Supplementary Material. The historical origin of modern science

F. Gargiulo[1], A.Caen[2], R. Lambiotte[1],T. Carletti[1]
1. Department of Mathematics and Namur Center for Complex Systems - naXys, University of Namur, rempart de la Vierge 8, B 5000 Namur, Belgium
2. DICE, Inria, Lyon, France

I. METHODS

In this document we describe the details of the statistical methods and the algorithms used in the main text.

A. Rank comparison

The Jaccard index is a statistical measure used to define the similarity between two unordered sets:

$$J(A,B) = \frac{A \bigcap B}{A \bigcup B}\,. \tag{1}$$

It is a measure simply based on the content of the set, returning the fraction of elements that are common between two sets with respect to the total number of involved elements. To measure the distance between ordered sets, on the other side it is often used the Kendall-Tau distance, that is a measure of the number of swaps one should apply between two classifications of the same set of elements to obtain the same order. In our case, the rankings in different time windows can contain (and it is usually the case) different sets of elements: a country, for example, can enter the ranking only at a certain time, being absent before. Therefore, in the present case the Kendall-Tau index cannot be a suitable solution, nor the Jaccard one because we want to preserve the importance of the ranking.
For this reason we introduce a new index for comparing rankings with unequal entries, that we define "extended Jaccard", $\tilde{J}$ and that is based on the Jaccard index. For each ordered classification r_A, of length L, we define the unordered extended set $\tilde{A}$, of cardinality $\tilde{L} = \sum_{i=1}^{L} i = L(L+1)/2$:

$$r_A = \{1:a, 2:b, 3:c, 4:d\} \rightarrow \tilde{A} = \{a,a,a,a,b,b,b,c,c,d\} \tag{2}$$

namely the set where each element i of the ranking appears exactly $(L - r_i)$ times. The modified Jaccard index is simply the Jaccard index applied to the unordered sets generated by this procedure:

$$\tilde{J}(r_A, r_B) = J(\tilde{A}, \tilde{B}) \tag{3}$$

In such a way, a shift between two adjacent elements in the ranking returns, for example, a value of $\tilde{J} = 2/(\tilde{L}+2)$. Whereas, shifts among elements with very different rank, can produce larger variations because of the different number of copies present. By its definition, it can be directly applied also in the case where the entries of the rankings are not the same.

In Figure 1 we plotted the modified Jaccard index and the Kendall-Tau index for the comparison of different strings of length $L = 10$, in the cases where the p-th letter is exchanged with the x-th letter (i.e. for $p = 0$, the first point of the line, the exchange with $x = 1$, corresponds to the comparison between the strings abcdefghij-bacdefghij). As we can see the behaviour is similar to the one of the Kendall-Tau index, and thus we can conclude in this case where both can be used.

In Figure 2 we represented the modified Jaccard index (in this case the Kendall-Tau cannot be applied) for the case where a new letter is exchanged with the letter at the position x (i.e. the first point corresponds to the situation abcdefghij-zbcdefghij).

B. Correcting the dates

In the database we reported that the 0.32% of the couples mentor-student were characterised by an error in the PhD defence dates reporting $\Delta t_{MS} = year(student) - year(mentor) < 0$. Moreover the 6% of the scientists are not associated at all to any date. To solve this problem we used an iterative process consisting of two phases:

