Arch. Tierz., Dummerstorf 44 (2001) 5, 521-530

Department of Animal Breeding and Husbandry, Faculty of Animal Science, Agricultural University of Athens, GR

ANTONIOS P. KOMINAKIS

Graph analysis of animals' pedigrees

Dedicated to Prof. Dr. E. Rogdakis

Summary

In the present work, an attempt to apply graph analysis on visual representations of animals' pedigrees is presented. Analysis of pedigree networks of moderate size (tens or hundreds of points) can substantially contribute to revealing the relational structures between animals. Partitioning graphic representations of pedigree networks to smaller parts (blocks) by means of network decomposition methods resulted in better handling and understanding of horse genealogical data. Analysis of pedigree networks could be used to estimate shortest kinship paths among animals, determinate all predecessors and successors of selected animals and finally to estimate inbreeding coefficients of selected individuals. Detection of families and animals with major gene contribution could substantially be facilitated. Graphic representation of pedigree networks provide simultaneous, dynamic and parsimonious representation of kinship, interrelations and constituent structures.

Key Words: pedigree, graph of pedigree, kinship, network, analysis of descent

Zusammenfassung

Titel der Arbeit: **Graphische Analyse von Zuchttierpedigrees**

Vorliegender Beitrag analysiert Möglichkeiten der graphischen Analyse und die Darstellung von Zuchttierpedigrees. Die graphische Analyse von Abstammungsnetzwerken kann wesentlich zum Verständnis bestehender Verwandtschaftsbeziehungen beitragen. Dargestellt werden verschiedene Methoden zur Zerlegung umfangreicher Netzwerke in kleinere Netzwerke, die eine bessere und leichtere Analyse z.B. von Pferdegenealogien ermöglichen. Die graphische Analyse von Pedigrees kann die Übersicht von Tierfamilien und die Schätzung von Inzuchtkoeffizienten von Zuchttieren unterstützen. Das Erkennen von Familien oder Zuchttieren mit größerem genetischen Gewicht wird erleichtert. Es wird gezeigt, dass die graphische Darstellung von Pedigreenetzwerken dazu dient gleichzeitig einfache, dynamische, übersichtliche Abbildungen von Verwandtschaftsverhältnissen sowie bestehende Wechselbeziehungen und Teilstrukturen zu erkennen.

Schlüsselwörter: Pedigree, graphische Analyse, Verwandtschaft, Netzwerk, Abstammungsanalyse

Introduction

The use of visual images is common in branches of science. Visualization provides powerful tools for the investigation of various relational structures, such networks, e.g. of computers, transportation, the Internet, communication, intra/inter organizational networks. Applications arise in economics (project management, work-flow diagrams), computer science (flow graphs of programs, data base modelling, algorithm animation), social science (social networks), natural science (large molecules and flow-diagrams in Chemistry, visualization of excavations in Archaeology) and communication networks (links among servers, usage of Internet, phone calls) (SNYDER and KICK, 1979; FREEMAN, 1988; BATAGELJ and MRVAR, 2000).

Two distinct forms of display can be used to construct images of networks, one based on points and lines (graphs) and the other on matrices. In a matrix display form, the rows and columns both represent points and numbers while symbols in the cells show the connections linking those points. Large networks (hundreds or thousands of points), based on matrix representation, cannot be treated efficiently by standard analysis tools. Application of standard network analysis is therefore limited to networks of moderate size (tens or hundreds of points).

The overwhelming majority of network images have involved the use of graphs. A graph is defined as a set of points and a set of lines that connect points, written mathematically as G=(V, E) or G(V, E). According to the branch of science, there are many synonyms for the term "point": node, vertex, actor and junction and for term "line": edge, link, branch, tie or arc (HARARY, 1969; FOULDS, 1992). In a genealogy or pedigree graph, the points represent persons or animals and the lines represent relationships among them. Graph representations of pedigrees are directed graphs (also known as digraphs) because lines have directions. Animals' pedigrees are usually large networks, consisting of thousands of points and lines.

