

A Case Study of Genealogical Networks from Network Science Perspective

Imre Varga^a

Department of IT Systems and Networks, University of Debrecen, 26 Kassai str., Debrecen, Hungary

Keywords: Genealogy, Networks Analysis, Pedigree Collapse, Social Network, Ancestral Network.

Abstract: In this paper, the analysis of a genealogical network is presented. The source database was constructed from the records of birth, marriage and death registers of a medium-sized Hungarian town covering some centuries. This genealogical network contains ca. 100.000 individuals. The topological features of this acyclic directed graph were analyzed by computer software in order to draw conclusions about the community. The results illustrate how network science can help the social sciences. A new measure is also defined to quantify the degree of pedigree collapse of a person having a partially known ancestor graph. The network was analyzed from the point of view of this ancestor-loss coefficient.

1 INTRODUCTION


Social networks, where different interactions of individuals are described by graphs, are broadly studied in the last decades. Different aspects were in the focus of scientific analysis for instance citation or co-authorship of scientists (Radicchi et al., 2012; Newman, 2004), sexual interactions (McDonald and Pizzari, 2017), membership of terrorist groups (Fellman, 2008), etc. Not just the real world social networks, but also online social networks are also investigated (Kumar et al., 2010; Howard, 2008). Besides their topological structure (Barabási, 2016), their dynamical properties and roles in spreading processes (Newman, 2010; Kocsis and Varga, 2014) are also examined empirically and theoretically.

Genealogy is an ancillary historical discipline. It means the study of family origins and history, and the tracing of their lineages. The word "genealogy" comes from two Greek words ("family" and "science"), thus is derived "to trace ancestry", the science of studying family history. Genealogists use historical records, genetic analysis, and other sources to get information about a family and to demonstrate ancestry and pedigrees of individuals. There are trials of computer-aided document processing, but this task is very complicated even for artificial intelligence (Malmi et al., 2017; Gellatly, 2015). In the broad sense, genealogy traces the descendants and the an-

cestors of one person. Genealogy research is performed for historical, scholarly, or forensic purposes as well. The results of such research are often presented in pedigree charts (BCG, 2019).

Family trees or ancestry charts are usually maintained as a binary tree data structure containing the ancestors of a person. In a simple assumption, everyone has 2 parents, 4 grandparents, 8 great-grandparents, 16 great-great-grandparents, and so on. Thus the number of ancestors in a given generation can be expressed by the powers of two. For example in the 30th generation theoretically, there are more than one billion people, which can be more than the total population of the Earth at that time. This conflict can be resolved by the fact that not all ancestors are unique. In genealogy, this phenomenon is called pedigree collapse (Wikipedia, 2020). It describes the situation caused by the reproduction between two individuals who share an ancestor. It is very rare in the short-term oral history of a family, but it is unavoidable in huge pedigree charts covering centuries. Due to pedigree collapse genealogists have to use graphs instead of tree data structures. It is quite frequent in royal families. A good example of pedigree collapse can be found in the ancestors of Charles II, the last Habsburg King of Spain. There were three uncle-niece marriages and three first cousins marriages besides other unions of his immediate ancestry. Between Habsburg Charles II and his ancestor Philip I of Castile, there are 14 different lineage relationships.

When not just a family, but a community is in the

^a  <https://orcid.org/0000-0003-3921-2521>

focus of genealogical research besides the size of the data source its structure also changes (Rannala, 1997; Kingman, 1982). Marriages and childbirths connect families (Koylu et al., 2021). Ancestry charts of a minor community cannot be represented by a forest of family trees, it is a general directed acyclic graph. In a small settlement especially in bygone years, the society was more closed than nowadays, thus families are densely interconnected. People who live in small communities often choose wives/husbands from the same community (villages, ethnic or religious minorities).

The "loss of lineage" (also called implex) can be characterized by a genealogical coefficient of a given genealogical tree, defined as the difference between the number of theoretical ancestors of a person and the number of his/her real ones in a given generation. For example, procreation between first cousins means 25% loss in the generation of great-grandparents of the offspring. This measure is not so useful in case of marriages between different generations or when some ancestors are unknown (Pattison, 2001; Pattison, 2007).

