# Evolution Over Time of the Structure of Social Graphs

**Clustering**

Sofiia Shelest

*Supervised by*

Frédéric Giroire

Nicolas Nisse

Małgorzata Sulkowska

Thibaud Trolliet

# Contents

# Abstract

LabEx and IdEx are research funding programs supported by the French government as a part of 'Investments for the Future' program. Both programs encourage the best French research centres to increase their scientific influence and to promote scientific collaborations by bringing together different research teams. In September 2018, was the beginning of the SNIF (Scientific Networks and IDEX Funding) project [17] that aimed to investigate the impact of above-mentioned programs on interdisciplinary collaborations and productivity of researchers. This Master thesis was conducted in the scope of the above-mentioned project. The purpose of this Master thesis is to develop an approach to define community productivity, then examine how the productivity of communities with PhD students that were granted LabEx scholarship for their PhD program changed over time, and finally, compare productivities of communities with LabEx or IdEx PhD students to those that had none. This investigation was performed on two networks: on a small network of Nice-Sophia Antipolis scientific network (about 10 000 members) and on a larger network of France Computer Science collaboration network (about 170 000 members). The final results on both networks have shown that pairs of communities that supervised PhD works are more productive than other communities.

# 1 Introduction

## 1.1 Motivations

The purpose of multiple funding programs, such as IdEx and LabEx, is to facilitate scientific collaborations between research teams. For example, the idea of the LabEx program is to fund PhD students that are supervised by different research teams to increase the number of collaborations between those teams in the future.

This Master thesis is a part of the Scientific Networks and IdEx Funding (SNIF) project conducted by INRIA Sophia Antipolis, I3S, GREDEG and SKEMA Business School. The goal of SNIF project is to measure the success of funding of the aforementioned programs by studying their influence on the evolution of collaborations among research teams.

In the scope of this project, it had to be investigated whether PhD works with LabEx funding had an impact on collaborations between communities that supervised those PhD works in the scope of a small network. The goal is to developed approach, and investigate an impact on France Computer Science network of other PhD works funded by IdEx program.

## 1.2 Objectives

*Collaboration networks* play a significant role in modern science. They attract researchers from different fields of study, such as psychology, sociology, economics, computer science, etc.

In our case, SNIF project [17] brings together researchers in computer science, economics, management and sociology from four partners of Université Côte d'Azur (INRIA Sophia Antipolis, I3S, GREDEG, and SKEMA Business School) to investigate the impact of funding programs on researchers' productivity.

We assume that promoting new international and/or interdisciplinary collaborations can bring new fresh ideas into collaborating teams, which can lead to an increase in researchers' productivity and strong partnerships.

The main objective of this thesis is to investigate an impact of PhD theses on collaboration network.

To achieve this goal we built *collaboration network* (Section 2.2), that will represent the network of scientific authors.

Then, using existing algorithm for community detection, specifically Leiden algorithm [20], identify communities/partitions in abovementioned network.

It is believed that human networks tend to have community structure. This property of *community clustering* of the network can be verified by computing *modularity value* [14] of this partition.

The high modularity value means that the nodes inside partition are much tightly connected within community than with other nodes from other communities, which can be observed in further sections.

When communities are identified in the network, we try different

methods for estimating the productivity of communities. We assume that community productivity can be reflected in probabilities to co-author publications with other communities.

And finally, we have to compare how the probability to make publication between communities evolved before they co-supervised a common PhD work and after. Also, we are curious to check the productivities of other communities without PhD theses within the same period of time, to be able to say that there is a difference in productivities of communities with and without PhD supervision.

## 1.3 Tools

For this research, we used the Python 3.9 programming language due to the wide selection of packages for research and the brevity of the language itself. For graph representation, we used NetworkX [13] library. It offers fast tools for working with graphs and CDlib [4] library provides fast algorithm for community detection which is called Leiden algorithm [20] that is compatible with NetworkX graphs.

