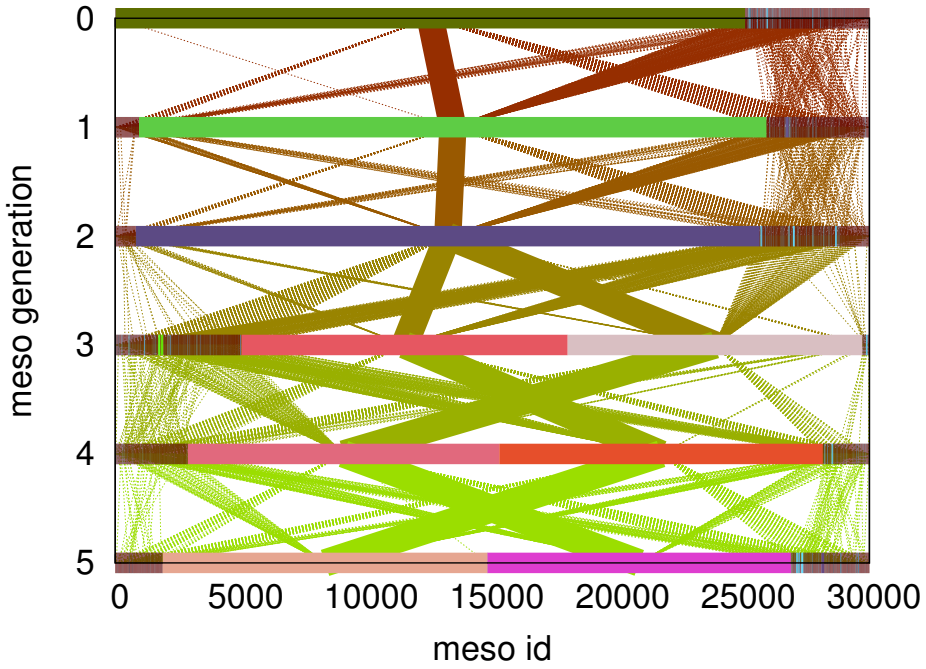


図 4.9. 一夫一妻制の拡張 Derrida モデルによる家系図を粗視化した図。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。各世代の個体数は $10,000$ であるので、各メソ世代は $30,000$ 個体が含まれている。サイズが 4 以下の連結成分と結ばれる辺は書かず、かつ窓内の個体数の 1 割以上が含まれる連結成分間の辺は太くし、それ以外の辺は点線で示している。

(2) 雌雄対称であることを考えれば、雄親の中に含まれるサイズ 3 の連結成分の割合 $q(3)^{male} = q(3)$ となる。ここでは $q(3)^{male}$ を求める。雄親の中から、子供の数が 1 である個体を選び、その子供の母親の子供の数が 1 であればよい。ここで、雄親、雌親はそれぞれ $N/2$ 個体存在し、雄親、雌親からはそれぞれ子世代に N 本の辺が延びていることに注意する。すなわち、

$$q(3) = q(3)^{male} = \frac{\frac{N}{2}p_1}{\frac{N}{2}} \times \frac{\frac{N}{2}p_1}{N} = \frac{{p_1}^2}{2} \tag{4.2}$$

(3) この連結成分も雌雄対称であることから、$q(4) = q(4)^{male}$ となることに注意する。$q(4)^{male}$ を求める。雄親の中から、子供の数が 2 である個体を選び、その個体の母親が同じ個体であればよい。すなわち、

$$q(4) = q(4)^{male} = \frac{\frac{N}{2}p_2}{\frac{N}{2}} \times \frac{2\frac{N}{2}p_2}{N} \times \frac{1}{N-1} = \frac{{p_2}^2}{N-1} \tag{4.3}$$

上式から、$q(4)$ がその世代の数 N に依存することがわかる。よって、N が無限に大きいとき、$p(4) = 0$ となる。

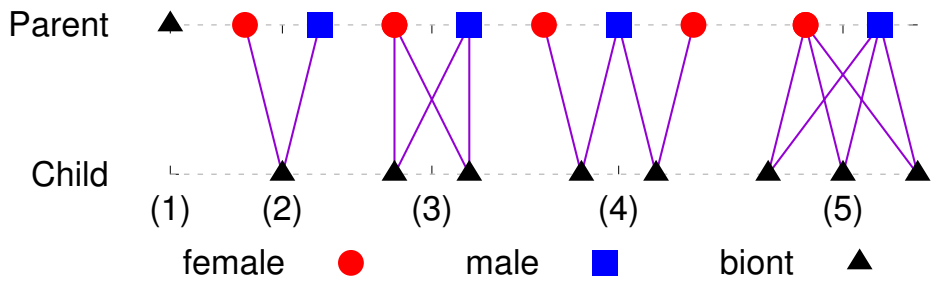


図 4.10. Derrida モデルで、$\Delta = 2$ の窓で切り出した時、得られるいくつかの連結成分を描いた。●は雌個体を、■は雄個体を、▲は雌雄どちらかの生物個体を示す。実線は親子関係を示す。左から、(1) 窓内の親個体が子供を持っていない場合のサイズ 1 の連結成分。(2) 両方の親個体が子供を 1 人だけ持っている場合のサイズ 3 の連結成分。(3) 雌雄同じ組み合わせに対して二つの個体が子供としているサイズ 4 の連結成分。(4) 子供の数は 2 であり、雄親が共通であるが、雌親が異なり、かつその両方の雌親が他に子供を生んでいないサイズ 5 の連結成分。この連結成分には親の雌雄が逆になっているパターンも存在することに注意。(5) 雌雄同じ組み合わせに対して 3 個体が子供であるサイズ 5 の連結成分。

(4),(5) $q(5)$ は (4) とその雌雄逆転と (5) の合計である。(4) とその雌雄逆転は、Derrida モデルが雌雄対称であることを考慮すれば、その割合は同じである。(5) は子供の親が常に同じになるように選ぶことを考えて、

$$q(5) = 2 \times \frac{\frac{N}{2}p_2}{\frac{N}{2}} \times \frac{{}_{Np_1}C_2}{{}_{N}C_2} + \frac{\frac{N}{2}p_3}{\frac{N}{2}} \times \frac{3\frac{N}{2}p_3}{N} \times \frac{1}{N-1} \times \frac{1}{N-2} \tag{4.4}$$

$$= 2{p_1}^2 p_2 \times \frac{1-\frac{1}{Np_1}}{1-\frac{1}{N}} + \frac{3{p_3}^2}{2(N-1)(N-2)} \tag{4.5}$$

$$\stackrel{N\to\infty}{\longrightarrow} 2{p_1}^2 p_2 \tag{4.6}$$

図 4.10 の (3) のように、連結成分にはループを持つものが存在する。ループを持つ条件について考察する。なおループを含む連結成分では、$N \to \infty$ では無視することが出来る。ループの定義は、連結成分内のある頂点から辺をたどっていき、同じ辺を通らず同じ頂点に戻ってくることができる経路が存在するとき、この連結成分はループを持つと定義する。また、その独立な経路の数がループの数となる。

頂点数 $|V|$、辺数 $|E|$ のグラフのオイラー標数 χ は $\chi = |V| - |E|$ で与えられる[*1]。すると、連結成分内のループの総数 L は、オイラー標数 χ を用いて、$1-\chi$ と表すことが出来る。我々が捉えたいループの数は $|E|$ を辺の数、$|V|$ を頂点数とすると、

$$L = 1 - \chi = |E| - |V| + 1 \tag{4.7}$$

で与えられる。ここで、連結成分内の親の数を n_1、子の数を n_2 とすると、

[*1] 面が存在する場合のオイラー標数はこれに面の数 $|F|$ を加えたものになるが、ここではグラフのループを数えるために $|F| = 0$ としている。

$|E| = 2n_2$、$|V| = n_1 + n_2$ が成り立つ。したがって、$L = n_2 - n_1 + 1 \leq 0$ となる。表 4.1 に示したように、$n_2 \leq n_1 - 1$ の条件を満たす時しか、連結成分は存在せず $n_2 > n_1 - 1$ を満たす時、少なくとも 1 組の親が 2 回選ばれる。すなわちループを持つ。したがって、ループを持たないのは、$n_2 = n_1 - 1$ の場合のみである。また、a を整数として $n_1 + n_2 = 2a$ で表される偶数だとすると、$L = (2a - n_1) - n_1 + 1 = 2(a - n_1) + 1$ となり、L は奇数となる。したがって、ループの数 L は 1 以上の奇数となり、偶数サイズの連結成分はループが存在することがわかる。ループが存在する連結成分に所属する親個体の親世代の個体数に対する割合には、同じ親を 2 回以上選ぶために N^{-1} のオーダーの項がある。よって、N が無限に大きい極限においては、ループをもつ連結成分に含まれる個体の親世代の個体数に対する割合は 0 となる。

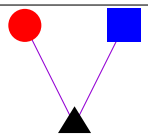
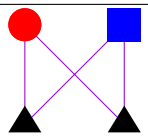
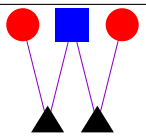
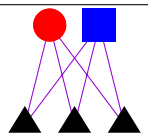
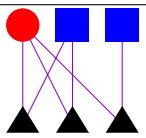
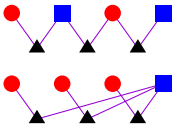
n_2 \ n_1	0	1	2	3	4
0	-	▲	-	-	-
1	-	-		-	-
2	-	-			-
3	-	-			
4	-	-	…	…	…

表 4.1: 親の数、子の数に対する連結成分の例。行が子世代の個体数 n_2、列が親世代の個体数 n_1 を示している。

Derrida モデルでは、子供の数の分布 p_k は $\lambda = 2$ のポアソン分布となる。表 4.2 には $\lambda = 2$ のポアソン分布の場合の $q(s)$ を実際に求めたものを載せている。なお、ポアソン分布は $p_k = \lambda^k e^{-\lambda}/k!$ である。

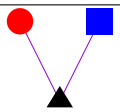
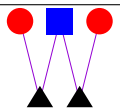
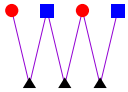
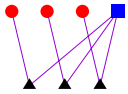
n_1	n_2	component	calculation
1	0		$p_0 = \frac{2^0 e^{-2}}{0!} = \frac{1}{e^2} \sim 0.1353353$
2	1		$\frac{{p_1}^2}{2} = \frac{1}{2} \cdot (\frac{2^1 e^{-2}}{1!})^2 = \frac{2}{e^4} \sim 0.0366313$
3	2		$\frac{3{p_1}^2 p_2}{4} = \frac{3}{2} \cdot \frac{2}{e^4} \cdot \frac{2^2 e^{-2}}{2!} = \frac{6}{e^6} \sim 0.0148725$
4	3		${p_1}^2 {p_2}^2 = \frac{8}{e^6} \cdot (\frac{2^2 e^{-2}}{2!}) = \frac{16}{e^8} \sim 0.0053674$
4	3		$\frac{{p_1}^3 p_3}{2} = \frac{2}{e^4} \cdot \frac{2}{e^2} \cdot \frac{2^3 e^{-2}}{3!} = \frac{16}{3e^8} \sim 0.0017891$

表 4.2: 親の数 n_1、子の数 n_2 に対する連結成分の $q(s)$ の値。p_k はポアソン分布で、N は無限に大きいと仮定し、ループがある連結成分は無視している。

4.3.2 巨大連結成分

次に、巨大連結成分が存在することを示す。巨大連結成分には“小さい連結成分以外が所属している”と考える。小さい連結成分の合計を見積もっていこう。先ほど述べたように、N が大きい極限では、$n_2 = n_1 - 1$ の場合のみの連結成分しか存在し得ない。$n_1 \leq 4$ では複数の形が存在したり、$n_1 = 3$ では雌雄非対称であることから、雌雄逆転した連結成分も考えられる。しかしながら $q(s)$ の計算の際には、子供に繋がっている親個体の子供の数を選び出す項 p_k をかけていくため、子供の数にしたがって、p_k の次数は増えていく。p_k の次数の合計は、$n_2 + 1$ となる。

