The classical origin of modern mathematics

Gargiulo, Floriana; Auguste, Caen; Lambiotte, Renaud; Carletti, Timoteo

Published in:
EPJ Data Science

DOI:
10.1140/epjds/s13688-016-0088-y
10.1140/epjds/s13688-016-0088-y

Publication date:
2016

Document Version
Early version, also known as pre-print

RESEARCH

The classical origin of modern mathematics

Floriana Gargiulo[1*], Auguste Caen[2], Renaud Lambiotte[1] and Timoteo Carletti[1]

*Correspondence:
floriana.gargiulo@unamur.be
[1]NaXys, University of Namur, Belgium, 8 rempart de la Vierge, Namur, Belgium
Full list of author information is available at the end of the article

Abstract

This paper introduces a data-driven methodology to study the historical evolution of mathematical thinking and its spatial spreading. To do so, we have collected and integrated data from different online academic datasets. In its final form, the database includes a large number ($N \sim 200K$) of advisor-student relationships, with affiliations and keywords on their research topic, over several centuries, from the 14th century until today. We focus on two different issues, the evolving importance of countries and of the research disciplines over time. Moreover we study the database at three levels, its global statistics, the mesoscale networks connecting countries and disciplines, and the genealogical level.

Keywords: Academic genealogies; History of mathematics; Directed acyclic graphs

Introduction

The statistical analysis of scientific databases, including those of the American Physical Society, Scopus, the arXiv and ISI web of Knowledge, has become increasingly popular in the complex systems community in recent years. Important contributions include the development of appropriate scientometric measures to evaluate the scientific impact of scholars, journals and academic institutions [13, 11, 12, 14, 15] and to predict the future success of authors [18, 19] and papers [20]. In parallel, the structure of collaboration has attracted much attention, and collaboration networks have become a central example for the study of complex networks, thanks to the high quality and availability of the datasets [10]. From a dynamical point of view, different papers [17, 16] studied the mobility of researchers during their academic career, showing that the statistical properties of their mobility patterns are mainly determined by simple features, such as geographical distance, university rankings and cultural similarity.

Limitations of the aforementioned datasets include their relatively narrow time window extension, at best, over 100 years and the difficulty to disambiguate author names, and thus to correctly distinguish career paths across time. The original motivation of this paper was to address these issues by performing an extended study of *The Mathematics Genealogy Project*, a very large, curated genealogical academic corpus [21]. The dataset, whose basic statistics have been already analysed elsewhere [4, 7], extends over several centuries and contains pieces of information allowing us to retrieve the direct genealogical mentor-student links, but also university affiliations at different points of a career and and the research domains. Data from the same website have already been used to assess the role of mentorship on scientific productivity [5] and to study the prestige of university departments [6].

Our main goal is to analyse the history of modern mathematics, through the processes of birth, death, fusion and fission of research fields across time and space. In particular, we focus on the temporal evolution of the roles and importance of countries and of disciplines, on the structure of "scientific families" and on the impact of genealogy on the development of scientific paradigms. As it is often the case when performing a data-driven analysis of historical facts [8, 9], the data set is expected to be incomplete and to present biases, mainly for the more ancient data. In the present case, the website collects the data in two ways: a participative method, based on the spontaneous registration of scholars (who can also register their students and their mentors), and a curated method, based on historical facts and performed by the creators of the web site.
The presence of biases calls for the use of appropriate statistical measures, in preference based on ranking instead of absolute measures. In this work, we have also introduced data-mining methods to correct and enrich the data structure. A first contribution of this work is thus methodological, with the design of a methodological setup that could be applied to other systems. We have then performed an analysis of the system at three levels of granularity. First, a global one investigates the fully aggregated "demography" (population in terms of countries and disciplines) of the database, with the aim to classify countries and disciplines according to their normalised activity behaviour. Tracking the evolution of the rankings helps identify transition points in the mathematical history, associated to emerging fields of research. Second, we have constructed directed weighted networks where nodes are scholars endowed with a set of attributes (thesis defence date, thesis defence location, thesis disciplines) and linked to other nodes using the genealogy associated to the mentor-student relation. This "mesoscale" network allows us to investigate the relationships between the attributes and to identify a strong hierarchical structure in the scientific production in terms of countries as well as its evolution in the course of time. Finally, using an approach typic of kinship networks studies [24, 25], we focus on the statistical properties of the tree structure of the genealogy in terms of family structures. We conclude by showing the presence of strong memory effects in the network morphogenesis.