Standard visualization approaches usually do not give satisfactory results in the case of large graphs. In the present paper, graph analysis has been implemented in an attempt to discern the internal structure(s) of a pedigree. Different methods of network decomposition are recruited to demonstrate how graph-theoretic analysis of pedigrees can contribute to: i) partitioning the relational structure of a pedigree, ii) estimate shortest kinship paths among animals, iii) determinate all predecessors and successors of selected animals and iv) to estimate inbreeding coefficient of selected individuals.

Material and Methods

Definitions

Graph theory

Graph theory has a wide repertoire of terms. A good reference on the subject offers the book of HARARY (1969). Only a brief description of the terms used in the present study will follow here. A *sub-graph* of a graph G is a subset of its points together with all lines connecting members of the subset. A *path* is an alternating sequence of points and lines beginning at a point and ending at a point, and which does not visit any point more than once. A *cycle* is just a path except that it starts and ends at the same point. A *bicomponent* of a graph G is a sub-graph of G with three or more points in which any pair of points is connected by two independent paths, hence, by a set of edges that form a cycle. The *length* of a path (or cycle) is defined as the number of lines in it. The shortest path between two points is called a *geodesic*. The *graph theoretic distance* or *geodesic distance* between two points is defined as the length of the shortest path between them. The *diameter* of a connected graph is the largest geodesic distance. A graph is *connected* if there exists a path (of any length) from every point to every other. A *connected component* is a maximal sub-graph in which all points are reachable from every other. Maximal means that it is the largest possible sub-graph. No point could be added in the sub-graph without violating its property. For directed graphs, like pedigrees, there are *strong* and *weak components*. A strong component is a maximal sub-graph in which there is a path from every point to every point following all the lines in the direction they are pointing. A weak sub-graph is a maximal sub-

graph, which would be connected if the direction of the lines were ignored. The *distance* of a connected graph is the maximum distance between two of its points. The *density* of a graph is the number of actually occurring lines as a proportion of the theoretically possible lines. In a directed graph, the *density* (d) is given by the formula $d = \frac{k}{n(n-1)}$ where k and n are the number of lines and points in the graph, respectively. *Abstraction* is the (recursive) factorization of a large graph into several smaller graphs. One of the approaches to support abstraction is *find* clusters, i.e. subsets of points, *extract* and *show* points that belong to the same clusters, shrink points in clusters and show relations among clusters. A *clique* is a subset of the graph in which the points are more closely and intensely tied to one another than they are to other members of the graph.

Network analysis

Decomposition

One major goal of network analysis is to discern fundamental structure(s) of networks in a way that: i) allows the knowing of its structure and ii) facilitates the understanding of network phenomena. The most used tool is called *blockmodelling* and does this through partitioning of networks according to well-specified criteria. Blockmodelling seeks to cluster together units having substantially similar patterns of relationships with the rest of units of the network. A blockmodel consists of structures obtained by identifying all units from the same cluster of the clustering C. The graph version of the blockmodel is a *reduced graph*. Blockmodelling is an empirical procedure, based on the idea that units in a network can be grouped according to the extent to which is equivalent, according to some meaningful definition of equivalence. Two definitions of equivalence were extensively treated in the last three decades: *structural* and *regular equivalence*. In structural equivalence (LORRAIN and WHITE, 1971) the units of the network are structurally equivalent if they are connected to the rest of the network in identical ways. In regular equivalence, two units are regularly equivalent if they are equivalently connected to equivalent others, i.e. they have the same type of neighbors (SAILER, 1978). Further work dealt with *automorphic* equivalence (BORGATTI and EVERETT, 1992). Points are automorphically equivalent if the graph can be permuted is such a way that exchanging the two points has no effect on the distances among all points in the graph. *Reduction* is the recursive deletion of all points of the network that have only 0 or 1 neighbor points.