得られた値を足して 1 から引くと、親世代の個体のうち表の連結成分に含まれていない個体の割合を表す。$1 - (0.1353353 + 0.0366313 + 0.0148725 + 0.0053674 + 0.0017891) = 0.8060044$ となる。大きい s に対する q の計算は、クラスタの構造が複数出て来るのでややこしくなるが q の値も (親の数だけ p_k がかかり) 小さくなると思われる。したがって適当なところで打ち切っても良い近似となると期待される。たとえば $q = 1/N$ では、1 個体が存在するかどうかになるので、$k = 1/N$ となる n_1 で十分だろう。すると、8 割近くの

個体がこれまで挙げたどの連結成分に所属しないことがわかる。それらはおそらくループをもった巨大連結成分を構成していると考えられる。

実際に、数値計算と理論計算を比較してみた。それを図 4.11 に示した。横軸は連結成分の種類を示している。エラーバーは標準偏差を示しているが、記号の大きさよりも小さいくらいであり、理論と数値計算の結果がよく合っている。

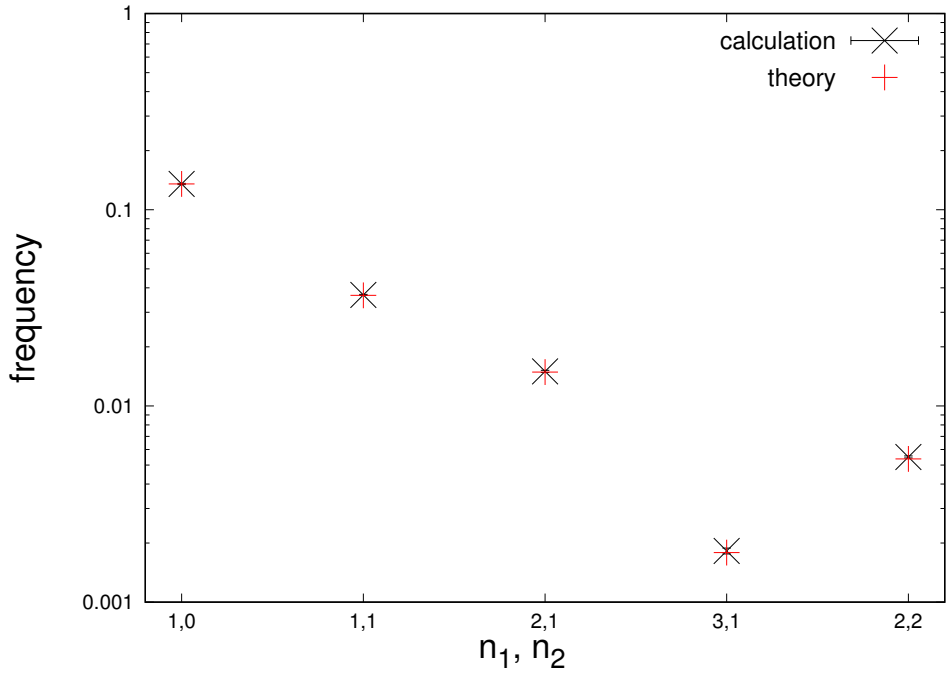


図 4.11. いくつかの小さい連結成分に対する $q(k)$ を比較した図。横軸が連結成分の種類を示しており、縦軸がその割合である。数値計算は 1 世代 100,000 個体で 10 回行い、その平均をとっている。エラーバーは標準偏差を示している。

第 5 章

実データの解析

この章では競走馬のデータを用いた実データの解析とその結果を述べる。

5.1　実データの取得と不完全性

この節では実データをどのように取得したかを述べ、実データに含まれる不完全性について解説する。我々の用いたデータは主に、生田の研究 [3] で使われたものと同様のものである。解説は生田の研究 [3] から多くを引用している。

実データにおける不完全な点を列挙してみよう。まず、データベースが閉じていない。先祖を辿っていくと遅かれ早かれ名前の分からない親に辿りつき、その以前の家系図は辿れない。最も過去にまで遡れたのは出生年が 1612 年の“ FAMILY NUMBER SIX ”であるが、それ以外の祖先は全てそれ以降で家系図を遡れなくなる。モデルでも最古の世代より遡ることが出来ないが、競走馬のデータの場合、名前の分からない親が出てくる世代は様々である。たとえば 2010 年に生まれた“ LITTLE ANGEL8”でも、その父親は“ unknown ”である。

次に、データベースは日々更新している。新しく生まれたものと思われる個体のみならず、古い時代のデータも書き換えられることがある。本研究で使用されたデータは 2013 年 11 月 6 日～2014 年 1 月 15 日の間に取得されたデータ及び、2017 年 11 月 22 日に取得されたデータである。ただし、その時点で登録されている全てのデータを取得できていないことにも注意しなければならない。

また、データには欠損や矛盾がある。出生年が特定できない個体や、親もしくは子の出生年と矛盾があるようなデータもある。データベース内の全競走

馬 1,729,747 頭のうち、出生年を特定できない競走馬は 14,598 頭（約 0.84 %）いる。また、親より子供が早く生まれている競走馬や、親が 100 歳以上で子供を産んでいるようなデータも存在する。

他にも生まれた子供が全てデータベースに登録されているわけではないと思われる。というのは生まれた子が全て競走馬になるわけではないからである。このように、競走馬のデータは不完全であり 100 %の信頼度がおけるわけではない。しかしながら、保存されたそのデータベースは膨大であり、我々はその親子関係を追跡することで個体数 1,729,747 頭のデータベースを作成することに成功した。これは質・量ともに類を見ないものであり、不完全なことを差し引いても解析すべき貴重なデータである。

5.2　有効なデータの範囲

前節で述べた通り、実データには少なからず不完全な点がある。この節ではそのことを踏まえて、有用なデータの範囲を検討する。

図 5.1 に生年に対する競走馬の頭数の推移を示した。横軸は競走馬の生年、縦軸は頭数を示している。データ数は 1,729,747 頭、雄が 655,511 頭で雌が 1,074,236 頭である。雌に対する雄の比率は約 61.0 %となる。競走馬の頭数は 1820 年代に生まれた馬の総計のように 10 年毎に総計をとっている。1820 年から 1980 年で最小二乗法を用いてフィッティングすると、時定数が 30.1（年）で漸近的標準誤差は 2.708 %となる。したがって、この範囲では指数関数的に増大していると考えられる。1990 年代から、競走馬の頭数が減っているのは探索が終わっていないからと考えられる。我々は比較的、一定の割合で頭数が増大している、1820 年代から 1980 年代に生まれた競走馬（雄 219,380 頭、雌 472,688 頭、合計 692,068 頭）を対象に研究を行った。

5.3　各特徴量

この節では第 5.2 節で述べたデータの範囲について、特徴量を計測した結果を示す。計測した量は、子数分布、親子間の年齢差分布、絶滅の個体数分布である。

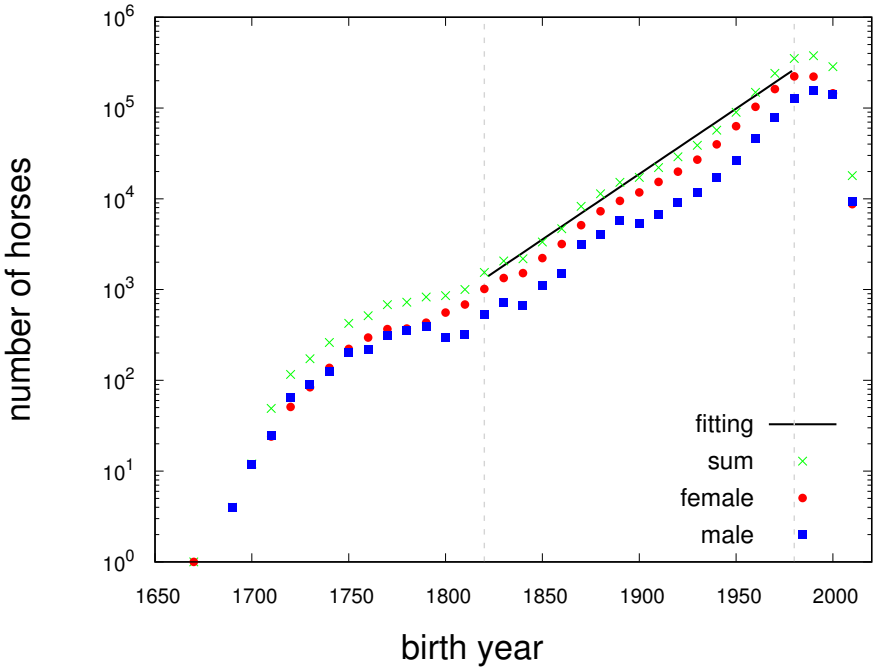


図 5.1. 成年に対する競争馬の個体数の推移を示した図。横軸が競争馬の生年。縦軸が競争馬の頭数を示す。プロットは 10 年毎である。縦軸は対数スケールになっている。1820 年から 1980 年に生まれた競走馬の頭数は指数関数的に推移している。本研究ではこの範囲のデータを主に用いる。

5.3.1　子数分布

家系図のネットワークの親子関係の線が子供から親へと向きがある有向グラフとみたとき、子数分布は in-degree の分布となる。また、子供の数という性質上、雌雄において大きな差が見られる。

雌の子数分布から見ていこう。図 5.2 に 1820 年から 1980 年に生まれた雌の競走馬の子数分布を片対数で示した。子供の数が 1 から 14 までに限れば時定数 2.53（年）、漸近的標準誤差 1.991 %でフィッティングすることができる。雌の子数は指数的に分布しているといってもよい。

次に雄の子数分布を見ていこう。図 5.2 に 1820 年から 1980 年に生まれた雄の競走馬の子数分布を両対数グラフで示した。雌の場合とは大きく違い、1,000 頭を超える子供を持つ雄馬も存在する。また、子供を一頭も持たない雄馬は全雄馬中 71.4 %にあたる 156,605 頭存在するが、両対数なので、図 5.3 には描かれていない。

以上のデータから雌雄が異なる子数分布に従っていることがわかる。表 5.1 に 1820 年から 1980 年に生まれた競走馬の子数の平均値、分散、標準偏差、数

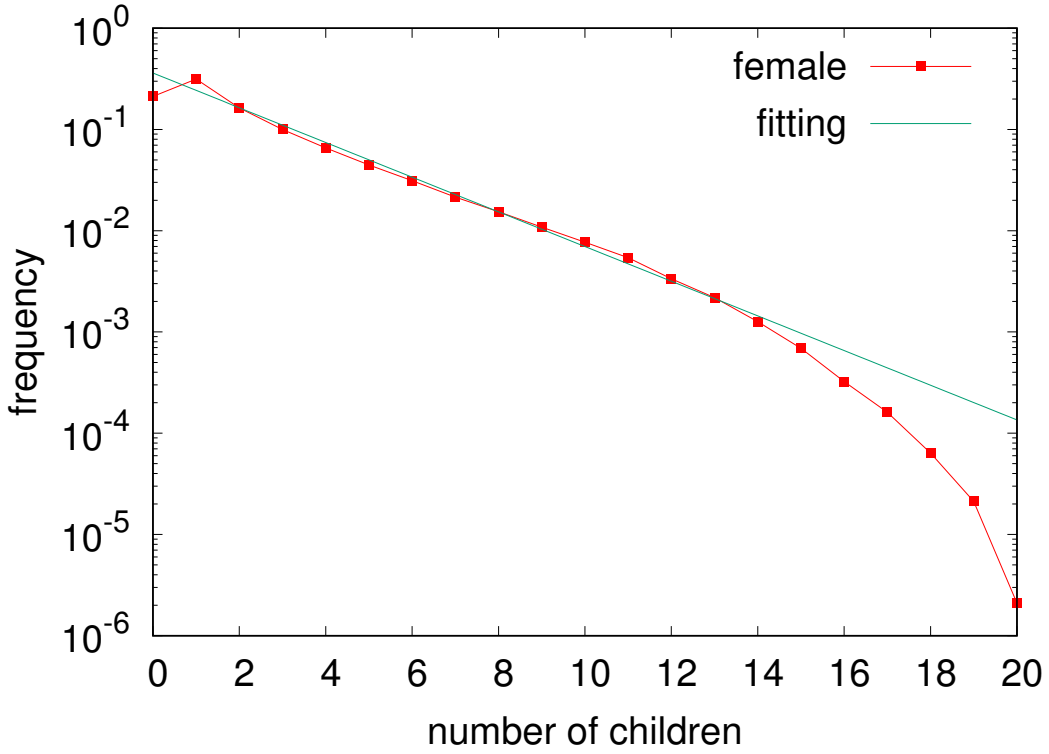


図 5.2. 1820 年から 1980 年までに生まれた雌の競走馬の子数分布。横軸は子供の数。縦軸はその割合。子供の数が 1 から 14 までに限れば時定数 2.53（年）、漸近的標準誤差 1.991 %でフィッティングすることができる。雌の子数は指数的に分布していることがわかる。