Charts, plots, etc. were visualized with MatPlotLib [12]. and graph visualization was made with Cytoscape tool [6].

To analyse big amount of data different libraries were used such as pandas [15], seaborn [21] and NumPy [8].

My personal machine was used for experiments and optimizations, which has the following characteristics.

- Processor: Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz 1.99GHz
- Installed RAM: 8.00 GB (7.89 usable)
- OS: Windows 10

## 1.4 Paper Organization

This thesis consists of 3 main sections:

Section 2 provides general information on the collaboration network that we were investigating, as well as the data that underlies the structure of the network.

Next Section 3 describes the approach that was used for investigation of impact of LabEx funded PhD works on Nice Computer Science network and obtained results. It provides general information on the network as well as the description of community structure of this network. Also, this section presents methods for analysing evolution of community productivity through years.

The final Section 4 introduces another more effective method for investigation of evolution of community productivity in large networks. This new approach was used to investigate the impact of PhD works in Computer Science on France Computer Science network and obtained results are presented in this section.

# 2 General Information

## 2.1 Data Origin

This Section provides the general information on the data, that was used for this investigation. There were two main sources of data.

1. Scopus database 2.1.1, which was used for building collaboration networks;

2. Theses.fr database 2.1.2, which was used as a source of the information on PhD theses, that were defended in France. This data is needed to study the impact of these works on productivities of communities in investigated networks.

### 2.1.1 Scopus database

The source of the data was the Scopus [16] database, which is the largest abstract and citation database owned by Elsevier. Scopus database covers journals and books from various publishers across multiple fields of science, technology, medicine, social sciences, and arts and humanities.

The extracted data contains publications that were published between 1990 and 2018 years across multiple disciplines. Since in current research, we need to investigate the influence of French funding

programs, each selected publication must have at least one author affiliated to France.

In total, there are 783 JSON files split among 27 folders. Each folder corresponds to one field of science such as agriculture, chemistry, computer science, etc and contains data on publications related to the corresponding field. Directories were named with the first four letters of the disciplines, such as COMP, MATH, PHYS, etc. JSON files also have the name of the discipline in the title and the publishing year.

### 2.1.2 Theses.fr database

The objective of this thesis is to develop and test investigation methods of evolution of productivity of pairs of communities. We expect that there will be two kinds of such community pairs: first – that co-supervised PhD thesis, and second – without co-supervised PhD work. However, we had not enough data on LabEx/IdEx PhD works, which were used for investigation of a small network (about 10 000 members). Thus, the other source of information was used, called Theses.fr [19].

Theses.fr is an online database that is constantly evolving. It consists of thousands of PhD theses on different disciplines that were defended in France since 1985.

For our purposes, 16 142 PhD works on Computer Science were retrieved from this database. These PhD works have 20 889 distinct names in total, including students and supervisors. PhD works that have only 1 supervisor and PhD works, where at least one of the supervisors is missing a name in the retrieved data, were excluded from the data that was used for analysis. In total, there were 10 200 PhD theses with two supervisors.

Thus, 5942 (36.81 % from all) PhD works were used for exper-

iments. They have 11 067 distinct names, which includes the name of the PhD student and names of their supervisors, which is 52.98 % from all authors.

For further work, each author had to be found in existing data (that we have) and to be assigned with their Scopus Id. However, there was a problem, that sometimes there was no one to one match of author name and Scopus Id. Because of this, the number of valid PhD data was reduced even more till 5595.

The key date for data from Theses.fr is year *2010*, since there are only 35 PhD works before this year, and we suppose that such small number of PhD works compared to the total number of utilized data on PhD theses will not have significant distortion in computations, and we still can get accurate results.

### 2.1.3 LabEx data

The data on PhD works that were granted with LabEx [1] scholarship was obtained from Polytech Nice Sophia [2]. Each of these PhD works were supervised and defended in Nice.