Connections

In a pedigree network, individuals may have many or few genetic ties. Furthermore, they may be 'sources' of genetic ties, sinks (receive ties but don't send) or both. The sum of connections from point v to others is called the *out-degree* of the point. The sum of connections to point v from other points is called the *in-degree* of the point. Animals with unusually high out-degree are influential or central animals in the network and have high *degrees of centrality*. The number and kind of ties that individuals have are keys to determining their embeddednsess in the pedigree. Examining local structures can reveal the way the individuals are embedded in a pedigree. The most common approaches here has been to look at *dyads*, i.e. sets of two points and *triads*, i.e. sets of three points (WASSERMAN and FAUST, 1988).

Index of connectedness (P)

$$P = \frac{k + m - n}{k + n - 2M}$$

where:
n – number of vertices
m – number of lines
k – number of weakly connected components
M – number of maximal vertices (vertices having output degree 0, M≥1)
with $0 \leq P \leq 1$. A graph having only one vertex has P=0. The highest connectedness (P=1) are for graphs representing matings between full-sibs.

Data

Pedigree data were available on Holsteiner stallion Libero through Data Horse Ltd. Data tracing back to 6 generations of the animal. They were retrieved via Internet through the web site of the above company (http://www.horses.nl/klaus2/6generat_e.htm). The total number of animals in the pedigree was 122 (including 121 ancestors and the animal Libero). The total number of sires and dams was 58 and 63, respectively. There were 61, 30, 16, 8, 4, 2 and 1 animals in generations 1–6, respectively. Two animals, (Loretto and Fanal) were used in two subsequent generations 1 and 2, respectively.

Pedigree analysis and display

Network analysis of the pedigree was performed by the computer program Pajek (BATAGELJ and MRVAR, 2000). Blockmodelling was according to algorithms described by DOREIAN et al. (1994). In most large networks the number of lines n is the same order as the number of the vertices. Such networks are considered *sparse networks*. The efficiency of an algorithm is accounted by its time T(n) and space S(n) complexity where T(n) and S(n) are estimates of the time and memory space needed to run it on instances of size n, respectively. Given the capabilities of nowadays computers, space complexity for storing sparse networks is not crucial anymore. Having much faster computers, however, does not help a lot in the case of high order time complexities. Most of the algorithms implemented in Pajek have subquadratic time complexities: O(n), O(nlogn), O(n$\sqrt{n}$), or are restricted to small sets of selected vertices. Graph representation of the whole pedigree as well as of its clusters was also performed by the Pajek program. Several standard algorithms for automatic graph drawing were used like spring embedders based on minimization of the total energy of the system KAMADA and KAWAI (1989) and drawing in layers.
The inbreeding coefficient of animal X was calculated by applying the classical formula

$$F_X = \sum_{CA=1}^{k} \left(\frac{1}{2}\right)^{n_1+n_2+1} (1+F_{CA}) \qquad (1)$$

where: CA= a common ancestor of the sire and dam of X, k = the number of common ancestors in X's pedigree, n_1 = the number of generations separating the common ancestor from the sire of X, n_2 = the number of generations separating the common ancestor from the dam of X, F_{CA} = the inbreeding coefficient of the common ancestor.

Results and Discussion

Figure 1 displays the graph E of the pedigree of animal Libero in genealogy layers. The total number of points and lines in this graph was 122 and 126, respectively. The resulting density (d) of the graph was thus d=126/(122*121)=0.00853 (~0.01) (Table 1). This means that only 0.85% of all possible relations are present which denotes a loose graph in terms of genetic relations between animals. Standard deviation (SD) of the connections in the graph was 0,09, i.e. substantially greater than the density. In terms of the coefficient of variation (CV (%) = SD / d) there is thus great variation in genetic relations between animals (CV (%)=0.09/0.01=900%). The index of connectedness (P) of this graph is P=0.04132 (Table 1).