比をまとめた。

	合計	雄	雌
子数の平均	3.15	5.15	2.23
子数の分散	225	699	5.95
標準偏差	15.0	26.4	2.43
数比	1.0	0.32	0.68

表 5.1: 1820 年から 1980 年に生まれた競走馬のデータから求めた子数の平均値、分散、標準偏差、数比。

5.3.2　親子間の年齢差分布

親子間の年齢差分布は、馬の生育に依存しており、窓内の連結成分の大きさを考える上で非常に重要である。しかし、そのデータの扱いには注意を要する。当然のことであるが、親子間の年齢差を知るためには、親子の両方の生年が正確にわからなくてはならない。データベースの中には、1960s のように、1960 年代に生まれたという情報しかないような競走馬のデータも存

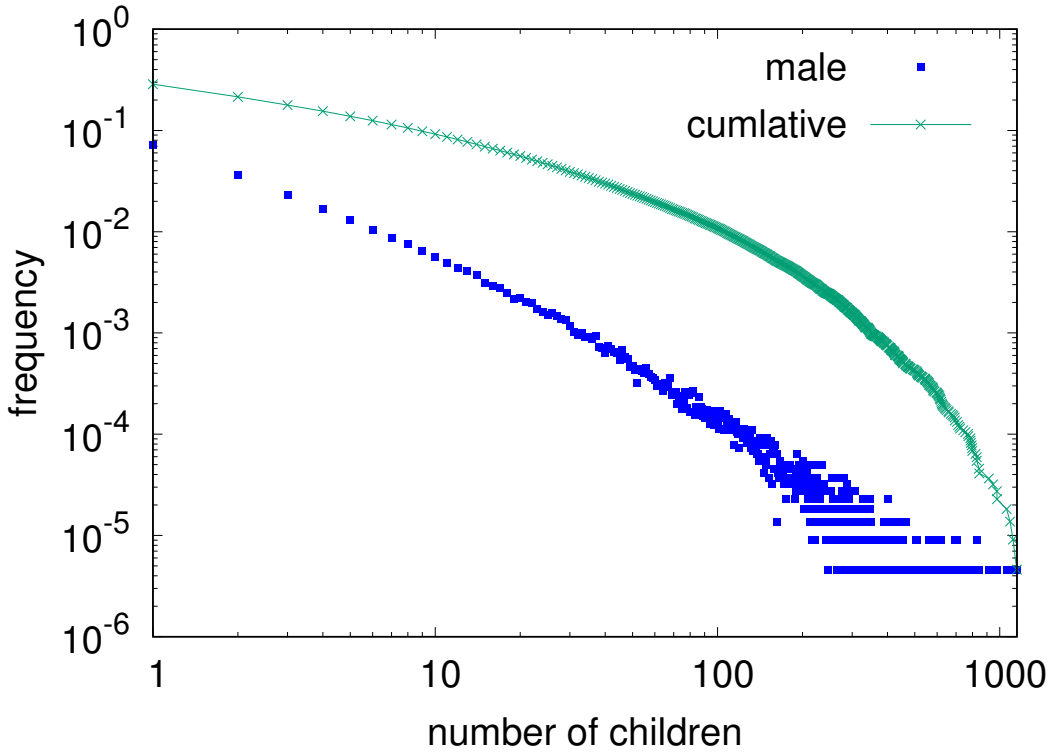


図 5.3. 1820 年から 1980 年までに生まれた雌の競走馬の子数分布及びその累積分布。横軸は子供の数。縦軸はその割合。両対数グラフである。雌の場合とは違い、1,000 頭を超える子供を持つ雄馬も存在する。

在する。競走馬のデータベース 1,729,747 頭のうち、自身の生年が正確にわからない競走馬は 14,798 頭（およそ 0.86 %）存在するが、年齢差分布の解析対象からは外している。データは 1820 年から 1980 年に生まれた競走馬が対象である。1820 年から 1980 年に生まれた競走馬は雌馬 695,849 頭、雄馬 348,065 頭、合計 1,043,914 頭存在し、雌馬の子供は 1,052,994 頭、雄馬の子供は 1,130,708 頭、合計 2,183,702 頭存在する。子供には母親と父親が存在するので、大部分が 2 回数えられていることに注意しなければならない。このうち、子の生年が正確にわからないものは、雌馬で 6,001 頭、雄馬で 10,611 頭、合計 16,612 頭、割合にして約 0.76 %であった。図 5.4 に親子年齢差分布を示した。図を見て明らかなように、親が生まれる前に子供が生まれているデータや、100 年以上たって子供が生まれているデータが散見される。これらは登録データの誤りであると考えられ、本解析では無視するのだが、無視するための条件をどう設定するか考察する。

当然であるが、自分が生まれるより前に子供が生まれることはありえない。子を産んだときの年齢が負になる 70 頭は無視することにした。馬の生態について記した [8] や競走馬について記した [9] によれば、雄馬の性成熟は 18〜

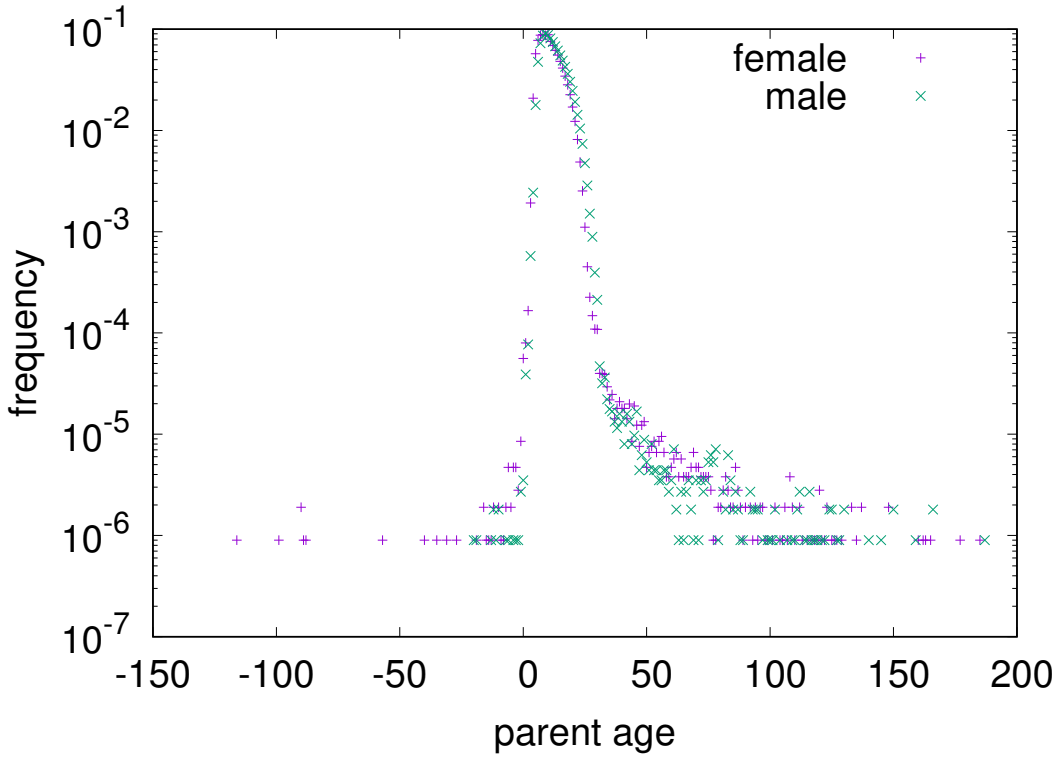


図 5.4. 雄雌それぞれの親子年齢差分布。横軸は子供を産んだときの親の年齢。縦軸はその割合。

24 ヶ月齢、雌馬は 1 歳では発情の徴候しか見せない。妊娠期間は 11 ヶ月強（平均 340 日）である。繁殖期は 4〜6 月であることを考えると、生年から 1 年後に子供を持つのはほとんど不可能である。2 年後もほぼ不可能であると考えられるが、繁殖期をずらす手法もあり不可能であるとは考えにくい。3 年後では子供を持つことも十分に可能であると推察される。以上のことから、1 年以内に子供が生まれているのはデータのミスであると考えられ、排除する。同年に生まれたデータ 63 頭、1 年差のデータは 128 頭であった。また、馬の寿命は長くても 50 年程度あり、親子年齢差 50 年以上のデータ 441 頭を除外した。以上、702 頭を除外した出産年齢差分布を図 5.5 に示した。

5.3.3 絶滅している個体数の割合

第 3 章で説明した絶滅している個体数を計算した。ある窓内のある連結成分が絶滅しているとは、子孫方向の窓外にでる親子関係の辺が存在しないことを意味している。以下これを、絶滅連結成分と呼ぶ。1820 年から 1980 年に生まれた競走馬に対して全ての Δ、全ての窓の取り方で、各連結成分が絶滅しているかどうかを調べれば絶滅している個体数の割合を算出できるように思える。この手段を手段 A とする。しかしながら、例えば、ある雌親と雄

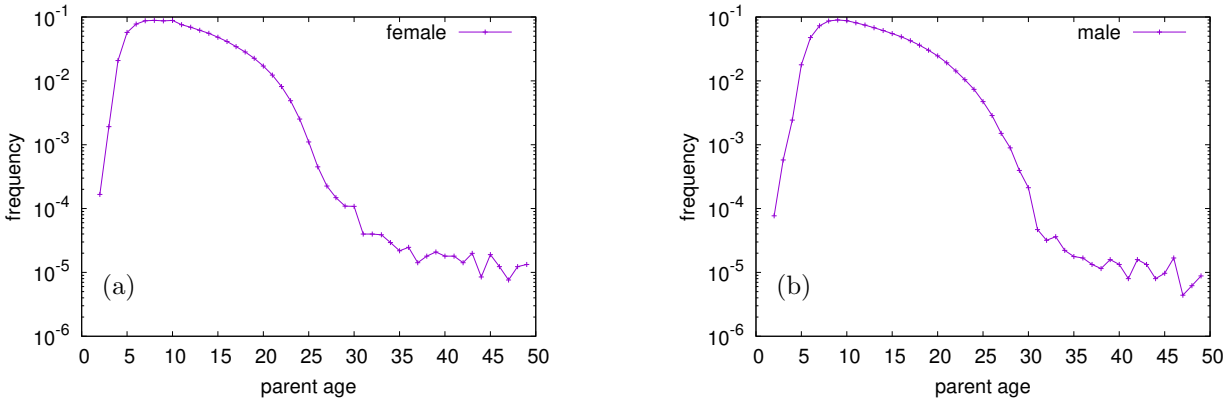


図 5.5. 雌雄それぞれの親子年齢差分布。(a) が雌で (b) が雄である。横軸は子供を産んだときの親の年齢。縦軸はその割合。親子年齢差 2 歳未満、または 50 歳以上は無視している。