In genetic genealogy, DNA analysis can be used to show out pedigree collapse (Tetushkin, 2011; Vince Buffalo, 2016). Generally, children inherit 50% of their DNA each from their parents, 25% from their 4 grandparents, and so on. Nevertheless, the actual amount of DNA inherited is random, the average amount of DNA inherited from an individual ancestor is halved going back to each generation level. Due to random inheritance, DNA analysis is an effective way of finding shared ancestors only within few generations. Nevertheless, this kind of research is quite expensive and involved.

Our goal is to build a directed network of people based on only registry records (without genetic test results) and then determine different metrics of the networks (Newman, 2010; Barabási, 2016), such as in-degree and out-degree distribution, average clustering coefficient, size of the giant component, average path length, etc. In this system, they have the social meaning as well. The characterization of pedigree collapse also requires network analysis. While the dataset is not complete a novel quantity is defined to illustrate the scale of pedigree collapse.

2 METHOD OF INVESTIGATIONS

Our research is based on a public dataset created by a Hungarian genealogist (Szepesi, 2020). He processed the available (civil and parish) birth, marriage and

death registers of a town (Hajdúböszörmény, Hungary) and other historical documents (census, burial records, etc) of the archives. The database contains different data fields appeared in the registry records: an ID, the name, the date of birth, marriages and death of the given person, names (and IDs) of his/her parents and name (and ID) of his/her spouse(s), etc. More than 100.000 individuals appear in the dataset mainly (but not exclusively) from the last three centuries.

Of course, the dataset is not complete due to the nature of the problem and the accuracy of the sources. Each person has two parents, but the source is restricted in time and space. Too old ancestors and too young descendants are unknown and migration is also not followed. In the 18th century, just the fathers were represented in registers.

The IDs of people and his/her parents were extracted from a dataset having Personal Ancestral File format and used to build up the genealogical chart i.e. an acyclic directed graph of depersonalized IDs. (Those few people who do not belong to any other individuals are eliminated.) A special graph analyzer program (Bordán, 2019) and a web-application (Széll et al., 2020) were applied to analyze the topology of this special social network. However, only one community was investigated in this case study we believe that the results and conclusions may be general.

2.1 New Characterisation of Pedigree Collapse

As it was highlighted in Section 1, the pedigree collapse cannot be properly characterised by the simple loss of ancestors in a given generation. That is why we propose a new quantity to measure the degree of the pedigree collapse. First a kind of auxiliary measure α_j is assigned to the given person i and to his/her known ancestors according to a recursive definition. The $\alpha_j = 1$ for the given person, so where $j = i$. For ancestors the value of α_j is given by the following form

$$\alpha_j = \frac{1}{N} \sum_{k=1}^N \frac{\alpha_k}{2}, \quad (1)$$

where k runs over all the N children of person j who are ancestors of person i . An example is shown in Figure 1. It was motivated by the inheritance. However it is a random process, approximately half of the genome comes from the father and the other half comes from the mother.

In order to define the new ancestor-loss coefficient of a person i (denoted by λ_i) the summation of auxiliary measure is needed according to the following restrictions:

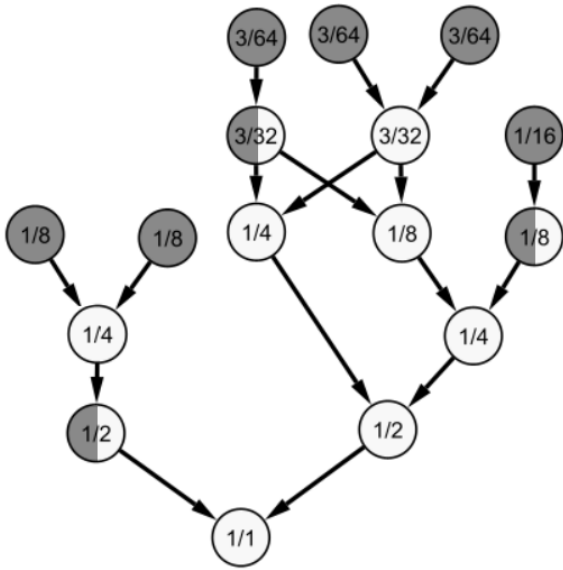


Figure 1: An example of auxiliary measures of known ancestors of the bottom person. The sum of values in gray nodes and the half values in gray-white (bicolour) nodes provides the ancestor-loss coefficient $\lambda = 0.1875$. It indicates incest, an uncle-niece marriage in the grandparent generation. (Colors represents the number of known ancestors).

- If ancestor j has not got known parents, his/her full auxiliary measure α_j is taken into account in the summation.
- If ancestor j has only one known parent, then the half of his/her α_j is added.
- If both parents of ancestor j are known, then his/her α_j value is disregarded.

Thus if the ancestor j of person i has P_j unknown parents ($P_j \in \{0, 1, 2\}$) then the λ_i ancestor-loss coefficient of a person i is defined as

$$\lambda_i = 1 - \sum_j P_j \frac{\alpha_j}{2}, \quad (2)$$

where j runs over all ancestors of person i . Since genome can origin from the starting points of lineages, that is why just the red and bicolor nodes of Fig. 1 are considered.

The λ can be interpreted as an extension of the common implex. In the case of simple situations (e.g. reproduction between first cousins, where all great-grandparents are known) $\lambda = A_R/2^i$, where A_R is the real identical ancestors in the i th ancestor generation. The definition of this quantity assumes that there is no pedigree collapse in the branches of unknown ancestors, thus λ coefficient determines just a maximum for the given person, in the case of more explored family history it can decrease. According to the definition $0 < \lambda < 1$. The $\lambda = 0$ indicates pure bloodline

in the investigated genealogical network. The $\lambda > 0$ implies the rate of incest (pedigree collapse). For instance, in the well-known case of Habsburg Charles II the $\lambda = 0.830295$.

3 RESULTS

It was found that our genealogical network contains 100.273 nodes (people) and 156.062 directed links (parent-child relations). If it would be a set of independent families we should see a forest of many trees of approximately the same size. Instead we found only 3840 independent clusters of the network, where most of them are relatively small, but there is a dominant cluster. This giant component contains the 85.3% of nodes (and 91.7% of links). In the remaining 14.7% of the system, most of clusters contain not more than 4 nodes ($S \leq 4$). The cluster size distribution is presented in Figure 2 excluding the giant component. As one can see it can be well fitted by a power-law form.

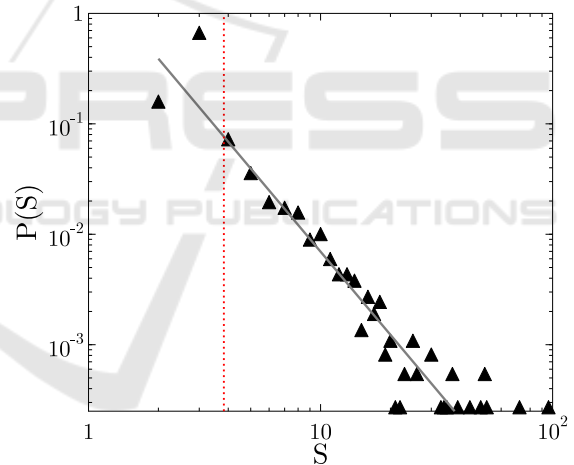


Figure 2: The cluster size distribution of the investigated genealogical network without the giant component. The solid gray line illustrates a power-law function with an exponent -2.5 . The dotted line refers to the average cluster size excluding the giant component.

In the case of small clusters, the available source documents probably did not contain enough information to identify individuals behind the registry records, so relationship was not found to other people. In the remaining part of the paper, the investigation is restricted only to the giant component, so the interconnected pedigree of 85536 people is in the focus of the study.