To summarise this paper has a twofold goal: first to propose, in the framework of data science, new tools to collect and analyse historical databases, second and in complement with the former, to provide a narrative on the history of mathematics as extracted from data. In both cases, this work opens interesting perspectives. Because of their generality, the presented tools could clearly be used to study different databases with a genealogical structure, for instance in the case of bibliometrics or Wikipedia studies. In addition, our results provide a first glimpse of the potential use of data and algorithms in the study of the history of science. An important future step would consist in complementing and interpreting this data-driven view with that of epistemologists and historians of science, as briefly outlined in the conclusions.

Dataset and associated networks

The core of our dataset has been extracted from the website "Mathematical Genealogy Project". It is one of the largest academic genealogy available on the web, consisting of approximatively 200K not–isolated scientists (186505) with information

on their mentors and students. The data cover a period between the 14th century until nowadays. For a majority of mathematicians, we have detailed information about his/her PhD, including the title (for 88% of the scholars), the classification according to the 93 classes proposed by the American Mathematical Society [1] (for 43% of the scholars), the University delivering the degree as well as the year of its defence. However, because a large part of the database is spontaneously filled by the scientists, the data is imperfect and attributes may be wrong or missing. A first step has thus consisted in comparing the database with additional data form Wikipedia [2]. In particular, we first downloaded, when available, the Wikipedia pages of all the scholars present in the Mathematics Genealogy Project database. Disambiguation of the names is assured by the fact that Wikipedia pages have a direct link to the Mathematics Genealogy Project site. In the text of the Wikipedia pages, we then searched the keywords associated to the AMS classification in order to expand the information about authors. This external dataset allowed to assign a discipline to the 54% of the mathematicians.
For more recent entries, we retrieved the affiliations with the Scopus profiles of scientists [3]. Notice that we only extracted the information required for our needs and that additional information, e.g. about their scientific impact or on their geographical links, could be collected in order to address other research questions.

It is worth noting that our analysis are biased by the actual scientific and socio-political environment. First, the countries' borders changed in time. In the Mathematics Genealogy Project, the location of the PhD defence is determined according to the position of the university in the current geo-political setting. We kept in our analyses this county classification, but it would be interesting in the future to consider, for example, the resilience of the system to borders shifts. Similarly, the concept of discipline is also very delicate to define on such a long time scale [32], and we decided here to use the current classification from the AMS for all authors.

After this preliminary phase, we have enriched the information available for the authors, by developing algorithms aimed at correcting the dates and assigning to each thesis a discipline. The algorithm for fixing errors in temporal entries is based on the topological structure of the genealogical network and uses the available statistics on the age difference mentor-student to identify and suitably correct wrong time sequences (e.g. the cases where the mentor has completed its PhD after its student, or where the time distance between mentor's and the student's PhD is too large). The missing disciplines (not previously extracted from Wikipedia) have been learned based on the thesis title using a bayesian supervised dictionary learning technique.

As previously stated, these algorithms, summarised in the supplementary material (sections I.B-D), are general and could be applied in other contexts. After the enrichment, all the scholars of the database have a corrected date, 88% of them have an associated discipline and 94% an associated country. As a next step, we have exploited the enriched database in order to study the geographical and temporal evolution of mathematics. Different data representations, described below, have been adopted to mine different typologies of information from the dataset.