親には子供が一頭だけ存在し、その子供が子供を持たなかった場合（図 4.10 の (3)）、これはサイズ 3 の絶滅している連結成分だと捉えることができるが、子供だけに着目すればサイズ 1 の連結成分だと考えることもできる。このように、ある連結成分が絶滅しているとき、連結成分内でもっとも新しい競走馬は必ず子供を持っていない。ならば全ての Δ で調べることにより、手段 A では連結成分の累積分布を出しているようにも思えるが、実際には同年に複数個体生まれる場合や、子供を持っていない子を複数持っている親がいる場合もあり純粋な累積分布であるとは言えない。例えば、表 4.2 の $n_1 = 3, n_2 = 2$ で子世代の個体が子供を持っていないとき、サイズ 1 とサイズ 3 の絶滅連結成分と捉えるが、サイズ 2 の絶滅連結成分とは捉えることができない。そこで我々は、全ての Δ、全ての窓の取り方で得られた連結成分が絶滅しているかどうかを調べた後、絶滅している連結成分の一部分が絶滅している場合は最大のものを残して全て除外することにした。それによって得られた絶滅の個体数割合を図 5.6 に示した。また、それによって得られたデータの累積分布も同図に示している。図を見る限り、テールのついた指数分布のように捉えることができる。全体の個体のうち、実に 4 割以上の個体が絶滅している個体であり、31.9 % にあたる 220,644 頭の親には別の子供がいるが自分には子供がいないサイズ 1 の絶滅連結成分である。サイズ 18 以上の連結成分で絶滅しているのは 6 個の連結成分のみであり、各サイズが 1 つだけ存在する。競走馬の家系図を粗視化した時に、絶滅の連結成分をどの程度とらえているか検討する必要がある。

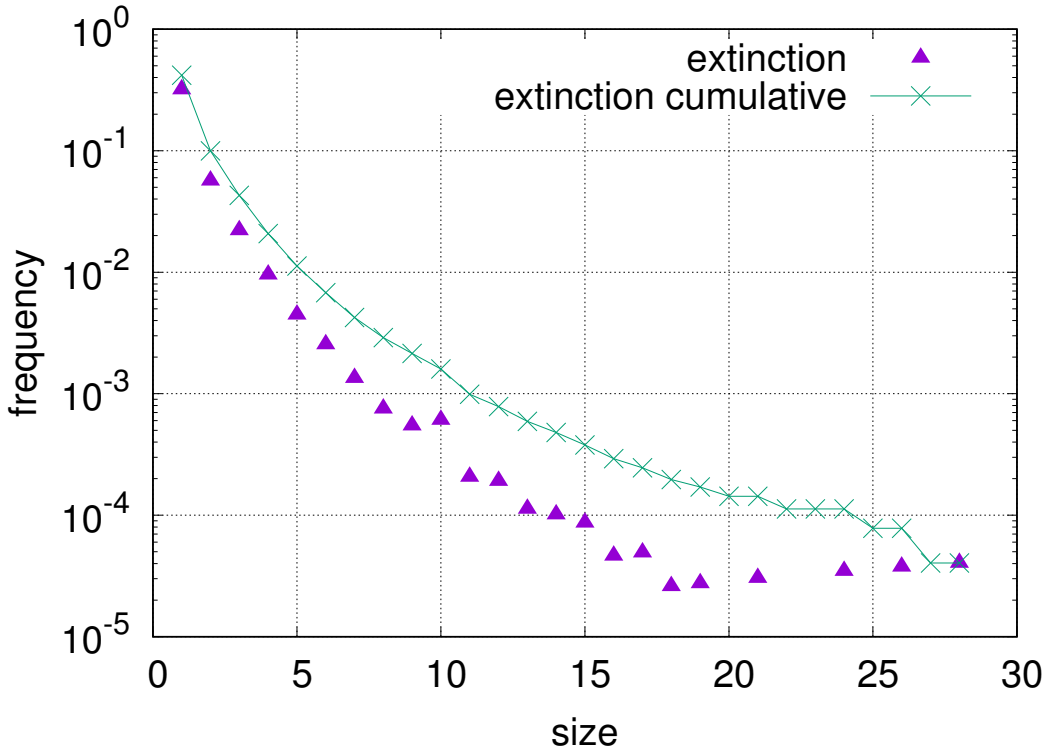


図 5.6. 1820 年から 1980 年に生まれた競走馬に対する絶滅した個体数の割合。横軸は連結成分のサイズ、縦軸は割合を示している。縦軸は対数スケールになっていることに注意。

5.3.4 連結成分のサイズ分布

第 3 章で述べたように、粗視化をする上でパラメータ Δ を決めなければならない。我々は、連結成分のサイズ分布を調べることにより、Δ を見積もった。図 5.7 は 1820 年から 1980 年までのデータに対して、Δ を 2 から 40 まで変えて、窓内の競走馬に対するその窓内の最大連結成分に含まれる競走馬の割合を取りうる全ての窓について調べ、平均を出したものである。Δ が増加するにつれて最大連結成分の平均は上がっていく。$\Delta = 10$ 程度から割合は急激に上昇し始め、$\Delta = 17$ で平均が 5 割を超え、$\Delta = 24$ で 8 割、$\Delta = 31$ で 9 割を超える。すなわち、$\Delta = 17$ ととることで窓内のおよそ 5 割の個体が一つの連結成分に属することになる。

次に、各 Δ に対して、連結成分の相対サイズ分布を調べた。いくつかの Δ に対する結果を図 5.8 に示した。$\Delta = 8$ のとき、どの窓内にも巨大連結成分は存在していない。Δ が 8 未満のとき、どのような窓をとろうとも、全ての連結成分のサイズの窓内の個体数に対する割合は、0.025 未満である。Δ が 8 以上 17 未満のとき、窓内の個体数に対する連結成分のサイズの割合が 0.025

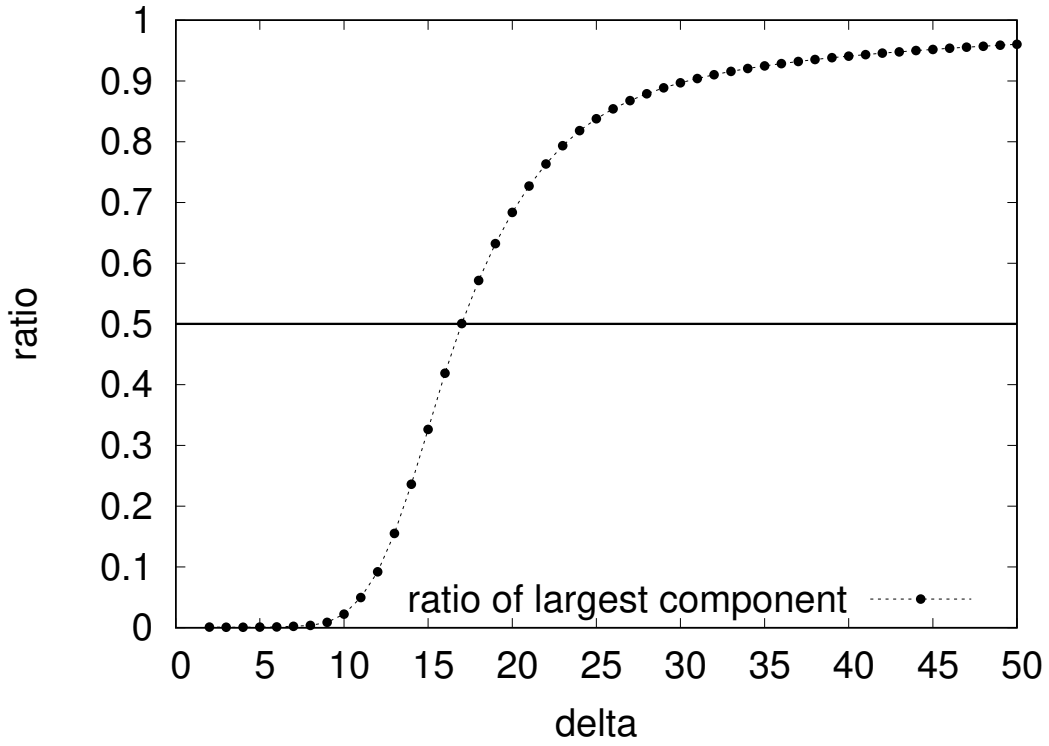


図 5.7. Δ を変えたときの、窓内の最大連結成分の割合の平均の推移。データは 1820 年から 1980 年に生まれた競走馬。窓は取りうる全てを取っている。Δ が増加するにつれて、最大連結成分の割合が上がっていることが確認できる。

を超えるものが存在する。しかし、巨大連結成分が存在するかどうかは窓の取り方に依存する。Δ が 17 以上のとき、窓内の個体数に対する連結成分のサイズの割合が 0.025 を超えかつ 0.2 を下回るようなサイズの連結成分は存在しなくなる。巨大連結成分が一つと、小さな連結成分がいくつもあるような窓の構造か、2 割か 3 割程度の連結成分が数個あるような窓が確認される。Δ が大きくなるにつれて、窓内の個体数に対する連結成分のサイズの割合は増えていく。Δ が 30 のとき、どの窓においても 8 割を超える巨大連結成分とそれ以外の連結成分にわけられる。

以上の情報から、我々は Δ を見積もった。我々の本来の目的は、親子関係の情報だけをもとにして、家系図の中の大まかな構造を捉えることである。もし、$\Delta = 30$ の場合のように、8 割を超えるような巨大連結成分が必ず存在し、粗視化をしたところで一本の太い線がつながっているようにしか見えない。逆に、$\Delta = 8$ の場合のように、連結成分のサイズが窓に含まれる競走馬の個数より著しく小さくなり、粗視化をしても情報量があまり減らず、ネットワークは入り組んだままとなる。