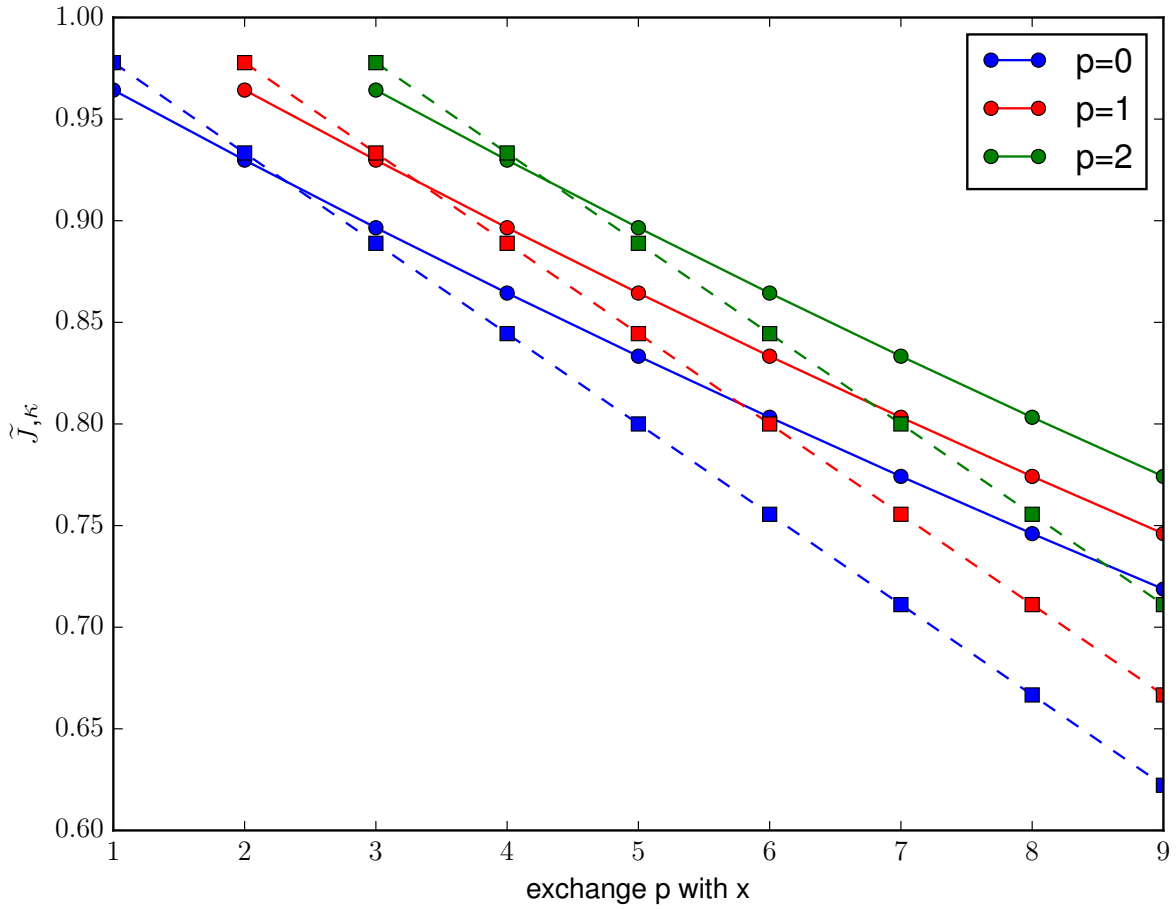


FIG. 1: Modified Jaccard index (continuous lines) and Kendall-Tau index (dotted lines) for the cases where the p-th letter is exchanged with the x-th letter.

- *Error detection and deletion*
  In the database two possible kinds of errors can be observed (see Figure 3).
  First we can observe errors in the date reporting. In this case the correction of the date of the mentor or of the student can generate a global coherent situation along the whole genealogical line. In this case we decide the date to change (the one of the mentor or the one of the student) according to the better p-scores in the mentor-student time difference distribution $P(\Delta t_{MS})$ comparing the mentor with its ancestors and the student with its students. If this kind of error occurs the less coherent date is suppressed. Secondly we can have a genealogical error, where the incoherence propagates on more than one generation and cannot be solved by changing a single date. In this case we suppress the filiation link.

- *Dates generation*
  Once the wrong dates are suppressed the missing dates are generated extracting a value from the mentor-student time difference distribution $P(\Delta t_{MS})$ when possible or using the time difference in the brotherhood when the first heuristic cannot be applied.

Of course also this new generation can produce errors. Therefore the two steps of this procedure are iterated until all the dates are fixed.
To test the statistical validity of our results we compared the new mentor-student time difference distribution $P(\Delta t_{MS})$ with the original one and we observed that the Kolmogorov Smirnov distance between the two distributions is very small ($d_{KS} = 0.007$).