Since it is a directed graph the in-degree and out-degree distribution can be important. In the case of

genealogical networks, the in-degree k^{in} means the number of known parents of a person. In this sample, 74.8% of the population has 2 known parents and 6.0% has only one parent. (Old registers contain just the name of the father and in case of a bastard child just the name of mother is documented.) Nodes with $k^{in} = 0$ refer to individuals where the parents are not known most likely due to the missing documentation.

The out-degree k^{out} of a node denotes the number of known children of a person. (It must be mentioned that the real number of children can be greater than the number of known ones.) The out-degree distribution is presented in Figure 3. Results show that an average parent has 3 children, but someone has much more ($\max(k^{out}) = 20$). Almost half of nodes have not got outgoing edges ($k^{out} = 0$). This can be explained by several things. On one hand, young adults can migrate mainly due to their marriages. On the other hand, the child mortality was high in former times. Last, but not least the first childbirth was later than the last public documents.

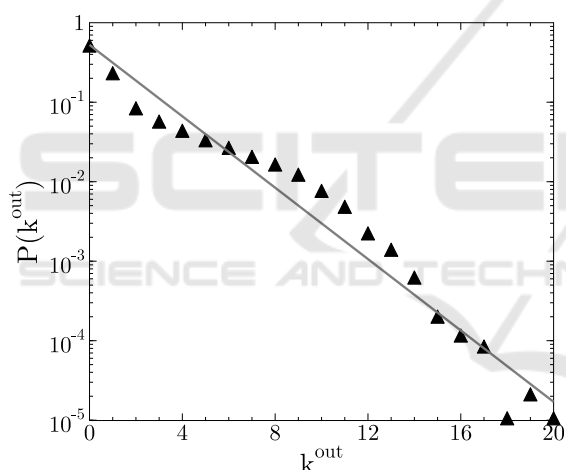


Figure 3: B) The out-degree distribution $P(k^{out})$ of the sample network. The solid gray line illustrates an exponential form on a semi-log scale.

There are two special subsets of the population. One of them includes people where the in-degree is $k^{in} = 0$. Since the age of the source documents is limited, probably they are the oldest people in the sample, they are the forefathers. The other group covers the $k^{out} = 0$ subset of nodes. They are either in the youngest generations or they are the end of lineages. From nodes of the former group to nodes of the latter one we can find multiple paths (along lineage). In order to characterize the network, these paths were discovered. These are the longest paths in that sense that there is no more known ancestor of the oldest person along the path and no more known descendant of

the youngest person along the path. The length distribution of these maximal paths is shown in Figure 4, while their average length is 6.46 generations.

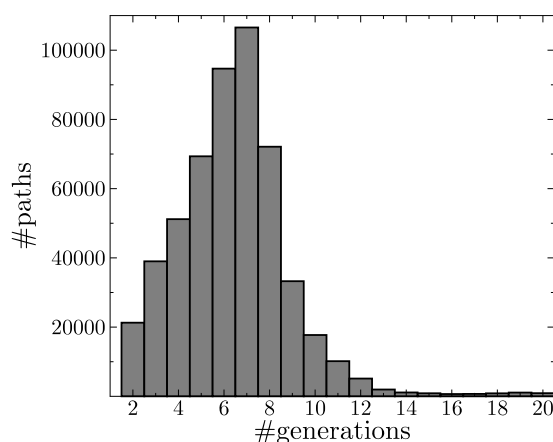


Figure 4: The number of maximal paths containing the given number of generations. As one can see the most paths cover 6-7 generations. Some of the lineages are much longer. These belong to nobleman families, because only they have so old documents (public registration started only in the 18th century).

The distribution of our ancestor-loss coefficient λ is shown in Figure 5. One can see that the majority of the people can be characterized by $\lambda = 0.0$, thus in case of them, it is not possible to figure out pedigree collapse based on the available registry records. It is consistent with the average ancestor-loss coefficient $\bar{\lambda} = 0.002986 \pm 0.021524$.