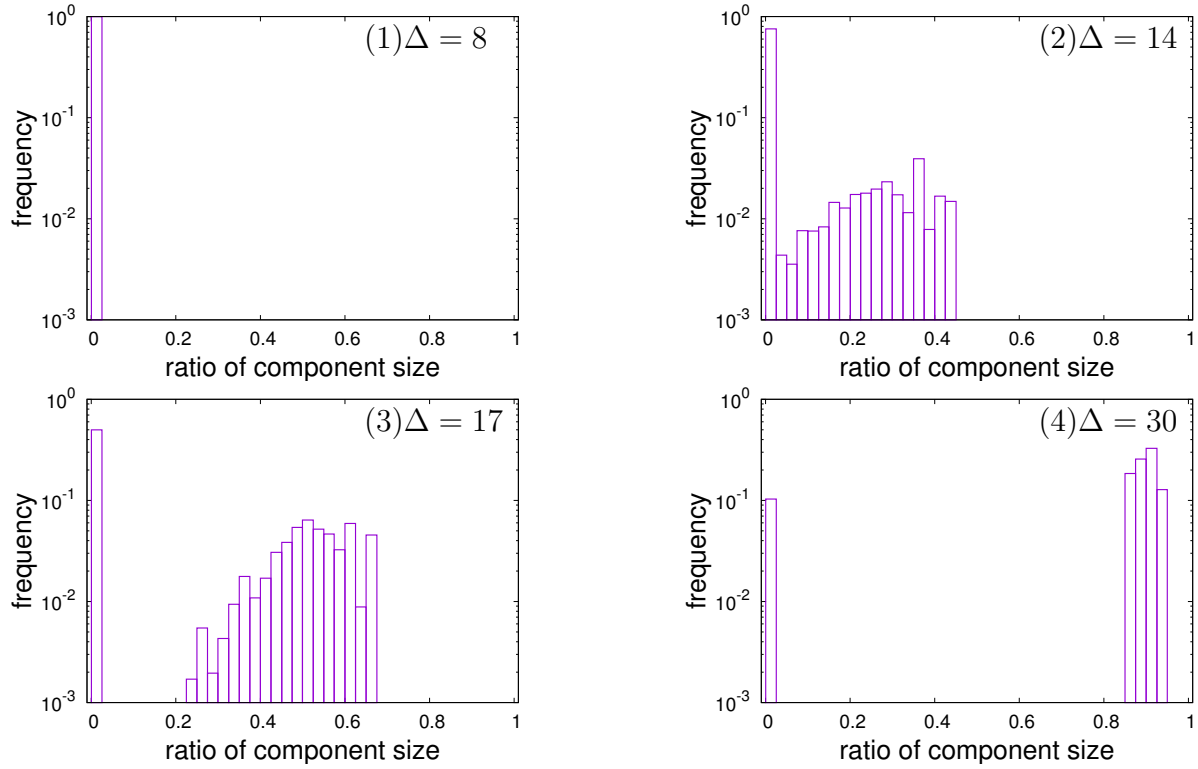


図 5.8. 窓内の連結成分の相対サイズ分布の平均を示した図。1820 年から 1980 年の取りうる全ての窓に対して連結成分のサイズ分布を計算し、その平均をとっている。横軸は相対サイズ、すなわち窓内の競走馬の頭数に対する、連結成分のサイズの割合。縦軸は、その連結成分に所属する競走馬の頭数の割合の平均。0 から 1 を 40 分割している。(1) は $\Delta = 8$、(2) は $\Delta = 14$、(3) は $\Delta = 17$、(4) は $\Delta = 30$ の場合である。

以上の考察から、粗視化パラメータ Δ は、巨大連結成分が存在せず、かつ、連結成分のサイズが窓内のサイズに対して 1 割、2 割程度の大きさを含んでいるような Δ が望ましい。$\Delta = 14$ 程度でいくつかの Δ がパラメータとして最適と考えられる。また、窓のずらす量 b の値を Δ より大きく設定すると、粗視化した家系図に含まれない競走馬が存在してしまい、情報を落とすことになるのでふさわしくないと考える。また、Δ より小さいと、同一の個体が別の連結成分で同時に含まれており、過不足なく数えられなくという観点からもふさわしくないと考える。以上の考察から、$b = \Delta$ であるとする。

5.4 粗視化

この節では競走馬の家系図を粗視化していく。まずは、粗視化する前の家系図を図 5.9 に示す。モデルの場合と同様に、粗視化する前では親子関係の線は複雑に絡み合い、構造を把握することは困難である。

第 5.3.4 節で考察したように、Δ は 14 程度、$b = \Delta$ であるとする。1820 年

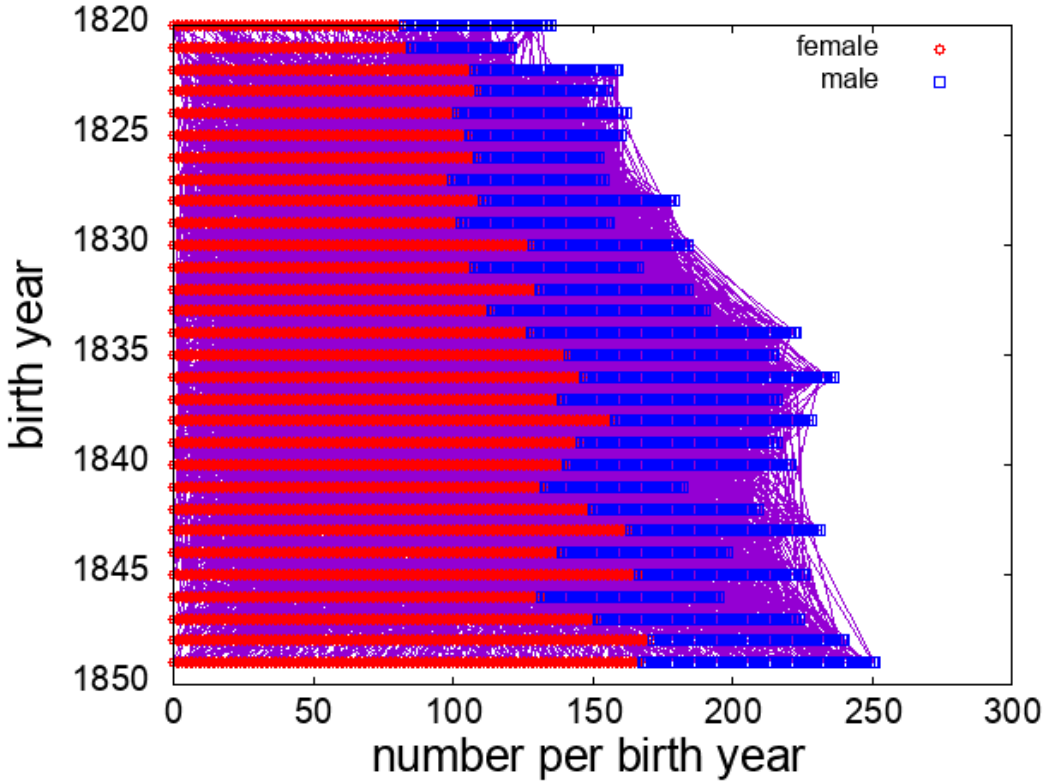


図 5.9. 1820 年から 1850 年に生まれた競走馬の家系図。縦軸は西暦で横軸は個体番号を表す。×は雌個体、＋は雄個体を示し、線は親子関係を示す。線が複雑に絡み合い、全体構造を把握することが困難な様子が見てとれる。

から 1980 年の競走馬のデータに対して粗視化を行った。それを図 5.10 に示した。各窓では、少数の大きな連結成分と多数の小さな連結成分に分割されていることがわかる。しかし、連結成分同士の辺は混み入っており、構造を把握することは容易ではない。

ここから、情報を削ったり、情報を強調することにより、大まかな情報を抽出することを考える。

第 5.2 章で示したように、競走馬の頭数は年代に対して指数関数的に増加する。メソ世代が 0 のところでは 2, 271 頭しかいないのに対し、メソ世代が 10 のところでは 230, 536 頭も存在する。ゆえに、図 5.10 をみると、メソ世代が 0 では頂点の大きさを把握しずらい。一つの図で表すことを考え、横軸の粗視化された頂点の大きさをそのメソ世代に対する割合で表すことにする。

また、家系図から大まかな構造を取り出すという目的から外れる情報を取り除いたり、情報を強調する。まず、相対サイズが 0.01 未満となる連結成分とつながっている辺は書かないことにする。また、相対サイズが 0.1 を超え

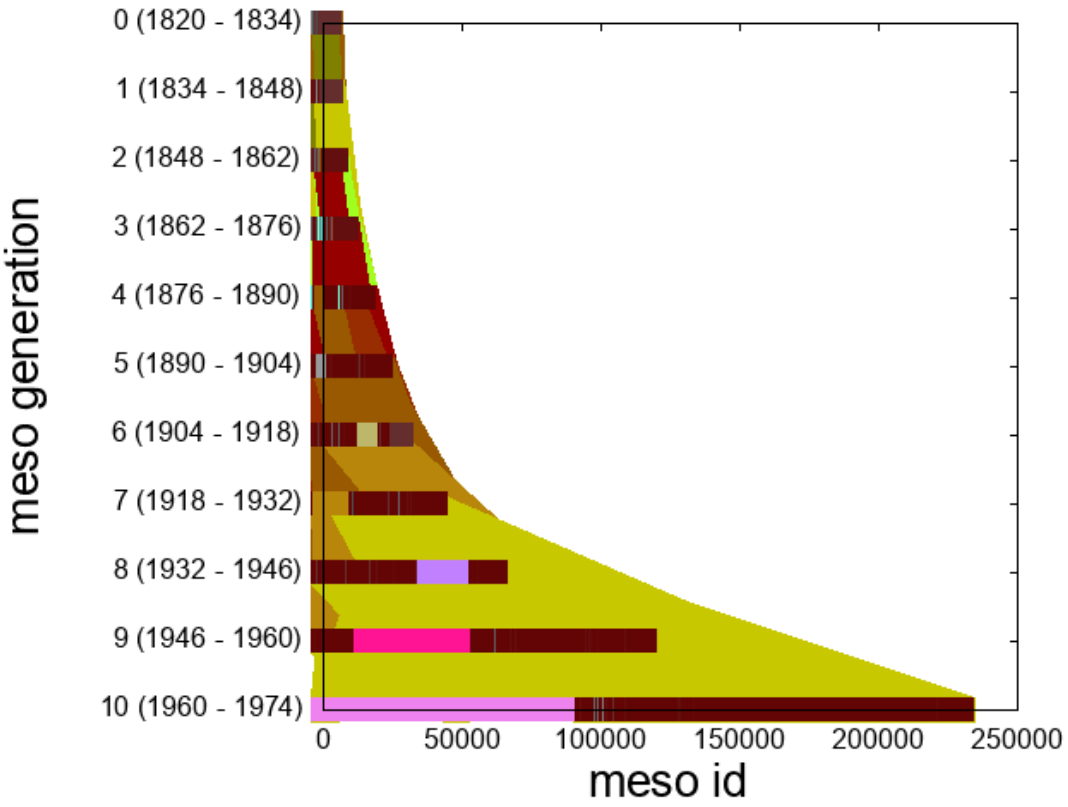


図 5.10. 1820 年から 1980 年に生まれた競走馬の家系図を $\Delta = b = 14$ で粗視化した家系図。縦軸は粗視化した世代、横軸には粗視化された頂点。一つの粗視化された頂点が一つの色で表されている。線は第 3 章で定義した、つながりがある頂点間をつないでいる。複雑に線が絡み合い、構造を把握することは困難である。

る連結成分同士を結ぶ線を強調する。さらに、相対サイズが 0.001 未満となる連結成分は描かないことにより、図 5.11 を得た。メソ世代が 0 か 1 では連結成分の相対サイズが 0.1 を超えない。世代を経るごとに、個体数は指数的に増大していくが、最大連結成分のサイズもメソ世代が 6 までは明らかに増大している。これは時代が進むにつれて、少数の種馬が多くの血統に関与し、血統の均一化が進んでいると考えられる。

血統の統一化をより定量的に表すため、各メソ世代における累積子数分布を求めた。それを図 5.12 に示した。時代を経る、すなわちメソ世代が増大するとともに多くの子供をもつ個体の割合が増えていることがわかる。$G = 0, 5, 10$ のそれぞれで子数の平均と標準偏差を調べると、表 5.2 のようになった。メソ世代が 0 から 5 になったときは、雄の平均、標準偏差共に値が小さくなっているが、メソ世代が 10 のときは値が大きくなっている。特に、標準偏差は大きな差がある。雌もメソ世代を経るごとに平均、標準偏差共に大きくなっている。