The mesoscale networks

As a first step, we built a multi-partite network, where the different kinds of nodes are scholars and their attributes extracted from our database, namely universities, cities, countries and disciplines. Hence we considered several possible "projections"; for instance considering the case of universities, we add a direct link between two universities-nodes, say A and B, if there exists in the database, and hence in the multi-partite network, a couple of scientists such that the one with attribute "university A" has been the PhD supervisor of the second one having attribute "university B". Observe that our data have time stamps, corresponding to the time of the defence of the PhD thesis, then under the assumption that a supervisor is in the same university as his/her PhD student, directed links are therefore a proxy for the mobility in time of a scholar, but also of the flow of knowledge between different places. In the case of disciplines, directed links correspond to a transfers of knowledge from one scientific discipline (the one of the mentor) to another one (the one of the student), from one generation to another, and how disciplines at a certain time may inherit, in terms of ideas and methods, from research fields at previous times. Let us note that the network has a large number of self-loops associated to the frequent situation when the supervisor and the student got their PhD in the same university or in the same discipline. Moreover the links are weighted by the number of researchers connecting attributes. The procedure is illustrated, for the case of flows between countries, in Fig 1. In the following, we will perform a longitudinal study of the system, by considering the evolution of networks observed in different time windows. Note here that the data are not uniformly distributed across time, with a strong bias towards recent times.

The genealogical tree and its partitions into families

The genealogical graph is the most obvious representation of our dataset, consisting in an oriented acyclic graph [28] linking a mentor to her/his students. This defines automatically the structure of hierarchical generations. Notice however that the structure of our data is not simply a tree due to the several cases where a student has two advisors. A very common process in kinship is to cut the genealogical directed acyclic graphs into linear trees (alliances) where each individual has a single progenitor (the mother for representing the uterine links and the father for the agnatic ones). In this representation, the links between alliances represent the matrimonial structures between the different alliances in the society. In our context, when a scientist has more than one advisor, it is not clear which links should be cut to retrieve the original ancestors (our dataset prevents us from identifying the principal supervisor from the secondary one, if any). We thus propose a method to reproduce the optimal ancestry lines and to identify the important families in the genealogy. The method, fully described in the supplementary material (section I.D), is based on the decomposition of the network into pure linear trees, and their statistical clustering based on probabilistic arguments; roughly speaking given two nodes A and B that can be linked in more than one way, thus implying the presence of non-trivial loops, we assign to every link in such paths the probability that A and B will be disconnected if the link is removed. We thus select links to be removed by maximising the probability that A and B are still linked. The resulting partition

of the graph into families identifies 84 families; remarkably, the 24 most populated families cover the 65% of the scientific population in the database. Let us observe that alternative methods for family identification do exist, see for instance [26, 27].

Results

Global statistics

Let us define the *relative abundance profile* of each country in different periods $f_I(t) = N_I(t)/N(t)$, where $N(t)$ is the total number of scientists whose country is known in the database at time t and $N_I(t)$ is the number of scientists in country I at time t. Notice that, at this stage, the genealogical information is not used. From such profiles we can asses the evolution of the importance of the countries had in the history of mathematics due to the their different historical dynamics. In order to compare the profiles of different countries independently from their total human capital, we normalise each of these profiles by their "volume"($\tilde{f}_I(t) = f_I(t)/\sum_t f_I(t)$) and then we classify them based on their Kolmogorov-Smirnov distance (see Fig. 2 where the results are reported using a dendrogram, and the SI where we reported the profiles for the top 10 countries in the database). We observe different prototypical behaviours: countries with a central role in the ancient history whose centrality has decreased in the last centuries (for instance Italy, France and Greece), countries with a central role before the world wars (e.g. central Europe countries), countries emerging after the world wars (such as Japan and India), countries recently emerging (among which China and Brazil). Because of the normalisation procedure we used, we obtain a cluster where USA is linked with ex-USSR countries and show a similar decreasing behaviour in the latest decades (impossible to observed in a non–normalised context).
Additional information on how the total number of mathematicians compares with that of scientists would make these results more significative, but this type of information is difficult to be retrieved in electronic archives.