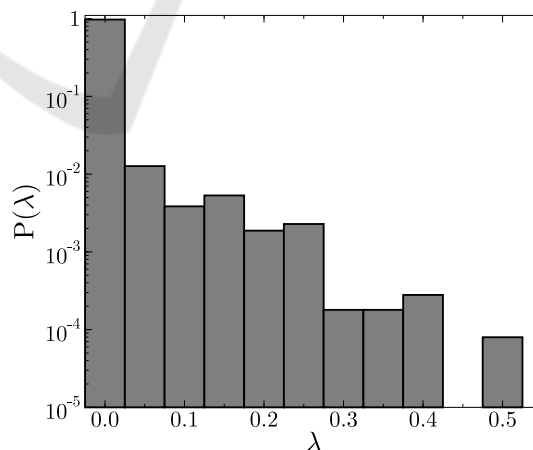


Figure 5: Distribution of the ancestor-loss coefficient λ . Pure bloodlines are very frequent, but there are some people with really low λ as well.

The most interesting finding of the research is the relatively large number of individuals affected by the pedigree collapse. In the studied population, 3943

people have an ancestor-loss coefficient greater than 0.0, thus they are available from another node of the graph along at least two distinct paths. It is the 3.93% of the investigated population, which is quite high if we consider that the average known lineages are only 6 or 7 generations long. At least 7.59% of ancestors of these people are lost, thus only 92.41% of the ancestors are unique in the last few generations. The lowest found λ was 0.5 in the case of 7 people (in 3 families). Some of them were children of a full siblings' marriage. If two brothers get married to two sisters and their children get also married (to each other) then the grandchild is also has $\lambda = 0.5$ coefficient. A real genealogical chart is illustrated in Figure 6 as an example of a significant pedigree collapse.

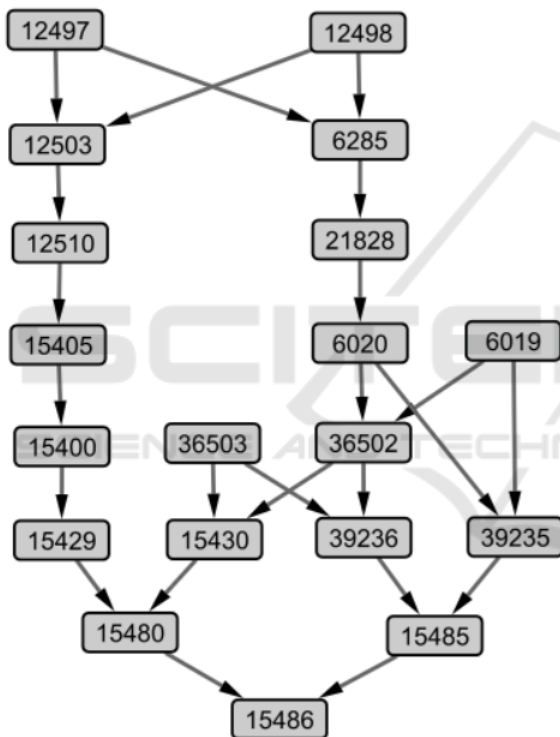


Figure 6: An example of the pedigree of a person with serious incest (pedigree collapse). The person represented by the bottom node (ID: 15486) has $\lambda = 0.475261$. However, his parents are not full siblings his ancestor-loss coefficient is close to 0.5.

In network science, the average clustering coefficient of the network is an important topological measure. The clustering coefficient gives the probability triangles, since it determines how often two neighbors of a node are also connected. In genealogical networks, this kind of "friend of my friend is my friend" situation refers to incest. In the investigated network, the average clustering coefficient $\langle C \rangle = 0.0$. It can be interpreted as total absence of procreation in father-

daughter or son-mother relationships.

4 CONCLUSIONS

In this work, a case study is presented in order to demonstrate how graph theory and computer science can be used in genealogical studies. A large dataset was created from records of birth, marriage and death (civil and parish) registers of a town. It contains the known parents of inhabitants covering a few centuries, thus a very complex genealogical network of 10^5 individuals serves as the object of analysis including several generations of many families. Naturally, it is not complete and full in the given time period because some missing or inaccurate records do not enable the full exploration of kinship. Nevertheless, the dataset is enough to discover huge interconnected pedigree charts. The graph analysis of them can provide interesting information for social sciences about the population (e.g. child number distribution).