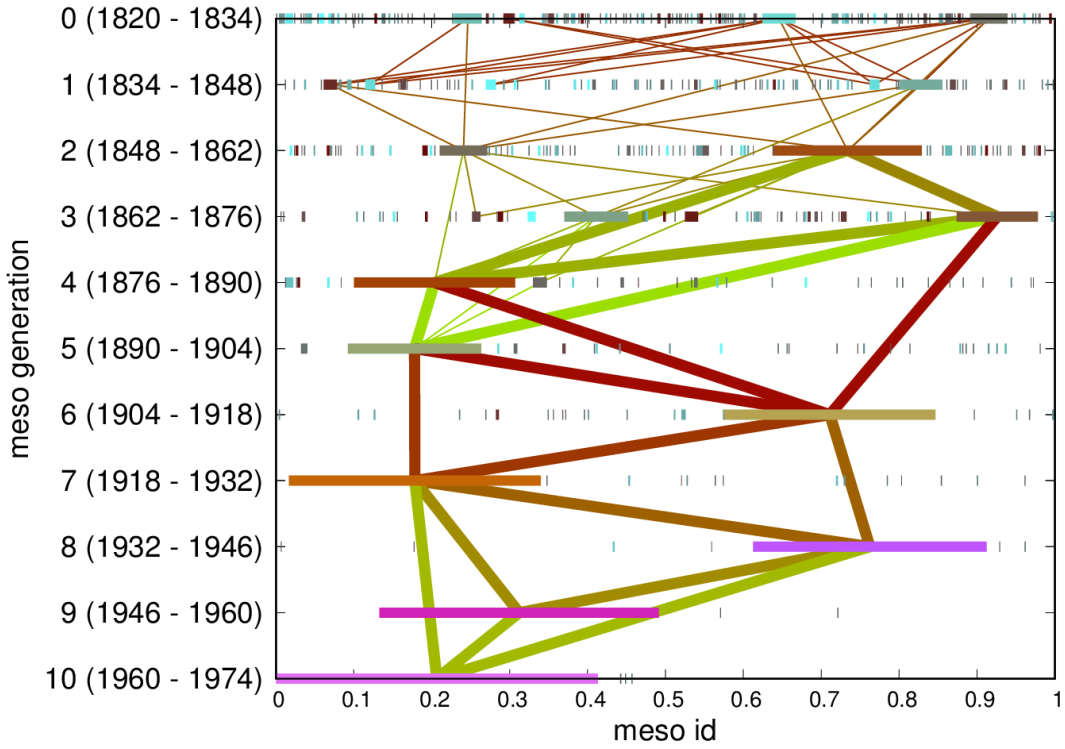


図 5.11. 1820 年から 1980 年に生まれた競走馬の家系図を $\Delta = b = 14$ で粗視化した家系図。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。相対サイズが 0.01 未満となる連結成分とつながっている辺は描かれていない。また、相対サイズが 0.1 を超える連結成分同士を結ぶ辺を太くして強調している。さらに、相対サイズが 0.001 未満となる連結成分は描かれていない。いずれも微細構造であるとしている。

性別	メソ世代	平均	標準偏差
雌	0	1.75	1.56
	5	1.78	1.79
	10	2.35	2.58
雄	0	4.96	15.8
	5	3.34	14.1
	10	5.61	29.0

表 5.2: メソ世代が $0, 5, 10$ のときの、雌雄の平均と標準偏差

次に、各メソ世代ごとに子供を持たなかった馬の割合を調べた。その結果が、図 5.13 である。メソ世代を重ねるにつれ、雌雄どちらも子供を持たない競走馬の割合が増加していることがわかる。雌の場合は緩やかな増加であり、またその割合も 1 割から 2 割程度であるが、雄の場合は 4 割から 7 割まで増加している。子供を持たない馬の割合が増え、その分が他の馬に集約されてい

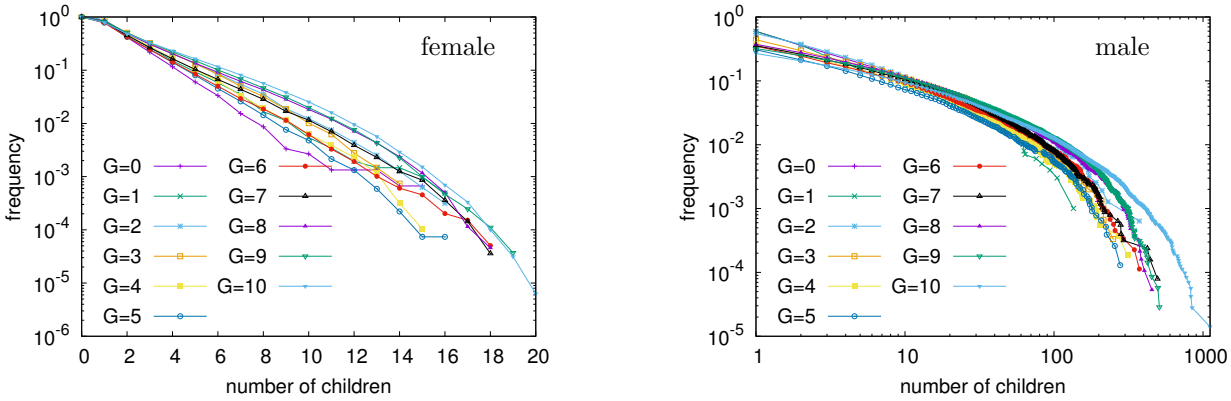


図 5.12. 雌雄それぞれのメソ世代ごとの累積子数分布。左図が雌で右図が雄である。両図とも、縦軸は対数スケールだが、横軸は雄のみ対数スケールになっていることに注意。横軸は子供を産んだときの親の年齢。縦軸はその割合。

ることがわかる。

なお、付録には 1821 年から 1834 年のそれぞれの年から粗視化した家系図も載せている。結果の要旨は変わらない。

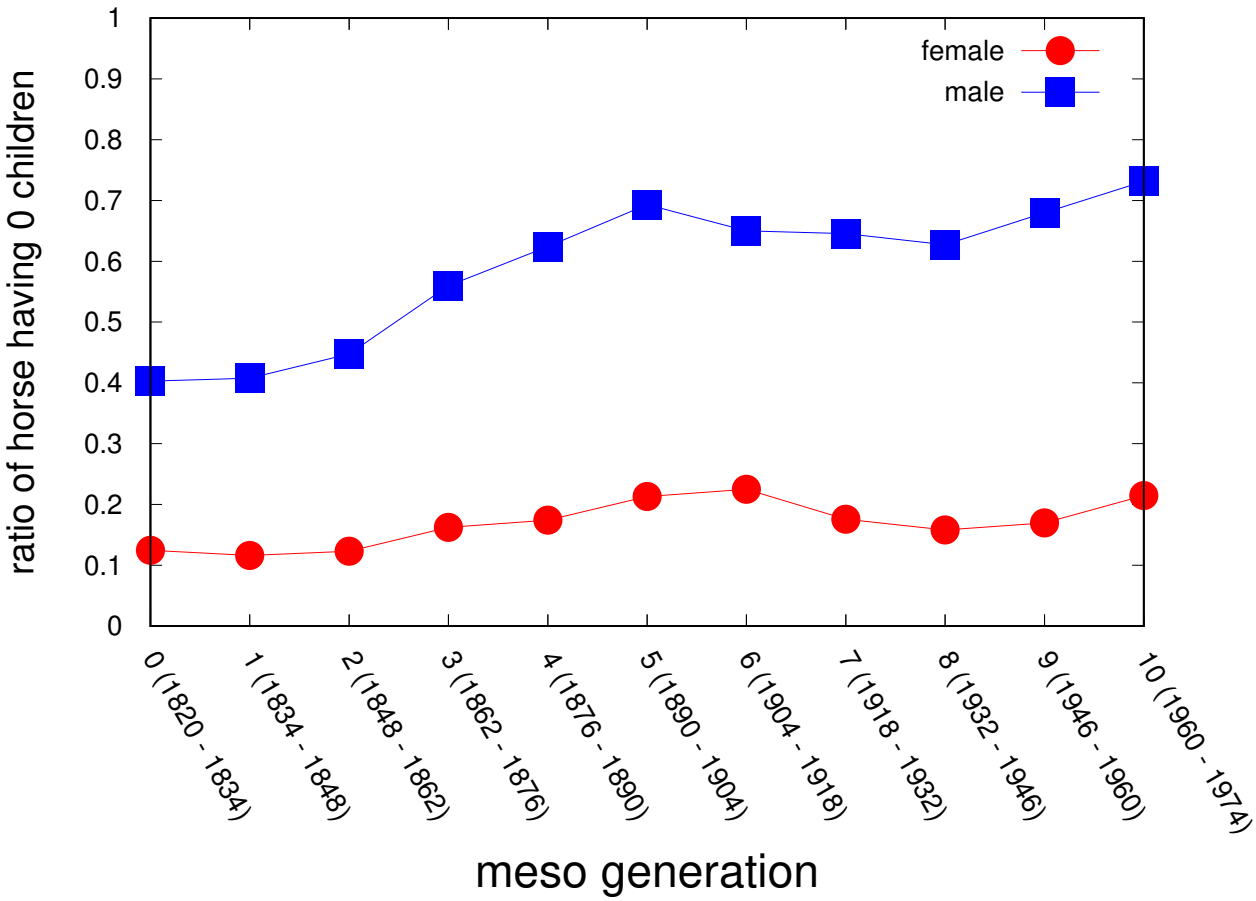


図 5.13. メソ世代ごとの子供を持たなかった競走馬の割合。横軸はメソ世代、縦軸は割合を示す。雌は赤色の●で、雄は青色の■で示されている。

第 6 章

まとめと今後の課題

本論文では、家系図の粗視化の手続きを導入した。その手続きは、次のようにまとめることができる。

1. 家系図の中の幅 Δ の窓に注目し、その中の連結成分、すなわち親子関係でつながっている個体集団を一つの塊としてとらえ、粗視化された頂点と呼ぶ。
2. 各窓をずらすことで各窓に粗視化された頂点が定義でき、粗視化された頂点に含まれる個体の親子関係を基にして、粗視化された頂点同士がつながるかどうかを定義した。

こうして得られたネットワークは、頂点同士のつながりも、もともとの家系図の親子関係によって決まっていることからもとの家系図を粗視化したものといえる。また、粗視化することで、家系図の大まかな構造を捉えることが可能になることを確かめるために、あらかじめどのような構造を持っているかがわかっているような生物集団の家系図を二種類作成した。そして、これらに対して粗視化手続きを適用した。具体的にはDerridaモデルを拡張し、2つの交配条件及び個性と相性関数を導入した。

乱婚制を用いて作成した家系図に対して $\Delta = 2, b = 2$ で粗視化を行ったところ、2つの小集団に分かれている様子が確認できた。一方、一夫一妻制を用いて作成した家系図に対して $\Delta = 2, b = 2$ で粗視化を行ったが、分岐構造は捉えることが出来なかった。しかし、$\Delta = 3, b = 3$ で粗視化をすると、分岐構造を捉えることが出来た。

そして実際の生物集団として競走馬を取り上げ、家系図を再構成した。その結果、競走馬は雌雄の数比に差があり、また子数分布がそれぞれ異なること、世代の重複が起きていることがわかった。粗視化の手法を適用することによ

り、競走馬の家系図を微細構造と大まかな構造に分けることに成功した。

今後の課題としては、モデルと実際の離散的な世代や子数分布など大きな違いがあり、モデルを実際の生物集団に近づけることや、Δが3以上での窓内の連結成分のサイズ分布を理論的に計算することが挙げられる。また、粗視化した競走馬の家系図に対するさらなる解析として、連結成分内の構造を調べることも挙げられる。他の生物集団に粗視化の方法を適用して、競走馬の家系図との違いやモデルとの差異を考察することも課題として挙げられる。

謝辞

本研究において、指導教官である水口毅准教授に心より感謝申し上げます。水口准教授には、研究活動において多くの助言や指導を賜りました。特に、修士論文の執筆においては数え切れないほどのご助力を賜りました。ここに厚く御礼申し上げます。また、大同寛明教授、堀田武彦教授、福田浩昭講師にも数多くの助言をいただきました。諸先生方には大変感謝しております。また、同期の城塚庸行氏とは数多くの意見を交わし多くを学ばせていただきました。植田智明氏には研究姿勢についてたくさんの刺激をもらいました。後輩の皆様には、研究活動以外の面でも多くの刺激を受けました。諸先生方、同期、後輩の皆様への感謝と、何不自由なく学生生活を送らせていただいた両親への感謝の意を表し、本論文の結びと致します。