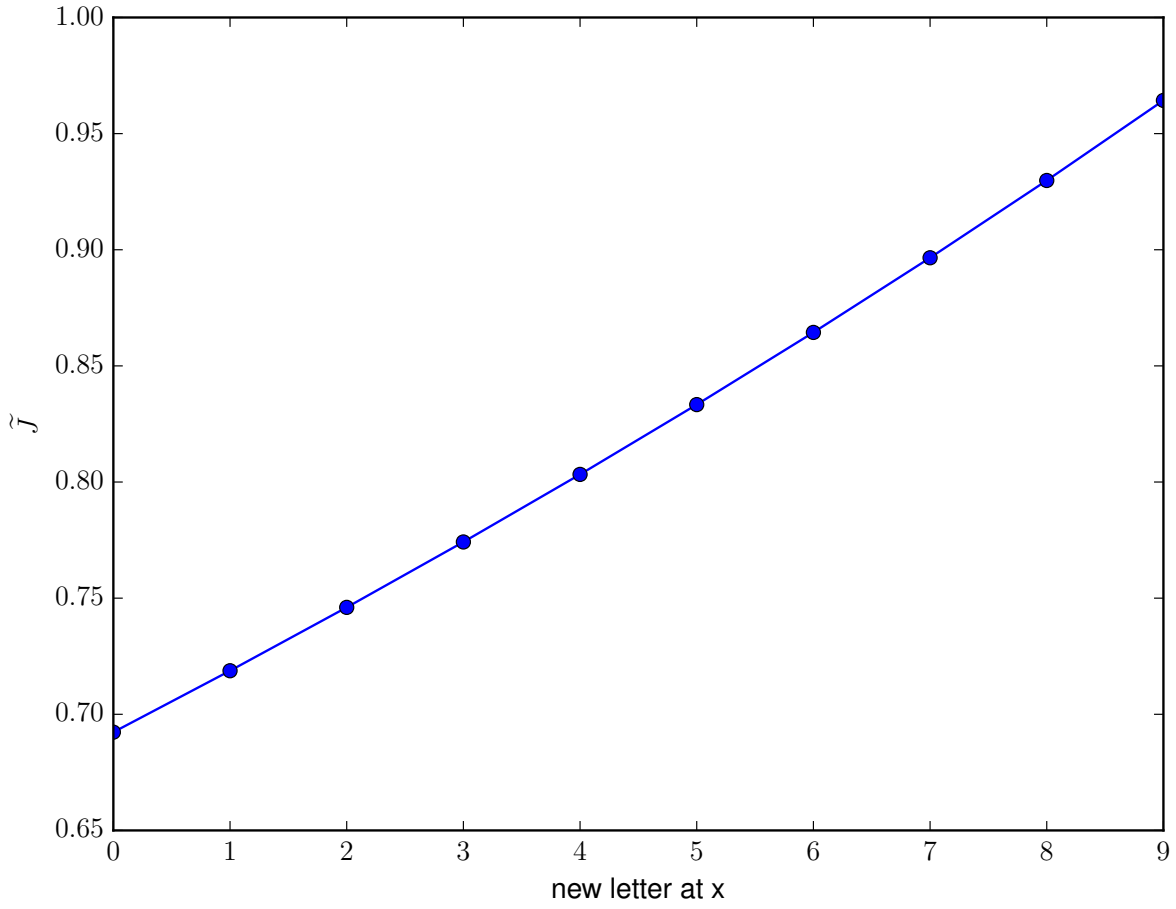


FIG. 2: Modified Jaccard index for the cases where the x-th letter is exchanged with a new letter.

C. Learning the disciplines

The 88% of the scientists presents the information concerning the title of the thesis. On the contrary only the 43% has an associated discipline code (one of the 63 classes by the AMS classification). We use the scientists with both the thesis title and the AMS code as a training set for a learning procedure aimed to learn an association between the thesis keywords and the codes.

First we translated all the not-English thesis titles to English using the Google Translate API. Second, we removed from the thesis all the stop words (the most common words in a language) using the Python NLTK library.

After this procedure the thesis title I is considered as a set of words, $T_I = \{w_1, w_2, ...\}$, and this holds for all thesis. The training set is the set of theses for which the domain is known $TS = \{T_I\}|_{D(T_I)=\mu}$.