The key data for LabEx PhD data is year *2013*, since the first student that was granted with this scholarship appeared in 2012 but had their first publication a year after.

## 2.2 Collaboration Graph

We consider a mathematical model of a collaboration network by representing the network as a *collaboration graph*, where nodes represent authors and edges represent collaborations between them (i.e., there is an edge between two authors if they collaborated on a single publi-

cation). Such a model allows utilising mathematical tools and various graph algorithms to study the structure of the network and calculate numerous metrics, such as clustering of a network, and see how it evolves over time.

To construct a collaboration graph, we use the data about publications that was extracted from the Scopus database 2.1.1.

A pseudocode of the method of building a collaboration graph is presented in Algorithm 1.

Algorithm 1: Build a Graph

**Input:** a list of publications $P$
**Output:** a graph $G$, an edge weight function $w$

```
1: G ← an empty graph
2: w(e) ← 0                                   ▷ for any e
3: for all publication p in P do
4:     for all pair of authors (a, b) of p do
5:         G ← G with an edge between a and b
6:         w(ab) ← w(ab) + 1,
7:     end for
8: end for
```

# 3 Impact of LABEX in Nice-Sophia

This Section describes investigated network, applied approaches for detection and comparison of changes in productivities of communities through years. A small network, that corresponds to collaboration network of Nice and Sophia-Antipolis, was used for this part of investigation.

## 3.1 Nice-Sophia Computer Science Network

*Nice-Sophia Computer Science network* is the network of researchers that was built based on Scopus papers and articles published between 1990 and 2018, where at least one of the authors was affiliated to Nice or Sophia Antipolis. In total, we have 10 821 such publications with 14 967 distinct authors.

The resulting network has the same number of nodes as the number of authors, which is 14 967 nodes. Two author-nodes are connected with an edge if those authors had at least one common publication, and the total number of edges is 116 684.

However, this network is not connected. When we look deeper

in the structure of this network we will find that there are multiple components in the network (Figure: 3.1). This may occur due to the fact that, there were some publications in some isolated groups that never collaborated with "outer" world.

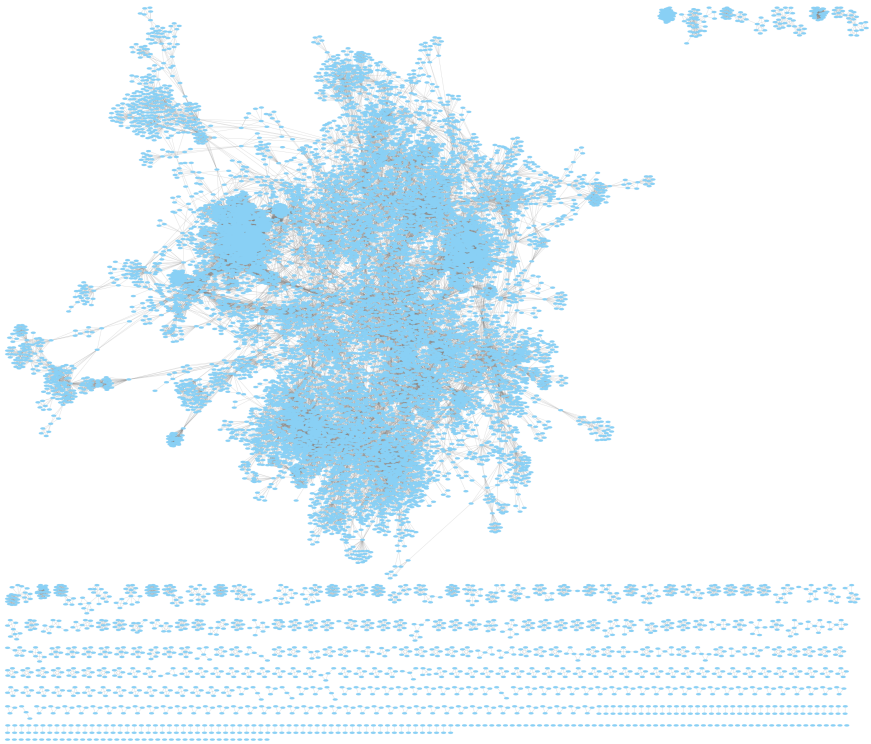


Figure 3.1: Shape of Nice Sophia Computer Science Network.