The same procedure is applied to disciplines and results reported in Fig. 3 allow to identify three main blocks of disciplines: the disciplines that were more central during the industrial revolution (before 1900) are associated to physical applications (such as thermodynamics, mechanics and electromagnetism). The disciplines reaching their maximum of expansion around the 1950 are more abstract, even if several links exist to applied topics, such as telecommunication and quantum physics. Finally, the last decades have witnessed the emerging dominance of applied mathematics (e.g. statistics, probability) and computer science. This last point shows the considerable impact of the computer revolution on the evolution of mathematics: the magenta subfield in Fig. 3 (Operations research; systems theory; category theory; computer science) has emerged at the expense of many other fields, but the yellow one, which may have also been helped by the introduction of computers.
To capture the rise and fall of countries or disciplines, we have compared the rankings of the top 10 countries and disciplines in different time periods. Standard indicators for rank comparison, such as the Kendall-Tau index, cannot be applied here since the elements in the top-k lists are not conserved in time [22]. For this reason, we have used a distance measure based on a modified version of the Jaccard index allowing to compare ranked sets, $J(rank_1, rank_2)$ (more information is provided in

the Supplementary material, section A). As for the original Jaccard index the modified version is such that $J(rank_1, rank_2) = 1$ when the rankings $rank_1$ and $rank_2$ are completely equivalent, and gives a value 0 when these latter are not correlated at all. This information is then transformed in a distance by taking $d_J = 1 - J$. Increases in the distance measure, d_J, indicate major reshaping of the rankings.

As we can observe in the upper plot of Fig. 4, corresponding to countries, we observe several transition points; for example a transition can be observed during the first World War, with the decreasing centrality of Austria and Hungary due to the end of the Austro-Hungarian emperor and the entering in the ranking of Russia. Another transition is connected with the European political reshaping during the second World War and with the massive migration of jewish and dissident scientists to US due to fascism in Europe. This is the period at which for the first time, USA surpasses Germany in the ranking. A third transition, around the 1960s, shows the increase of centrality of the Soviet Union (testified by the presence of several east-european countries in the ranking). Finally, more recently, we observe the decline of Russia and the emergence of new countries such as Brazil.

A similar analysis can be performed for disciplines. Results reported in Fig. 5 show the presence, among the others, of three significative tipping points. The first one is connected to industrial revolution and to the emergence of disciplines related to the physics of machines (such as thermodynamics and electromagnetism). The second one is connected to the emergence of fields linked to telecommunication and cryptography (e.g. number theory, spectral functions) during the second World War period. Finally the third one, in the 80's, concerns the emergence of computer science and statistics.

Mesoscale networks

Network of countries

The countries network can be used to represent the *knowledge flows* from one country to another one, associated to the transition of a student in a country, becoming a professor and PhD supervisor in another country. The network presents few important hubs, that are the gravity centres of the scientific research (USA, Germany, Russia, UK). Each of these hubs tends to be surrounded by a community of countries. These communities can be associated to historical divisions, for instance a large block connected to USA scientific production, the Commonwealth nations, the ex-Soviet block, the central European countries. The betweenness of countries allows to detect countries at the interface between different communities, such as France connecting the central European countries with the USA-centred community or Poland connecting European research and the ex-Soviet area.

Another important index of the countries network is the weighted *in(out)*–degree of the nodes and the number of self loops. The *in*–degree of a country represents the number of scientists obtaining their PhD elsewhere and mentoring a PhD student in that country. Therefore a high *in*–degree is associated to a country with a strong capacity of attracting scholars and absorbing knowledge from abroad. On the contrary, a high *out*–degree represents a country producing scholars and exporting knowledge elsewhere. Each country can be therefore characterised by three

normalised quantities, the fraction of scientist formed inside the country and remaining there, the fraction of scientist formed inside and leaving and the fraction formed abroad and absorbed by the country. In Fig. 6A we display the different positions of the most important countries with respect to these indexes, the closer a country is to one of the triangle vertices the larger is the associated index. The size of the dot is a measure of the production of a country, i.e. estimated by its number of PhDs. The most productive countries tend to be the most *scientifically autarchic* ones, with a large fraction of self-loops. The most important *exporters* are Russia and the UK. Countries with small scientific production show a tendency for *importing* scholars. Observe that these indexes evolve in time (see Fig. 3 of the SI). An important inversion point between $k_{in}(t)$ and the $k_{out}(t)$ is often observed around the second World War. Moreover, a key signature of emerging scientific countries seems to be the presence of $k_{in}(t) > k_{out}(t)$.