Using the training set we reconstruct the word-domain matrix $M_{i\mu}$ where the index i identifies the words contained in the thesis titles and the index μ the thesis domain. At the beginning we set $M_{i\mu} = 0$ for all the values of the matrix. The algorithm to construct the matrix is schematically hereby reported:

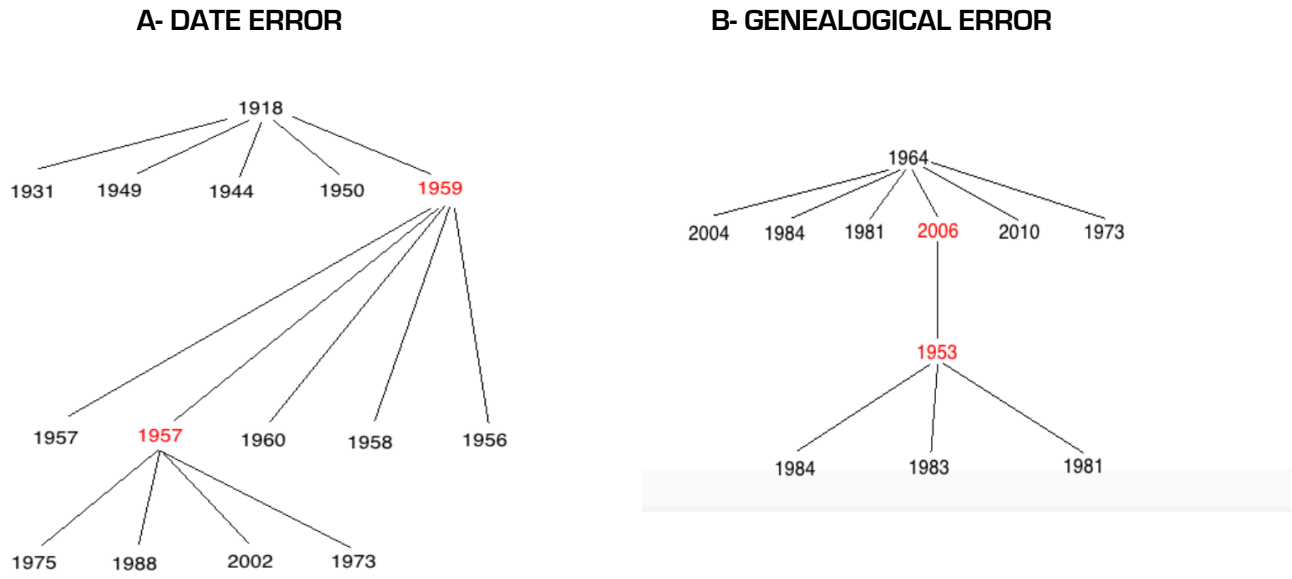


FIG. 3: Plot A: Example of a date error. In this case the date of the mentor has less coherence of the date of the student according to the context. Plot B: Example of a genealogical error. The dates of the student and of his mentor are respectively coherent respect to their brotherhoods and descendencies. Changing one of the two dates would not solve the error.

Data: Thesis with known domain T
Result: Words-Domain matrix $M_{i\mu}$
for *thesis in TS* **do**
  for *word in thesis* **do**
    $M_{word,Dthesis} = M_{word,Dthesis} + 1$
  end
end

Algorithm 1: Build the words-domain matrix.

Using the word-domain matrix we associate a score for each thesis I in each domain μ:

$$S_{I\mu} = \frac{1}{NT_\mu} \sum_{w \in T_I} \frac{M_{w\mu}}{\sum_\nu M_{w\nu}}\,, \tag{4}$$

where NT_μ is the number of thesis I in the domain μ.

We thus assign to a thesis the domain that maximises the score

$$AD(T_I) = \max_{\mu \in Domains} S_{I\mu}\,. \tag{5}$$

Such AD function can be used to predict the missing thesis domains. Applying the AD function to the thesis of the training set we obtain the 52% of positive results.