We created an acyclic directed unweighted graph and then we investigated its features. The system is dominated by a giant component, other clusters are negligible. The out-degree distribution of nodes reflects the number of children in families. It can be roughly fitted by an exponential distribution. Due to the discovery of directed paths, we found that most lineages include 5 – 7 generations. We introduced a new quantity to measure the degree of pedigree collapse in a not complete ancestor chart. The distribution of this ancestor-loss coefficient shows the prevalence of incest within the given community even if the dataset covers just a bit more generations than the oral history of an average family. These results cannot be obtained without the tools of network science.

ACKNOWLEDGEMENTS

The author would like to express his sincere gratitude to Imre Szepesi for his valuable registry research and the creation of the genealogical database used (Szepesi, 2020). The author expresses great appreciation to Imre Bordán for technical assistance.

REFERENCES

- Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.
- BCG (2019). *Genealogy Standards*. Turner Publishing Company.

- Bordán, I. (2019). Genealógiai hálózatok számítógépes elemzése. Master's thesis, University of Debrecen, Faculty of Informatics.
- Fellman, P. V. (2008). The complexity of terrorist networks. In *Proc. of 12th International Conference Information Visualisation*, pages 338–340. IEEE.
- Gellatly, C. (2015). *Population Reconstruction*, chapter Reconstructing Historical Populations from Genealogical Data Files, pages 111–128. Springer.
- Howard, B. (2008). Analyzing online social networks. *Communications of the ACM*, 51(14-16):11.
- Kingman, J. F. C. (1982). On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43.
- Kocsis, G. and Varga, I. (2014). Investigation of spreading phenomena on social networks. *Infocommunications Journal*, 6(3):45–51.
- Koylu, C., Guo, D., Huang, Y., Kasakoff, A., and Grieve, J. (2021). Connecting family trees to construct a population-scale and longitudinal geo-social network for the u.s. *International Journal of Geographical Information Science*, 35(12):2380–2423.
- Kumar, R., Novak, J., and Tomkins, A. (2010). *Link Mining: Models, Algorithms, and Applications*, chapter Structure and Evolution of Online Social Networks, pages 337–357. Springer.
- Malmi, E., Rasa, M., and Gionis, A. (2017). Ancestryai: A tool for exploring computationally inferred family trees. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 257–261.
- McDonald, G. C. and Pizzari, T. (2017). Structure of sexual networks determines the operation of sexual selection. *PNAS*, 115(1):E53–E61.
- Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *PNAS*, 101(1):5200–5205.
- Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.
- Pattison, J. E. (2001). New method of estimating inbreeding in large semi-isolated populations with application to historic britain. *Homo*, 52(2):117–134.
- Pattison, J. E. (2007). Estimating inbreeding in large, semi-isolated populations: effects of varying generation lengths and of migration. *American Journal of Human Biology*, 19(4):495–510.
- Radicchi, F., Fortunato, S., and Vespignani, A. (2012). *Models of Science Dynamics*, chapter Citation Networks, pages 233–257. Springer.
- Rannala, B. (1997). Gene genealogy in a population of variable size. *Heredity*, 78:417–423.
- Szepesi, I. (2020). Hajdúböszörményi családardó. <https://gw.geneanet.org/szepesi>.
- Széll, M. C., Becsei, M., and Kocsis, G. (2020). Introduction to dina: An extendable web-application for directed network analysis. In *Proceedings of the 5th International Conference on Complexity, Future Information Systems and Risk*, pages 129–135. SciTePress.
- Tetushkin, E. (2011). Genetic aspects of genealogy. *Genetika*, 47(11):1451.
- Vince Buffalo, Stephen M. Mount, G. C. (2016). A genealogical look at shared ancestry on the x chromosome. *Genetics*, 204(1):57–75.
- Wikipedia (2020). Pedigree collapse. https://en.wikipedia.org/wiki/Pedigree_collapse.