To characterize the mobility of scientists across countries, we show in Fig. 6B the fraction of the total production and the total absorption of *migrant* scientists for the first r countries respectively in the *in* and *out*–degree ranking. The distribution is highly skewed, as one observe from the fact that the top 7 countries produce 80% of the total international scholars. On the contrary, the curve concerning the total absorption has a lower slope, depicting a larger worldwide spreading around the world. This shows a strong hierarchical structure in academic research where few countries ensure a large share of the worldwide diffusion of scientific knowledge. This scenario obviously evolved in time, as we observe in Fig. 7. Remark that scientific leaders changed at different points in time, but also that the scientific leadership group (countries producing the 80% of the whole scientists production) is more restricted in recent times. It is interesting to notice that the minimal size of the scientific *elite* has been reached in the sixties during the world bi–polarisation resulting from the cold war. Since then, the size increased again with the emergence of globalisation.

More information about this network, in particular the properties of the aggregated transition networks concerning the whole historical period, can be found in the Supplementary Information (section III).

The transition network of disciplines

The transition network of disciplines represents transfers of knowledge from one scientific discipline (the one of the mentor) to another one (the one of the student). The structure of this graph is quite homogeneous in terms of degree and four major topological communities can be identified using standard community detection algorithms working on the topological structure of the weighted network [35]: computer science, geometry, analysis and physics. Each community represents the disciplines exchanging more knowledge between them than with other research fields, and therefore can be interpreted as the scientific paradigms (according to Thomas Kuhn definition) at a certain period.

In Fig. 8 we show the normalised mutual information (NMI) between the community structures obtained from different temporal slices of the network. The NMI

index varies between one, when the two partitions are equal, and zero, when the two classifications are completely disjoined. A low value of the NMI indicates a "revolution" in the sense of a strong reorganisation of the knowledge structures, previously non interacting research fields start to exchange knowledge. The figure shows two important points where the NMI is low. The first transition, observed between 1930 and 1940, can be associated to the period when Statistics and Probability merged together, attracting then more applied disciplines like information theory, game theory and statistical mechanics, and leading to the emergence of the field of applied mathematics. The second transition is between 1970 and 1980, where computer science and statistics form one community, together with dynamical systems and applications in other fields of science. The latest transition is expected to be a spurious effect, due to a lack of data in recent years (the last time window starts in 2010 and therefore can contain data only for 5 years). Another potential approach, alternative to the measure of the NMI, to identify the structural changes in these structure could be the one proposed in [29].

The genealogical structure

This last section is devoted to the study of the genealogy tree reconstructed from our data and of its relevance in the evolution of the history of mathematical science.

The first result is the presence of a strong memory effects in the network morphogenesis, as students very often do research in the same discipline their mentor did. To quantify this idea we analysed the genealogical chains where the "filiation" link connects a mentor with a student maintaining the same discipline of the mentor. We call these objects iso-discipline chains. Let us observe that our analysis is data driven and that we only have information about "filiations" present in the Mathematics Genealogy Project. In Fig. 9 (left panel) we show results concerning the conditional probability of having a chain of length $n+1$ given a chain of length n, in other words the probability to have one more descendant working in the same research field of the whole chain, aggregating data over all the disciplines. The first point thus represents the probability for a student to have the same discipline of his/her mentor, one can clearly appreciate that as the chains get longer the probability to continue the same iso-discipline increases. This very marked memory effect in the network can be associated to the existence of "schools" where a long tradition in a discipline exists such that new students are attracted and continue the tradition. Observe however (Fig. 9 right plot) that this phenomenon strongly depend on the discipline .