This procedure has been improved considering common sequences of words inside the thesis instead of single words (for example: "ALGEBRAIC SPACES" instead of {"$ALGEBRAIC$","$SPACES$"}). With this prelimiray improvement we obtaind the 75% of right classifications.

D. Finding the families

In order to cut the genealogical tree into families we use an algorithm, based on the topology of the network, able to assign to each node with more than one "parent" its more probable family membership. To deal with the multiplicity of the parents, the basic idea is to generate random binary trees (namely random cutting one of the parents links

for each scientist) and to look, for each configuration if the removed parent and the son still fall in the same connect graph. Applying this procedure on all the possible random trees, we can extract for each node with multiple parental links, the probability for each link that its removal will not influence the positioning of the removed parent in the same family of the son.
Since this procedure is computationally heavy, we performed a hierarchical grouping of the nodes in super-nodes structures: the super-nodes are the connected subgraphs of the network where the in-links of the nodes with multiple parents have been all removed (Figure 4). Using this reduced structure, the calculation of the most probable filiation links is now strongly simplified.

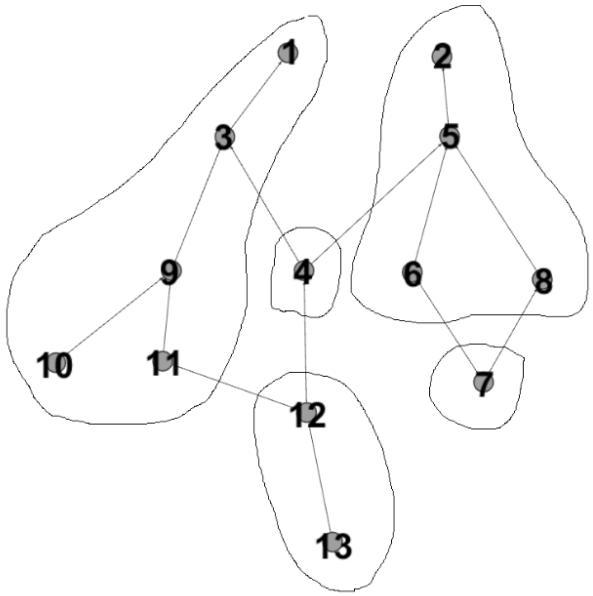


FIG. 4: Identification of the supernodes

II. THE HISTORICAL PATTERNS

In this section we show some prototypical "participation" patterns for different countries and disciplines, namely the profiles of the relative presence in each time window, for the 10 most significant countries 5 and disciplines 6.

From Figure 5 we can clearly notice the loss of centrality of countries like Italy and the Netherlands, and partially France. The rise of US and Russia coincides with the decline of Germany after the Second World War. Finally we can evince the fast growing trends for emerging countries like China, India and Brazil.
For the disciplines, in Figure 6 it is interesting to note the growth and fall of quantum theory and the parallel growth of group theory that is strongly inherent the physics of the 60's. Other disciplines like number theory and logic are quite uniformly distributed in the time.

III. THE MESOSCALE NETWORK STRUCTURES

In Figure 7 we display the structure of the time-aggregates static mesoscale networks. As it can appear from the graph, the strengths of the country network is strongly heterogeneous, varying from 6500 (USA) to 1 for several peripheral counties appearing just once in the database. The heterogeneity is less marked for the disciplines. As we can observe, both for countries and disciplines the degree centrality top list is not equivalent to the betweenness. For the countries the comparison between the top-10 countries in terms of degree and in term of betweenness put in evidence the very strategical role of France and est-European countries (Russia, Ukraine, Poland) in connecting information flows between the network. The difference between the two rankings is still more astonishing for the disciplines where just two disciplines (combinatorics and statistics) appear in the two top-10 lists. This scenario