Such detached communities are not useful for this investigation of future collaborations of communities with PhD students. Recall that we are interested in communities that collaborate with other communities, and we want to investigate how collaboration changed when two different communities supervised one PhD student that was granted with LabEx scholarship. Thereby, these small isolated groups will be omitted in further work.

Thus, it was decided to keep only the largest connected component of the network. In such manner, when all nodes are assigned to their community, each of these communities will have at least one connection with some other community. The largest connected com-

ponent of Nice Sophia Computer Science network has 12 901 nodes and 109 636 edges.

Leiden algorithm [20] split all 12 901 authors into 45 disjoint communities. *Modularity* [5] of this partition was 0.85, which indicates that the network exhibits strong community structure. The notion of modularity was introduced by Mark Newman and Michelle Girvan [14]. Its value ranges between $-\frac{1}{2}$ and 1. If the modularity value is high, this means that the current partition of the nodes has many edges inside the communities that makes communities very dense, while the number of edges placed between communities is significantly lower.

Sizes of obtained communities vary from 15 till 915 members. In figure 3.2 we can see the distribution of community sizes in partition obtained by Leiden algorithm.

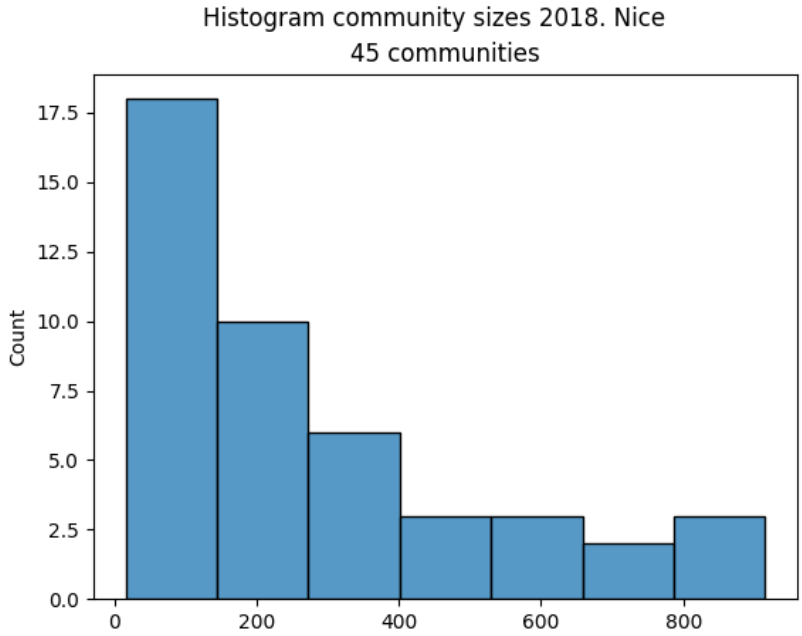



Figure 3.2: Distribution of community sizes in Nice Sophia Computer Science Network.

Thinking ahead, when we move to the larger network there will be much more communities and to consider all of them is not necessary to

obtain accurate results. It was decided to work only with first largest communities that make in total about 80% of the network and test this approach on a smaller network.

For example, we have some network with partition of nodes as in Figure 3.3. This network has 28 nodes, and we remove the smallest community like in Figure 3.4, we will have about 85% of the network.

We believe that considering only the largest communities we work also simply with the most active ones. Also, other approaches could lead to disconnected network. For example, if we took 80% of the network and then did partitioning we would break the original structure of communities, some of them could be disconnected.