As previously explained, we have partitioned the network into disjoint families of scholars. Fig.10A shows that the 65% of the scientists can be divided into 24 macroscopic families with size $S > 500$. The largest family is the one originated in 1415 by the Italian medical doctor, Sigismondo Policastro. The second one, is the family originated by the Russian mathematician Ivan Petrovich Dolby, at the end of the 19th century. The large size of this family, born more recently than other families and geographically located mostly around Russia, is due to a high "fecundity rate" in the Russian school of Mathematics.

The aggregated network between the families, reminiscent of kinship of the alliance networks defined in [23, 24], can be described using some typical topological indicators [25]: 1) the endogamy index, ϵ_0 describing the fraction of loops in the network (links between the same family); 2) The concentration index c_x denoting the heterogeneity of the concentration of links between pairs of families ($c_x = 1$ when all links are concentrated on a single pair and $c_x = 1/n^2$ when links are homogeneously distributed among the n families); 3) the network symmetry index s_x that varies from 0 in case of total link unbalance, namely the outgoing flux and the ingoing one are very different each other, to 1 in case of perfect symmetry of fluxes. To asses the relevance of such indicators computed for our genealogy network, we compared them with the expected values for a random multinomial reshuffling - null model - (see Fig. 10B), we can observe that, while the symmetry is a structural property, being unchanged by the reshuffling, the endogamy and the concentration are typical signatures of this network and moreover they are much higher than in traditional kinship networks [25]. These results imply that the obtained scientific families are structurally very distant between them and that their relationships are very hierarchical (being these mediated by the largest families).
This strong separation between the genealogical families can be a signature of the existence of tacit knowledge in mathematics [36]. It would be interesting to study the historical development of the kinship structure in order to better address this phenomenon.

Finally, we studied the distribution of families across countries and disciplines. As shown in Fig. 11A, with the exception of few cases, the most important countries (in term of production) are present in all the families, while the remaining countries are represented in a very low number of families (from 1 to 3). This feature implies a strong correlation between the genealogical structures and the geography. A similar behaviour can be observed for disciplines (Fig. 11B) even if, in this case, the curve describing the number of families with members working in a given discipline is smoother. We can therefore conclude that the genealogical families are strongly specialised in terms of geography and epistemic content.

Conclusions

In this paper, we have presented a data-driven study of the history of mathematical science, based on the Mathematical Genealogy Project. A first important aspect has been the cleaning and correction of the incomplete and sometimes inaccurate dataset. This operation was performed by means of machine-learning and by incorporating data from other sources, including Wikipedia.

We have then considered three different approaches to analyse the data: a demographic approach analysing the time evolution of the prevalence of certain attributes (i.e. country or disciplines); a mesoscale network approach focusing on the connections between these attributes; a "kinship" approach based on the clustering of genealogical trees. Our analysis reveals important transition points in the history of mathematics and allows us to categorise countries according to their capacity to attract, export and self-maintain knowledge. Moreover, the community structures of the network of disciplines allows us to better describe the transformation of knowledge across time. Finally, we have also identified important scientific families,

associating them to their founder, and described their geographical and disciplinary distribution.

Interesting lines of research for the future include the integration of additional datasets, based on different methodologies, to extend the scope of this work beyond the mathematical sciences.

Another research direction still connected to history of mathematics, could be to analyse how the scientific labor market reacts to exogenous events [30, 31] or to study the innovation dynamics due for instance to the impact of the computer age, of the internet, of the peer review practices, etc. in the disciplinary prevalence.

Finally it would be worth also to build an abstract agent based models of innovation diffusion, that could be calibrated and implemented on this framework, in order to forecast future events and thus to add a predictive character to this dataset.

Other interesting research directions could include the analysis of gender roles in scientific production, using methods similar to the ones proposed in [33, 34].

Competing interests
The authors declare that they have no competing interests.

Author's contributions
F.G. collected the data and performed the analyses. A.C. developed the pre-treatment algorithms for enriching the database. F.G., T.C. and R.L. wrote the paper.

Acknowledgements
The work of F.G., T.C. and R.L. presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office.