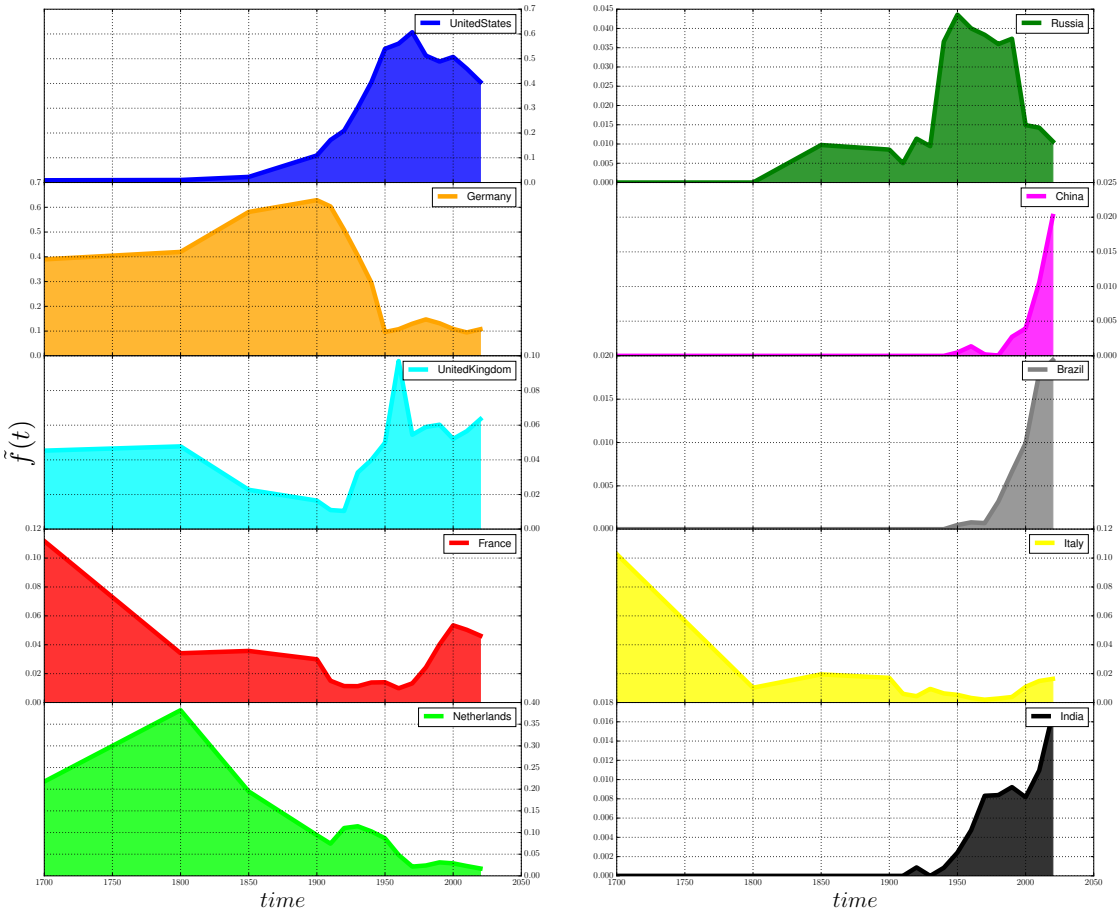


FIG. 5: Relative abundance profiles for 10 most significative countries in the database.

suggests the existence of very well structured epistemic communities around the top connected disciplines, where the connections between the communities is performed by peripherals nodes of each community. In Figure 8 we represent the time evolution of the fraction of the total links that loops and that are pointing in and out a country. As we can observe the emerging countries are characterised by a significantly high in-degree and a very small out-degree. The case of Italy is quite interesting showing a strong exodus during the second world war, followed by an inversion around the 50s due to the scientific politics of favouring the coming back of the emigrated scientists. The same maximum of the out-degree due to the second world war can be observed in Germany, while the opposite sign can be noticed in USA and UK. Notice also the following inversion of in and out degree prevalence for US after the 50s. This inversion is even more pronounced for Russia.