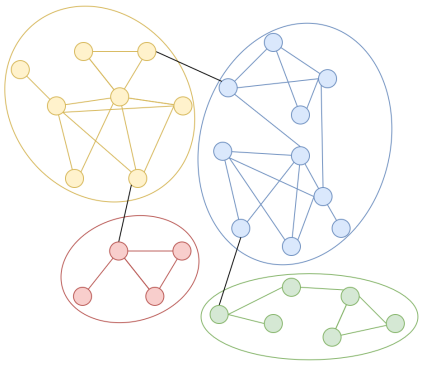


Figure 3.3: Original partition given by Leiden algorithm.

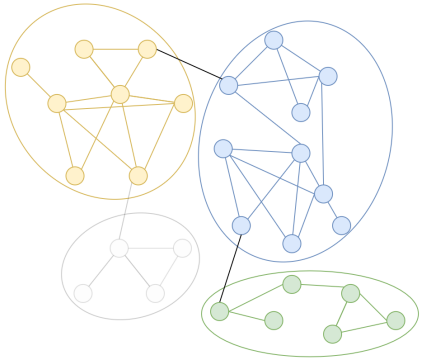


Figure 3.4: From the original partition, only largest communities are kept.

The number 80% is chosen because if we take less communities, for example that make 70% of the network, we lose some communities that had supervised PhD student with some other community, and if we keep more communities we would have to take into account numerous small communities, which would not have much impact on obtained results, but it would take more computational time to process them.

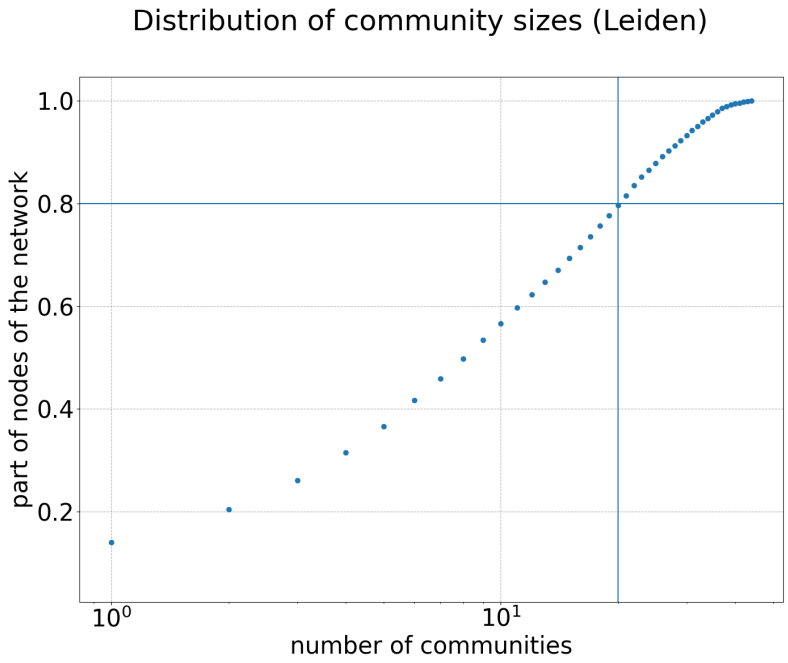



Figure 3.5: Distribution of sizes of the largest communities in Nice Sophia Computer Science Network. It shows what number of the largest communities should be taken to get 80% of the network.

To get about 80% of the network according to Figure 3.5 we can take only 20 communities, which is less than a half from all communities of the largest connected component of Nice Sophia Computer Science Network. Due to removal of small communities now community sizes range from 261 till 915 members in this network.

These 20 largest communities consist of 10 018 members in total. Modularity of this partition is 0.71 which is lower than original. Decrease of modularity could be caused by changes in structure of the network as some nodes with edges were removed, and remaining nodes could be organized a bit differently. However, we are interested in primary partition structure that was detected in the original network. In this way, the impact of excluded nodes on partitions will be considered.