Author details
[1]NaXys, University of Namur, Belgium, 8 rempart de la Vierge, Namur, Belgium. [2]DICE, Inria, Lyon, France, 15 parvis René Descartes, Lyon, France.

References

1. http://www.ams.org/msc/msc2010.html
2. http://www.wikipedia.org
3. http://www.scopus.com
4. $http://people.maths.ox.ac.uk/porterm/research/priya\_thesis\_final.pdf$
5. Dean MR, Ottino JM, Nunes Amaral LA (2010) The role of mentorship in protégé performance. Nature 465.7298: 622-626.
6. Myers SA, Mucha PJ, Porter MA (2011) Mathematical genealogy and department prestige. Chaos-Woodbury 21.4: 041104.
7. Engin A, Gunes MH, Yuksel M (2011) Analysis of academic ties: a case study of mathematics genealogy. GLOBECOM Workshops (GC Wkshps), 2011 IEEE. IEEE.
8. Schich M, Song C, Ahn YY, Mirsky A, Martino M, Barabási AL, Helbing D (2014). A network framework of cultural history. Science, 345(6196), 558-562.
9. Sinatra R, Deville P, Szell M, Wang D, Barabási AL (2015). A century of physics. Nature Physics, 11(10), 791-796.
10. Newman ME (2001) The structure of scientific collaboration networks. Proc. Natl. Acad. Sci. USA 98
11. Ramasco JJ, Dorogovtsev SN, Pastor-Satorras R (2004) Self-organization of collaboration networks. Phis. Rev. E 70, 036106
12. Pan RK, Kaski K, Fortunato S (2012) World citation and collaboration networks: uncovering the role of geography in science. Sci. Rep. 2
13. Bergstrom C T, West JD, Wiseman MA (2008) The Eigenfactor metrics. J. Neurosci. 28.45
14. Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientisandts. Phis. Rev. E 80, 056103
15. Radicchi F, Castellano C (2011) Rescaling citations of publications in physics. Phis. Rev. E 83, 046116
16. Gargiulo F, Carletti T (2014) Driving forces of researchers mobility. Sci. Rep. 4, 4860
17. Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabási AL (2014). Career on the move: Geography, stratification, and scientific impact. Scientific reports, 4
18. Wang D, Song C, Barabási AL (2013). Quantifying long-term scientific impact. Science, 342(6154), 127-132.
19. Acuna DE, Allesina S, Kording KP (2012) Future impact: Predicting scientific success. Nature 489.7415 , 201-202.
20. Shen HW, Wang D, Song C, Barabási AL (2014) Modeling and predicting popularity dynamics via reinforced poisson processes. In AAAI 2014, 291-297.
21. The mathematical genealogy: http://genealogy.math.ndsu.nodak.edu/index.php

22. Fagin R, Kumar R, Sivakumar D (2003). Comparing top k lists. SIAM Journal on Discrete Mathematics, 17(1), 134-160.
23. Hamberger K, Houseman M, White D (2011) Kinship network analysis. In: Carrington, P., Scott, J. (Eds.), The Sage Handbook of Social Network Analysis. Sage Publications, London, pp. 533–549.
24. White D, Jorion P (1992) Representing and analyzing kinship: a new approach. Current Anthropology 33, 454–462.
25. Roth C, Gargiulo F, Bringé A, Hamberger K (2013). Random alliance networks. Social Networks, 35(3), 394-405.
26. Karloff H, Shirley KE (2013). Maximum entropy summary trees. In Computer Graphics Forum (Vol. 32, No. 3pt1, pp. 71-80). Blackwell Publishing Ltd.
27. Clough JR, Gollings J, Loach TV, Evans TS (2015). Transitive reduction of citation networks. J. Complex Netw. 3(2), 189-203.
28. Karrer B, and Newman M (2009). Random acyclic networks. Physical review letters 102(12)
29. Rosvall M, Bergstrom CT (2010). Mapping change in large networks. PloS one, 5(1), e8694.
30. Borjas GJ and Doran KB (2012), The collapse of the Soviet Union and the productivity of American mathematicians. Q. J. of Econ. 127: 1143-1203
31. Moser P, Voena A, and WaldingerF (2014). German-Jewish Emigres and US invention. American Economic Review 104: 3222-3255
32. Sugimoto CR and Weingart S (2015). The kaleidoscope of disciplinarity. Journal of Documentation, 71(4), 775-794.
33. Larivière V, Ni C, Gingras Y, Cronin B, Sugimoto CR (2013). Bibliometrics: Global gender disparities in science. Nature, 504(7479), 211-213
34. King MM, Bergstrom CT, Correll SJ, Jacquet J, West JD (2016). Men set their own cites high: Gender and self-citation across fields and over time. arXiv:1607.00376.
35. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008). Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10), P10008.
36. Polanyi M. (1969) Personal knowledge. Towards a post-critical philosophy. Chicago University Press