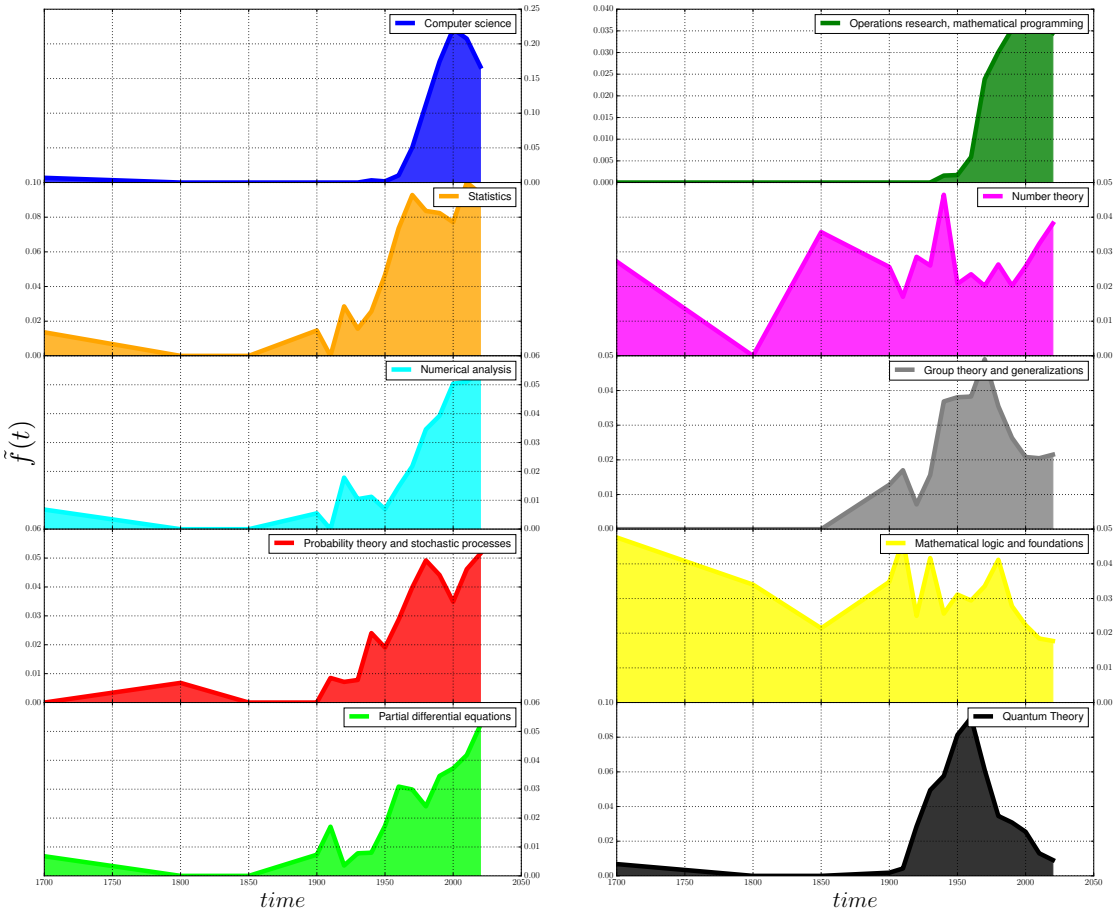
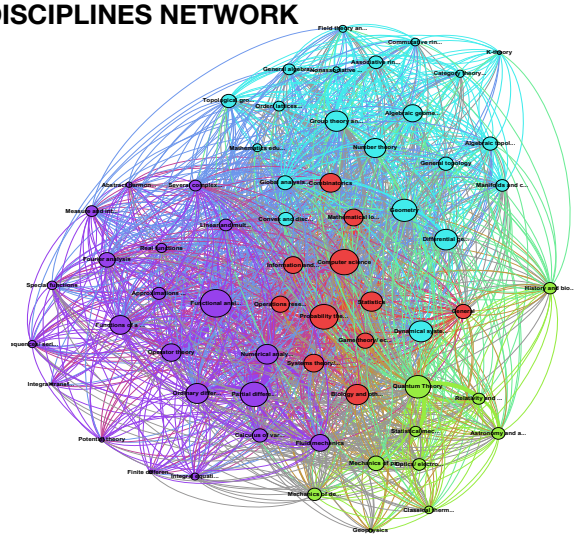


FIG. 6: Relative abundance profiles for 10 most significative disciplines in the database.