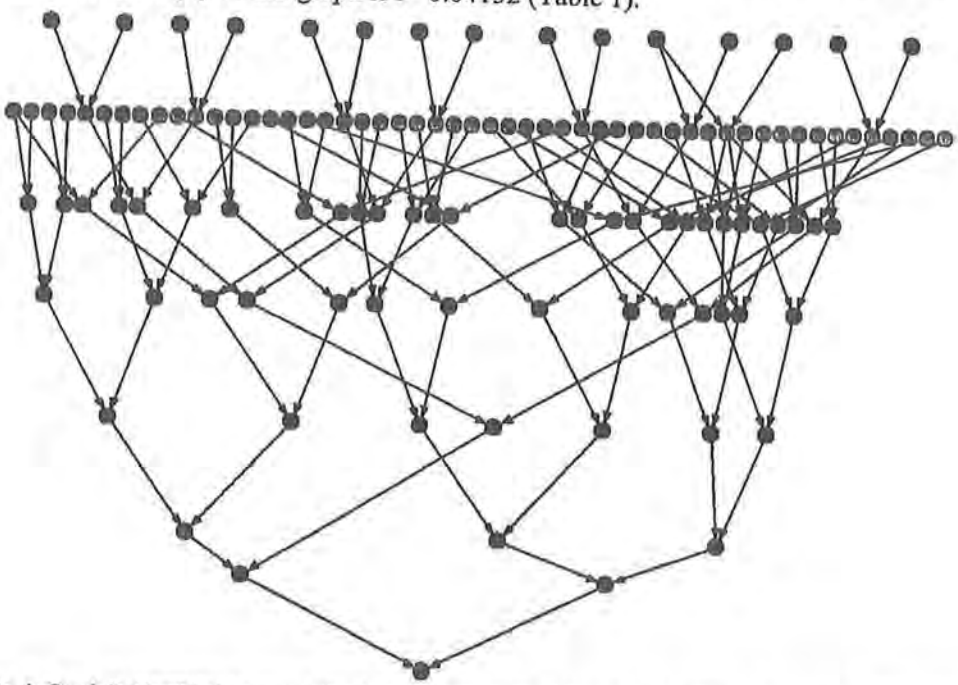

Fig. 1: Graph E: Graphical representation in genealogy layers of the original pedigree of the horse Libero (down edge)

Table 1
Density (d), diameter (δ) and index of connectedness (P) of the three graphs

Graph	Density (D)	Diameter (δ)	P
E	0.00853	6	0.04132
E1	0.0423	6	0.18519
E2	0.0605	6	0.21053

The diameter of graph E is 6 estimated on the following path that connects the ancestor Phalaris in generation 1 with Libero: Phalaris→ Colorado→ Loaningdale→ Lone Beech→ Ladykiller → Landgraf I→ Libero (Fig. 2).
Furthermore, the animals detected on all paths from animal Phalaris to Libero are: Phalaris, Fairway, Colorado, Blue Peter, Loaningdale, Sailing Light, (Lone Beech, Ladykiller, Landgraf I, Libero (Fig. 2).