付録

A　粗視化を始める年をずらした粗視化

本付録では粗視化を開始する年代を 1825 年から 1834 年まで変えた図を載せている。それを図 A.1 から図 A.4 に示した。メソ世代を経るごとに、連結成分のサイズが大きくなるという結果は変わりないが、1829 年から粗視化した家系図では、メソ世代が 0 でも相対サイズが 0.1 を超え、メソ辺が太く示されている。粗視化の開始年が数年変わる程度は結果に大きな差を与えないことがわかったが、10 年程度ずらすと結果も変わってくることが観察された。

B　子数のランク

この付録では、1820 年から 1980 年に生まれた競走馬の中で子数の上位 100 番までを列挙する。表 B.1 に示した。性別は雄に限られ、子数が 1000 を超えるのは 4 頭であった。また、馬の生年は 1970 年代に集中しており、そのうちわけは 1930 年代生まれが 1 頭、1950 年代生まれが 2 頭、1960 年代生まれが 28 頭、1970 年代生まれが 69 頭となっている。

ランク	名前	性別	生年	子数
1	ghadeer	Male	1978	1149
2	mr+prospector	Male	1970	1114
3	deputy+minister	Male	1979	1087
4	danzig	Male	1977	1057
5	miswaki	Male	1978	977
6	seattle+slew	Male	1974	976
7	crafty+prospector	Male	1979	946
8	runaway+groom	Male	1979	914
9	clever+trick	Male	1976	846

ランク	名前	性別	生年	子数
10	nijinsky2	Male	1967	844
11	kris+s	Male	1977	829
12	green+dancer	Male	1972	829
13	riverman	Male	1969	828
14	affirmed	Male	1975	820
15	sir+tristram	Male	1971	799
16	alleged	Male	1974	798
17	conquistador+cielo	Male	1979	794
18	be+my+guest	Male	1974	789
19	clackson	Male	1976	787
20	irish+river	Male	1976	786
21	cipayo	Male	1974	774
22	raise+a+native	Male	1961	771
23	lyphard	Male	1969	755
24	nureyev	Male	1977	743
25	roselier2	Male	1973	724
26	fast+gold	Male	1979	719
27	cure+the+blues	Male	1978	712
28	far+north	Male	1973	705
29	kris	Male	1976	705
30	be+my+native	Male	1979	693
31	sir+ivor	Male	1965	692
32	strong+gale	Male	1975	692
33	majestic+light	Male	1973	685
34	mr+leader	Male	1966	679
35	tumble+lark	Male	1967	671
36	fit+to+fight	Male	1979	659
37	silver+hawk	Male	1979	658
38	thatching	Male	1975	648

ランク	名前	性別	生年	子数
39	lord+avie	Male	1978	642
40	vaguely+noble	Male	1965	638
41	arctic+tern	Male	1973	633
42	coxs+ridge	Male	1974	628
43	damascus	Male	1964	626
44	storm+bird	Male	1978	623
45	habitat	Male	1966	620
46	tom+rolfe	Male	1962	620
47	deep+run	Male	1966	617
48	northern+dancer	Male	1961	617
49	known+fact	Male	1977	614
50	relaunch	Male	1976	612
51	halo	Male	1969	611
52	spectacular+bid	Male	1976	611
53	distinctive+pro	Male	1979	607
54	persian+bold	Male	1975	606
55	northern+taste	Male	1971	602
56	valid+appeal	Male	1972	598
57	shirley+heights	Male	1975	596
58	pirates+bounty	Male	1975	587
59	marscay	Male	1979	583
60	alydar	Male	1975	581
61	forli	Male	1963	580
62	ringaro	Male	1979	580
63	grosvenor2	Male	1979	578
64	sham	Male	1970	576
65	northfields	Male	1968	570
66	star+de+naskra	Male	1975	568
67	pleasant+colony	Male	1978	563

ランク	名前	性別	生年	子数
68	believe+it	Male	1975	562
69	ack+ack	Male	1966	561
70	grey+dawn2	Male	1962	561
71	stop+the+music	Male	1970	557
72	northern+guest	Male	1977	556
73	well+decorated	Male	1978	552
74	vice+regent	Male	1967	551
75	mountdrago	Male	1977	546
76	apalachee	Male	1971	545
77	secretariat	Male	1970	544
78	wolf+power	Male	1978	542
79	northern+baby	Male	1976	533
80	executioner	Male	1968	530
81	busted	Male	1963	528
82	his+majesty	Male	1968	525
83	st+chad	Male	1964	519
84	star+way2	Male	1977	518
85	private+account	Male	1976	514
86	graustark	Male	1963	513
87	robellino	Male	1978	512
88	stalwart	Male	1979	510
89	olden+times	Male	1958	509
90	temperence+hill	Male	1977	509
91	salt+marsh	Male	1970	506
92	key+to+the+mint	Male	1969	506
93	raja+baba	Male	1968	500
94	nashua	Male	1952	498
95	hyperion	Male	1930	495
96	wavering+monarch	Male	1979	491

ランク	名前	性別	生年	子数
97	silver+buck	Male	1978	483
98	silent+screen	Male	1967	481
99	explodent	Male	1969	480
100	bold+ruckus	Male	1976	476

表 B.1: 1820 年から 1980 年に生まれた競走馬の中で子数の上位 100 頭のデータ

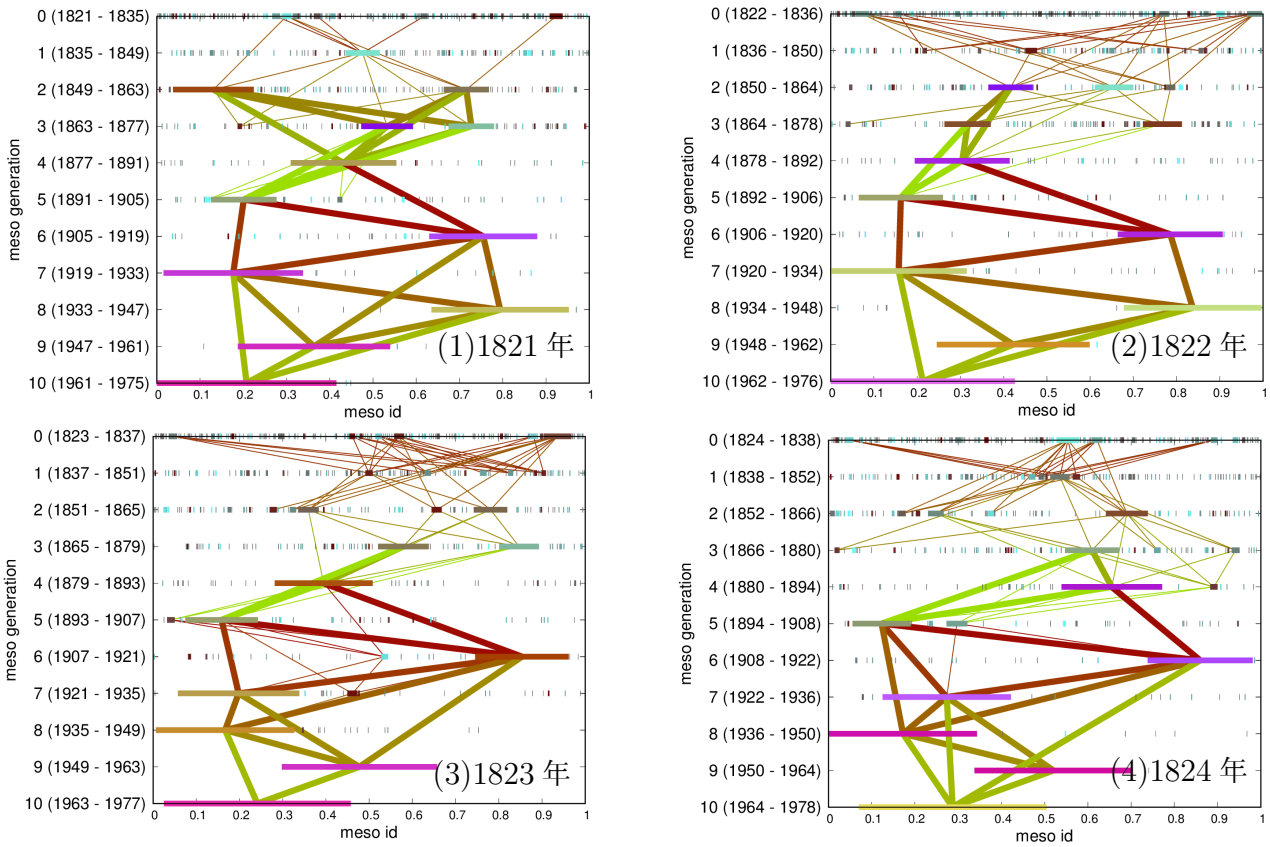


図 A.1. 競走馬の家系図を $\Delta = b = 14$ で粗視化した家系図。（1）1821 年から 1980 年に生まれた競走馬（2）1822 年から 1980 年に生まれた競走馬（3）1823 年から 1980 年に生まれた競走馬（4）1824 年から 1980 年に生まれた競走馬が対象となっている。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。相対サイズが 0.01 未満となる連結成分とつながっている辺は描かれていない。また、相対サイズが 0.1 を超える連結成分同士を結ぶ辺を太くして強調している。さらに、相対サイズが 0.001 未満となる連結成分は描かれていない。いずれも微細構造であるとしている。