Figures

Figure 1 Mesoscale networks An example of the procedure to derive the mesoscale network from the genealogical data.

Figure 2 Country prevalence dendrogram Clustering of the countries according to their time prevalence profile $\tilde{f}_I(t)$. The lines in the plots on the right describe the average prevalence profile for the cluster $\langle \tilde{f}_C(t) \rangle = \sum_{I \in C} \tilde{f}_I(t) / (\sum_{I \in C} 1)$. The shadowed area is included between the minimum value and the maximum value of the prevalence profile on the cluster $(\min_{I \in C} \tilde{f}_I(t), \max_{I \in C} \tilde{f}_I(t))$

Figure 3 Discipline prevalence dendrogram Clustering of the disciplines according to their time prevalence profile $\tilde{f}_I(t)$. The lines in the plots on the right describe the average prevalence profile for the cluster $\langle \tilde{f}_C(t) \rangle = \sum_{I \in C} \tilde{f}_I(t) / (\sum_{I \in C} 1)$. The shadowed area is included between the minimum value and the maximum value of the prevalence profile on the cluster $(\min_{I \in C} \tilde{f}_I(t), \max_{I \in C} \tilde{f}_I(t))$

Figure 4 Countries' centrality tipping points Modified Kendall-Tau index comparing the countries' rankings in different periods.

Figure 5 Disciplines' centrality tipping points Modified Kendall-Tau index comparing the disciplines' rankings in different periods.

Figure 6 Panel A: Relative position of the countries between scientifically autarchic/exporting and absorbing behaviours. Panel B: Fraction of scientists produced and absorbed from the first countries in the rankings respectively of k_{in} and k_{out}. In the boxes are displayed the countries producing and absorbing the 80% of scientists. Both the panels concern the temporal aggregate of the network.

Figure 7 Temporal network. Fraction of countries producing (and absorbing) the 80% of the scientists in different historical periods.

Figure 8 Mutual information between the communities structures of the discipline network in different historical periods.

Figure 9 Conditional probability of having an iso–discipline chain of length $n+1$, having a chain of length n. Left panel: aggregated data all disciplines together; right panel: some selected disciplines.

Figure 10 Panel A: Relative size of the different families and family's initiator name. Panel B: Table with the values of the topological indicators for the real (observed) network in the first row and for the randomised model (expected) in the second row.

Figure 11 Panel A: countries are set on the horizontal axis, ranked by the relative presence in the database, while in the vertical axis, we report the families ranked by their size. A point at the intersection of the country-column and family-row indicates that the country is present in this family. The upper plot, is the column-marginal of the matrix represented in the central plot, representing the number of families where each country appears. The right plot is the row-marginal of the matrix represented in the central plot, representing the number of countries present in each family. Panel B: On the horizontal axis we put the disciplines, ranked by the relative presence in the database, in the vertical axis, the families ranked by the size. A point at the intersection of the discipline-column and family-row indicates that the discipline is present in that family. The upper plot, is the column-marginal of the matrix represented in the central plot, representing the number of families where each discipline appears. The right plot is the row-marginal of the matrix represented in the central plot, representing the number of disciplines developed in each family.