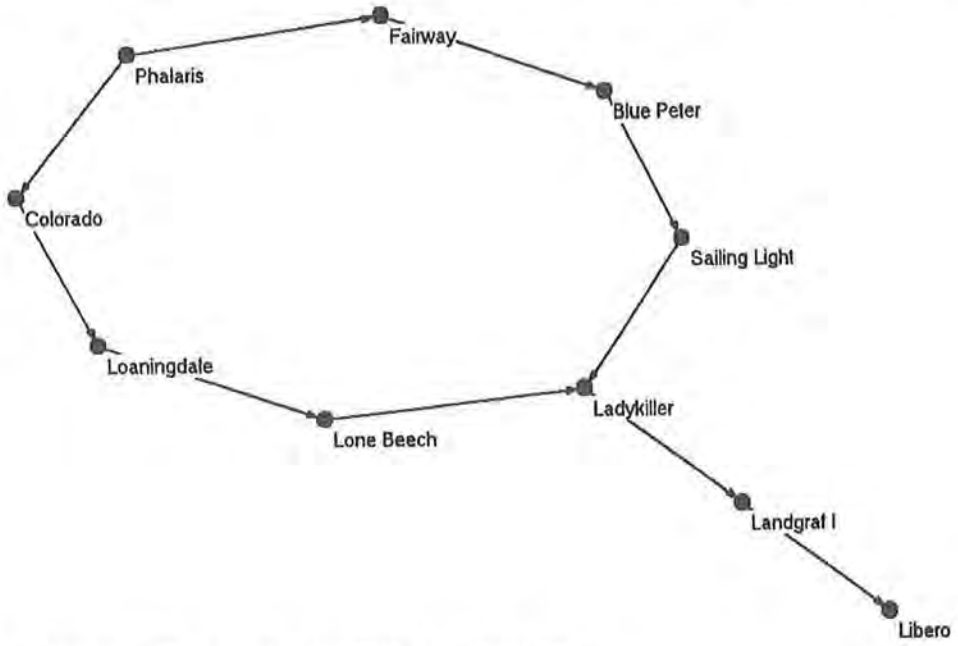

Fig. 2: Animals on all paths from animal Phalaris to animal Libero

Table 2
Type and number of triadic relations in the three graphs

Type	Graph E n	E1 n	E2 n
(1) 003	280319	2489	759
(2) 012	14722	742	348
(3) 021D	6	6	5
(4) 021U	63	7	5
(5) 021C	130	32	23
Sum(2-5)	14921	787	381
Total	310161	4063	1521
Sum/Total(%)	5.32	31.62	50.20

Most of the triadic connections of graph E are of type 1 (Table 2) which denotes a loose graph. This result is in agreement with the density value estimated above. There

are 14722, 6 and 63 relations of type 2, 3 and 4, respectively. Triads of type 3 denote a common parent, whose offspring are half-sibs. Triads of type 4 denote matings between two animals, sire and dam, resulting in full-sibs. Those triads can be considered representing families. Furthermore, there are 130 triadic relations that include members of three generations, e.g. grandparent sire, parent sire, sire.