Degree centrality	Betweeness
UnitedState	France
Germany	Germany
UnitedKingdom	UnitedStates
Canada	Russia
France	Canada
Russia	UnitedKingdom
Switzerland	Australia
Netherlands	Ukraine
Australia	Netherlands
Italy	Poland



Degree centrality	Betweeness centrality
Computer science	History and biography
Probability theory and	Functions of a complex
Partial differential	Number theory
Numerical analysis	Biology and other natural
Functional analysis	Systems theory
Quantum Theory	Differential geometry
Statistics	Combinatorics
Operations research-	Fourier analysis
Group theory and	Operator theory
Combinatorics	Statistics

FIG. 7: Static network representation for countries and disciplines. The node size is proportional to the total degree. Some centrality measures are shown in the tables.

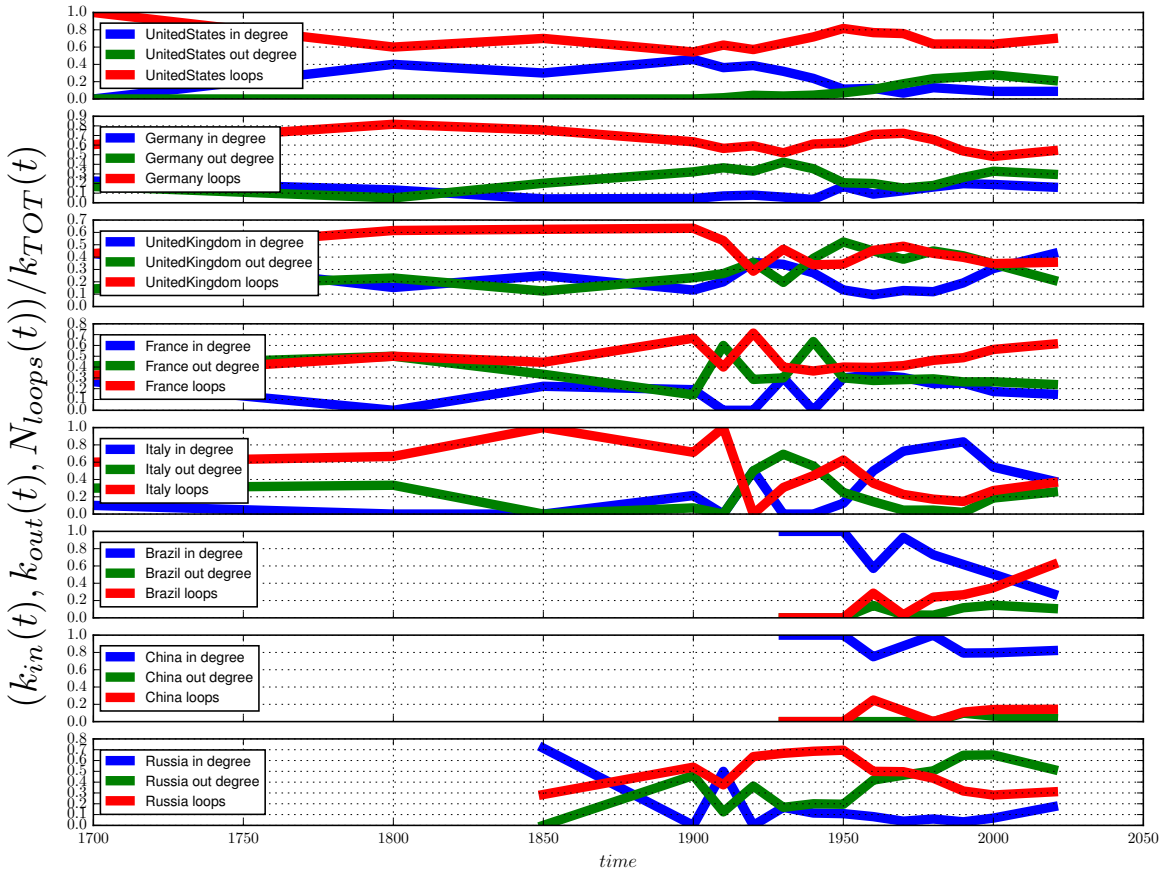


FIG. 8: Fraction of In-links,out-links and self-links for some significative countries in the dataset.