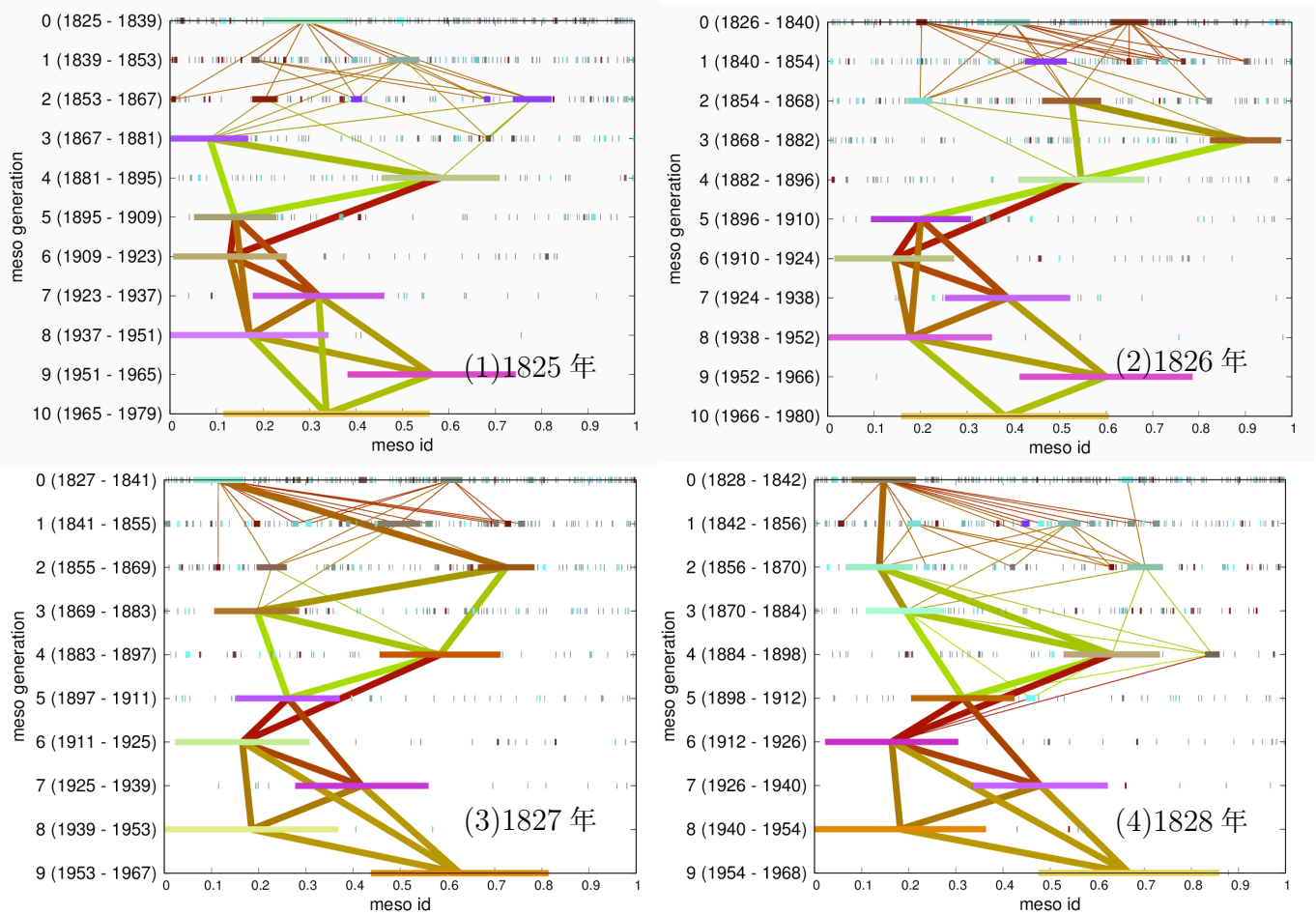


図 A.2. 競走馬の家系図を $\Delta = b = 14$ で粗視化した家系図。（1）1825 年から 1980 年に生まれた競走馬（2）1826 年から 1980 年に生まれた競走馬（3）1827 年から 1980 年に生まれた競走馬（4）1828 年から 1980 年に生まれた競走馬が対象となっている。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。相対サイズが 0.01 未満となる連結成分とつながっている辺は描かれていない。また、相対サイズが 0.1 を超える連結成分同士を結ぶ辺を太くして強調している。さらに、相対サイズが 0.001 未満となる連結成分は描かれていない。いずれも微細構造であるとしている。

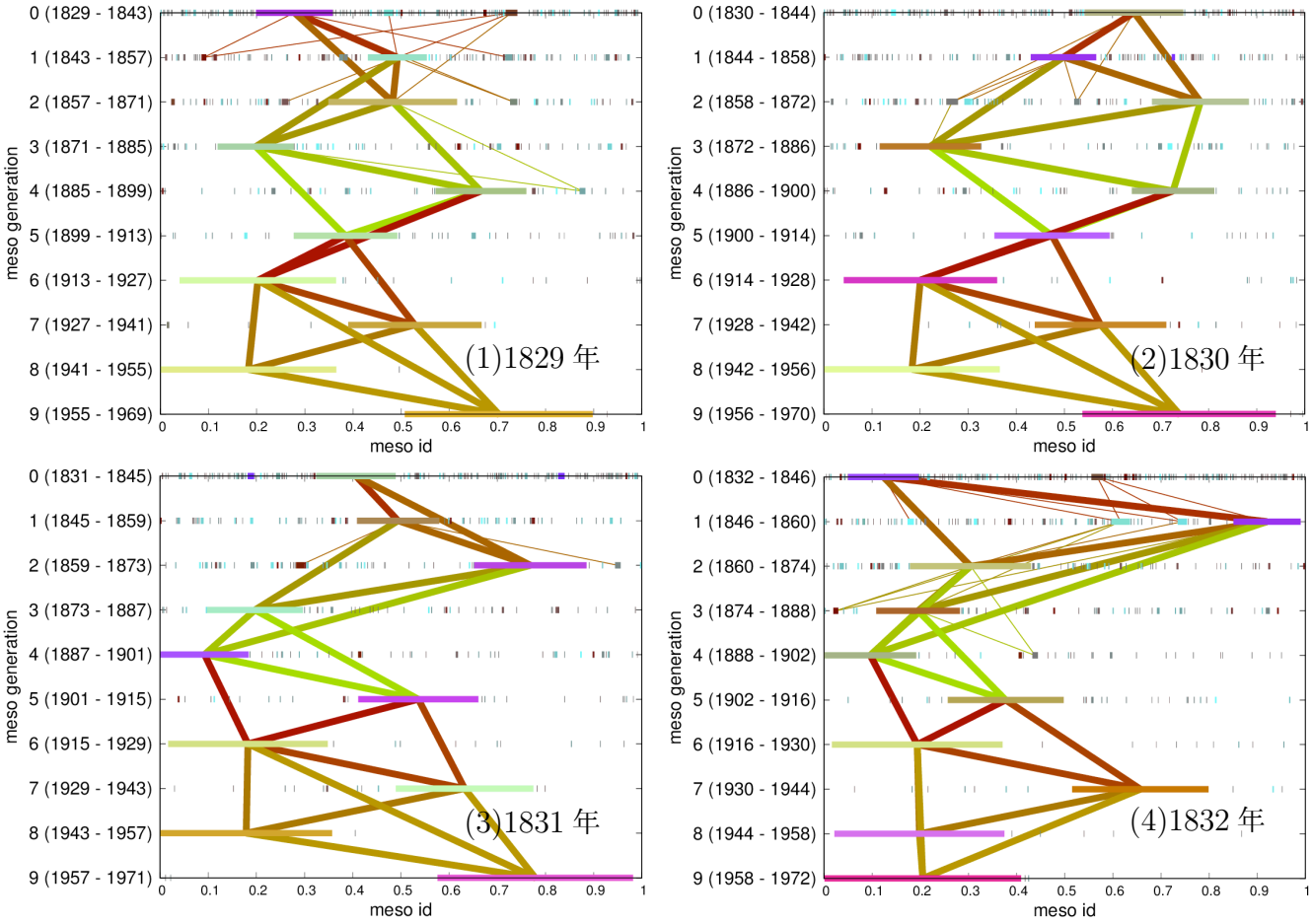


図 A.3. 競走馬の家系図を $\Delta = b = 14$ で粗視化した家系図。（1）1829 年から 1980 年に生まれた競走馬（2）1830 年から 1980 年に生まれた競走馬（3）1831 年から 1980 年に生まれた競走馬（4）1832 年から 1980 年に生まれた競走馬が対象となっている。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。相対サイズが 0.01 未満となる連結成分とつながっている辺は描かれていない。また、相対サイズが 0.1 を超える連結成分同士を結ぶ辺を太くして強調している。さらに、相対サイズが 0.001 未満となる連結成分は描かれていない。いずれも微細構造であるとしている。

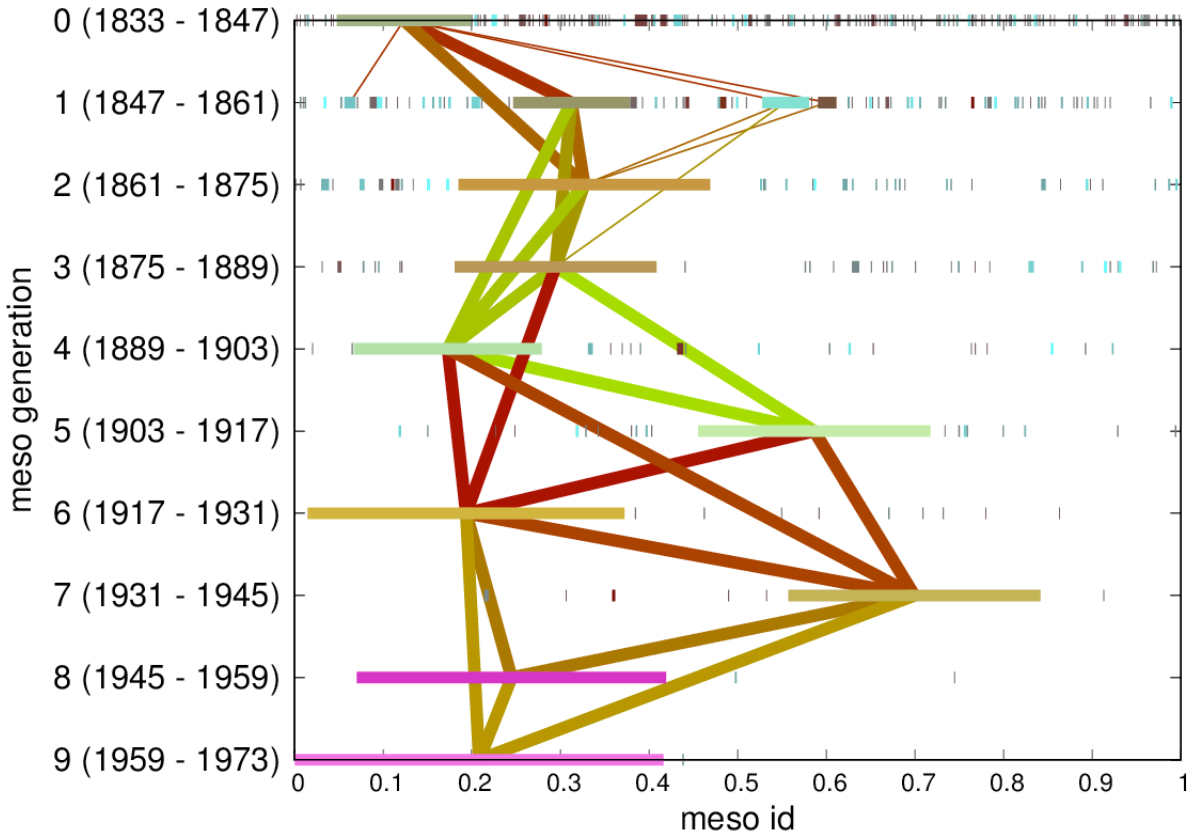


図 A.4. 競走馬の家系図を $\Delta = b = 14$ で粗視化した家系図。1833 年から 1980 年に生まれた競走馬が対象となっている。横軸がメソ個体番号で、縦軸はメソ世代を示す。辺は粗視化された頂点 V 同士の関係を示している。横軸の一つの色の線が一つの連結成分を示している。相対サイズが 0.01 未満となる連結成分とつながっている辺は描かれていない。また、相対サイズが 0.1 を超える連結成分同士を結ぶ辺を太くして強調している。さらに、相対サイズが 0.001 未満となる連結成分は描かれていない。いずれも微細構造であるとしている。

参考文献

[1] B. Derrida, S. C. Manrubia and D. H. Zanette, "On the Genealogy of a Population of Biparental Individuals", *J. theor. Biol.*, **203**, pp. 303-315 (2000).

[2] 堀内陽介, 「家系図ネットワークの構造解析」, 大阪府立大学修士論文 (2011).

[3] 生田成望, 「家系図ネットワークにおける継承過程と構造解析」, 大阪府立大学修士論文 (2014).

[4] 伏尾佳悟,「家系図ネットワークの粗視化 -モデルによるアプローチ-」, 大阪府立大学卒業論文 (2015).

[5] 三中信宏, 「生物系統学」, 東京大学出版会 (1997).

[6] 伏尾佳悟, 三浦圭二, 水口毅, 「有性生物の家系図ネットワークの粗視化」, 日本物理学会第 72 年次大会講演概要集, 19pB18-9 (2017).

[7] M. E. J. Newman, "The Structure and Function of Complex Networks", *SIAM REVIEW*, **45**, pp. 167-256 (2003).

[8] JRA 日本中央競馬会競走馬総合研究所,「新 馬の医学書」, 緑書房 (2012).

[9] 日本ウマ科学会, 「競走馬ハンドブック」, 丸善出版 (2013).