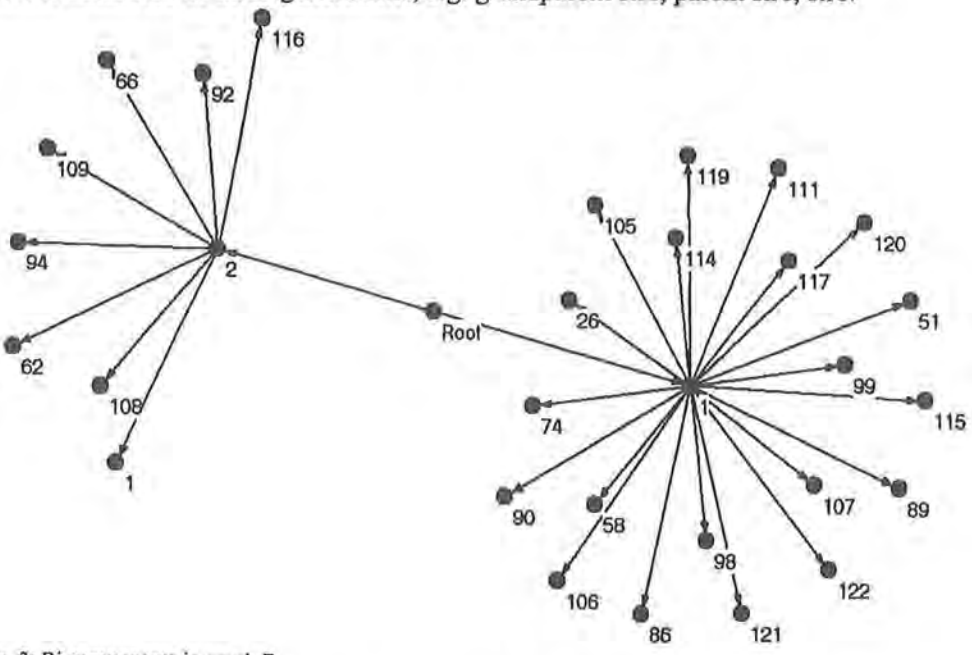

Fig. 3: Bi-components in graph E

There exist two bi-components in graph E (Fig. 3). One of them is large (C1), shown right of the graph-root and is containing 20 animals (animal Libero with code number 122 is including here) (figures before animal-names represent input coding numbers of animals).

C1: {(51)Loretto, (26)Lorbeer, (74)Oina, (98)Fangball, (89)Oresta, (106)Lowevenjager, (115)Loni, (119)Gelonika, (107)Emmia, (90)Fant I, (58)Fanal, (111)Schneenelke, (120)Landgraf I, (117)Warthburg, (99)Blümchen, (86)Lohengrin, (105)Valet, (114)Komet, (121)Oktave, (122)Libero}. The other bi-component, C2, is smaller and is containing 8 animals:

C2: {(1)Phalaris, (62)Fairway, (92)Blue Peter, (108)Sailing Light, (116)Ladykiller, (109)Lone Beech, (94)Loaningdale, (66)Colorado}.

The two bi-components were detected as p-cliques after blockmodelling of the original pedigree E. The respective cliques are displayed in sub-graph E1 (Fig. 4). In this reduced graph, the number of vertices and the number of arcs are 28 and 32, respectively, resulting in a density of 4.23% (Table 1). The index of connectedness of this sub-graph is P=0.18519. According to definition, animals in a clique are more closely and intensely related to one another than they are to other members of the pedigree. Indeed, the sum of the number of all triadic genetic relations in graph E1, expressed as a ratio to all number of triads, has increased up to 31.62% (Table 2).

Centers detected in sub-graph E1 with respective degrees of centrality (c) are shown in Table 3. Animals with c higher than 1 are: Schneenelke (c=2.0), Landgraf I (c=2.5),

Fangball (c=3.0), Loni (c=3.0), Ladykiller (c=3.0), Gelonika (c=3.0) and Loretto (c=4.5). Loretto has the highest degree of centrality; he has out- and in-degrees of 3 and 1, respectively. It is the most connected animal in sub-graph E1 implying a central point.

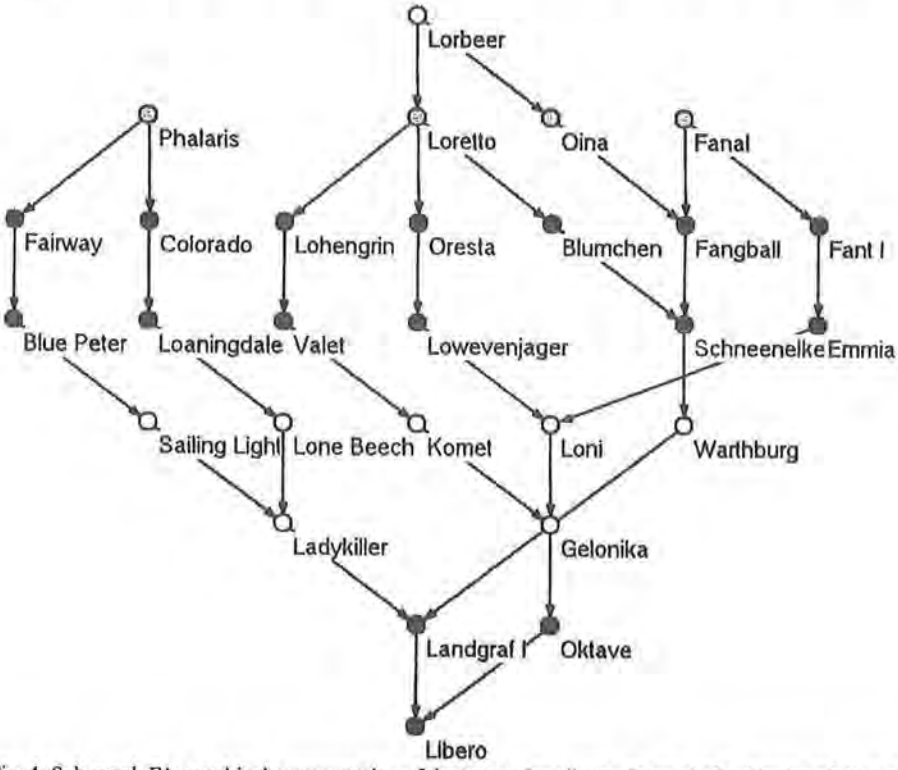

Fig. 4: Sub-graph E1: graphical representation of the two major cliques detected after blockmodelling of graph E

The left part of the sub-graph E1 is a cycle starting from animal Ladykiller and ending to the same animal after visiting the animals: Sailing Light, Blue Peter, Fairway, Phalaris, Colorado, Loaningdale and Lone Beech. Phalaris is the common ancestor (generation 1), followed by Colorado and Fairway (generation 2), Blue Peter and Loaningsdale (generation 3), Sailing Light and Lone Beach (generation 4) and Ladykiller (generation 5). Finally, Ladykiller is the dam of Landgraf I, the sire of Libero. All animals in clique 1 are therefore direct or indirect descendants of animal Phalaris. The inbreeding coefficient (F_X) of animal Ladykiller can be easily obtained by applying formula 1 resulting in F_X=0.0078125.

The second clique is more complex and it consisted of 20 animals. This clique is shown in sub-graph E2 (Figure 5). Since animal Libero is in this partition, a further analysis is needed. The number of vertices and arcs in sub-graph E2 is 20 and 23, respectively. The density of this sub-graph is thus d=0.0605 (6.05%) (Table 1). The index of connectedness of this sub-graph is P=0.21053. It is thus a dense graph. The lines without arcs from Loretto to Gelonika, Loni, Fangball and Schneelke represent indirect connections between Loretto with other animals (note, for instance, the path: Loretto, Lorbeer, Oina, Fangball).

Table 3
Centers and degrees of centrality (c) in sub-graph E1

Degree centrality (c)	Name
1.0	Phalaris
1.0	Fairway
1.0	Colorado
1.0	Fant I
1.0	Blue Peter
1.0	Loaningdale
1.0	Valet
2.0	Schneenelke
2.5	Landgraf I
3.0	Fangball
3.0	Loni
3.0	Ladykiller
3.0	Gelonika
4.5	Loretto

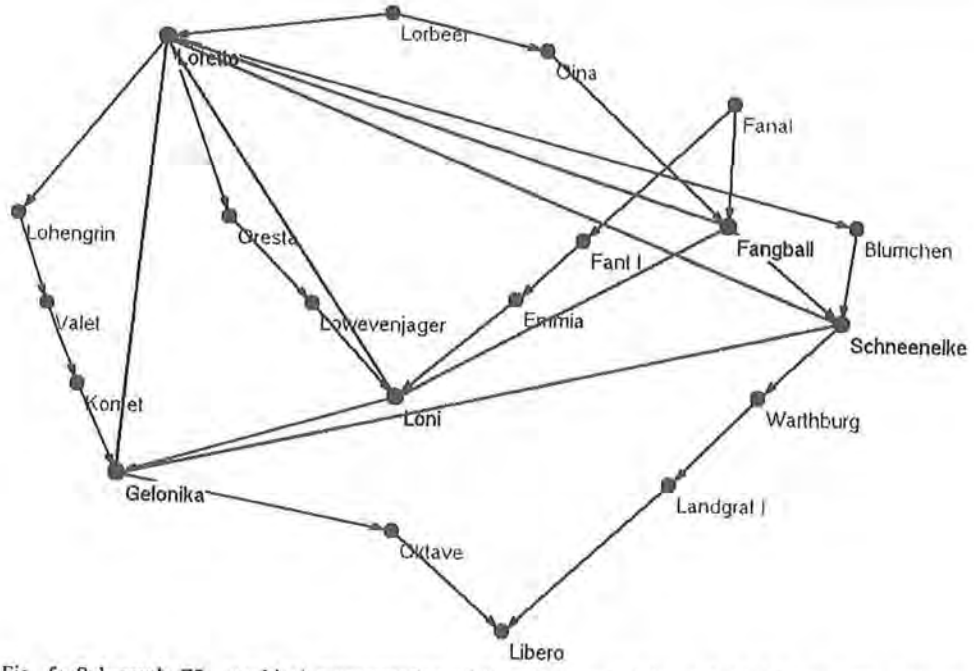

Fig. 5: Sub-graph E2: graphical representation of the clique containing animal Libero (lines without arcs represent indirect connections between animals)

In this paper, a new perspective to analyze animals' pedigrees, based on graph theory, is presented. Visual representation of pedigrees and the associated graph analysis has proved to be successful in detecting the internal structure of pedigrees of moderate size. Visualization of pedigrees has provided a useful tool for detecting animals' families as well as isolating single animals with major gene contributions to the

following generations. This implies a further possibility for better detection of bottleneck effects in a population. When the problem is the estimation of inbreeding coefficients of single animals, like best sires, graph analysis of pedigrees could also be used in this direction. The use of valued graphs, i.e. graphs with numbers attached to lines that indicate the strength or the frequency or the quantity of the ties between points could also facilitate the examination of the genetic relations between sub-divided or differentiated populations.

References

BATAGELJ, V.; MRVAR, A.:
Pajek-Program for Large Network Analysis. User's Notes, 2000

BORGATTI, S. P.; EVERETT, M. G.:
Notions of positions in social network analysis. In: Sociological Methodology, Ed. P.V. Marsden, San Francisco (1992): Jossey-bass, pp. 1-5

DOREIAN, P.; PARAGELJ, V.; FERLIGOJ, A.:
Partitioning networks based on generalized concepts of equivalence. Journal of Mathematical Sociology 19 (1994), 1-27

FREEMAN, L.:
Computer programs in social network analysis. Connections 11 (1988), 26-31

FOULDS, L.R.:
Graph theory applications, Springer Verlag, 1992

HARARY, F.:
Graph theory. Reading, Massachusetts Addison – Wesley, 1969

KAMADA, T.; KAWAI, S.:
An algorithm for drawing general undirected graphs. Information Processing Letters 31 (1989), 7-15

LORRAIN, F.; WHITE, H.C.:
Structural equivalence of individuals in social network analysis. Journal of Mathematical Sociology 1(1971), 49-80

SAILER, L.D.:
Structural equivalence: meaning and definition, computation and application. Social Networks 1 (1978), 73-90

SNYDER, D.; KICK, E. :
Structural position in a world system and economic growth 1955-70: a multiple network analysis of transnational transactions. American Journal of Sociology 84 (1979), 1096-1126

WASSERMAN, S.; FAUST, K.:
Social network analysis. New York: Cambridge University Press, 1994

Received: 2000-08-03

Accepted: 2001-06-20

Author's address
Dr. ANTONIOS P. KOMINAKIS
Department of Animal Breeding and Husbandry, Faculty of Animal Science,
Agricultural University of Athens,
Iera Odos 75, 11855, Athens, GR

E-Mail: acom@aua.gr