## 3.2 Community Stability

To investigate how the productivity of the community changed in several years, we need to identify this community in each collaboration networks that correspond to network states of investigated years. Thus, the first step was to build several networks for each of the considered years. As a reminder, in the Nice-Sophia Computer Science investigation, the core years are 2013 when the first PhD student was granted with LabEx scholarship (see Section 2.1.3), and 2018 – as the most recent year to compare to find changes in collaborations (see Section 2.1.1).

When Nice-Sophia Computer Science networks that correspond to network states for 2013 and 2018 were built (later on Nice-Sophia-2013 and Nice-Sophia-2018 networks for short) it was found that the number of communities can vary from year to year. Even if after several runs of the Leiden algorithm we could achieve the same number of communities through all years, we could not guarantee that these communities were the same. Thus, we needed to check whether the communities are stable over years 2013-2018, so the same communities from the past could be compared to their future versions.

As a first step, it was decided to try to find the best matching communities by comparing each community from 2013 to each community from 2018.

Communities can be matched by members of this community. This means if we take each community from 2013 we check whether we can find a corresponding one in 2018 with the significant number of the same members. This way we can get our match. Algorithm 2 describes this idea.

Algorithm 2: Compare communities by members

**Input:** list of communities in 2013;
list of communities in 2018.
**Output:** M – matrix of matching coefficients

1: $communities_{2013} \leftarrow$ list of communities in 2013;
2: $communities_{2018} \leftarrow$ list of communities in 2018
3: $M[m \times n]$, where $m$ – number of communities in 2013, and $n$ – number of communities in 2018
4: **for all** community $c_i$ in $communities_{2013}$ **do**
5: **for all** community $c_j$ of $communities_{2018}$ **do**
6: $M_{i,j} \leftarrow \frac{|c_i \cap c_j|}{|c_i|}$
7: **end for**
8: **end for**

As an example in Figure 3.6 in 2013 there was one community with members $[1, 2, 3, 4, 5]$. Then, in 2018 there are 2 communities. If we compare each community from 2018 to community in 2013 we will get the following matching coefficients:

- $\frac{|[5,..,..]\cap[1,2,3,4,5]|}{|[1,2,3,4,5]|} = \frac{1}{5} = 0.2$
- $\frac{|[1,2,3,4]\cap[1,2,3,4,5]|}{|[1,2,3,4,5]|} = \frac{4}{5} = 0.8$

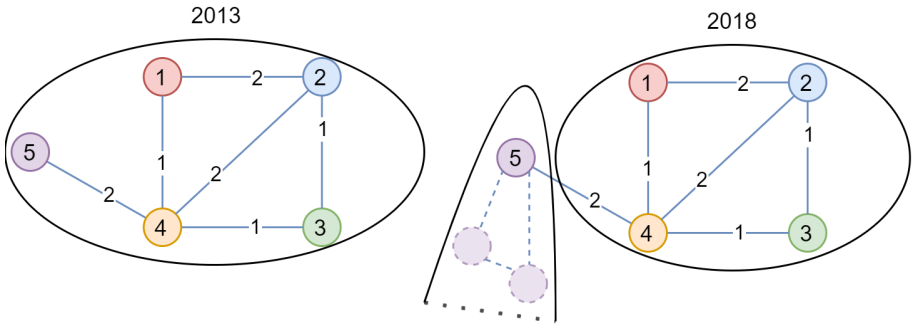



Figure 3.6: Matching communities.

However, when matching matrix was computed the following results were obtained (see Figure 3.7). The columns of matching matrix

in Figure 3.7 correspond to communities detected in the network Nice-Sophia-2013 and each column is labelled with its community number. Same for rows, but they correspond to communities detected in the network Nice-Sophia-2018. The values of this matrix represent matching coefficient which can take value [0, 1].

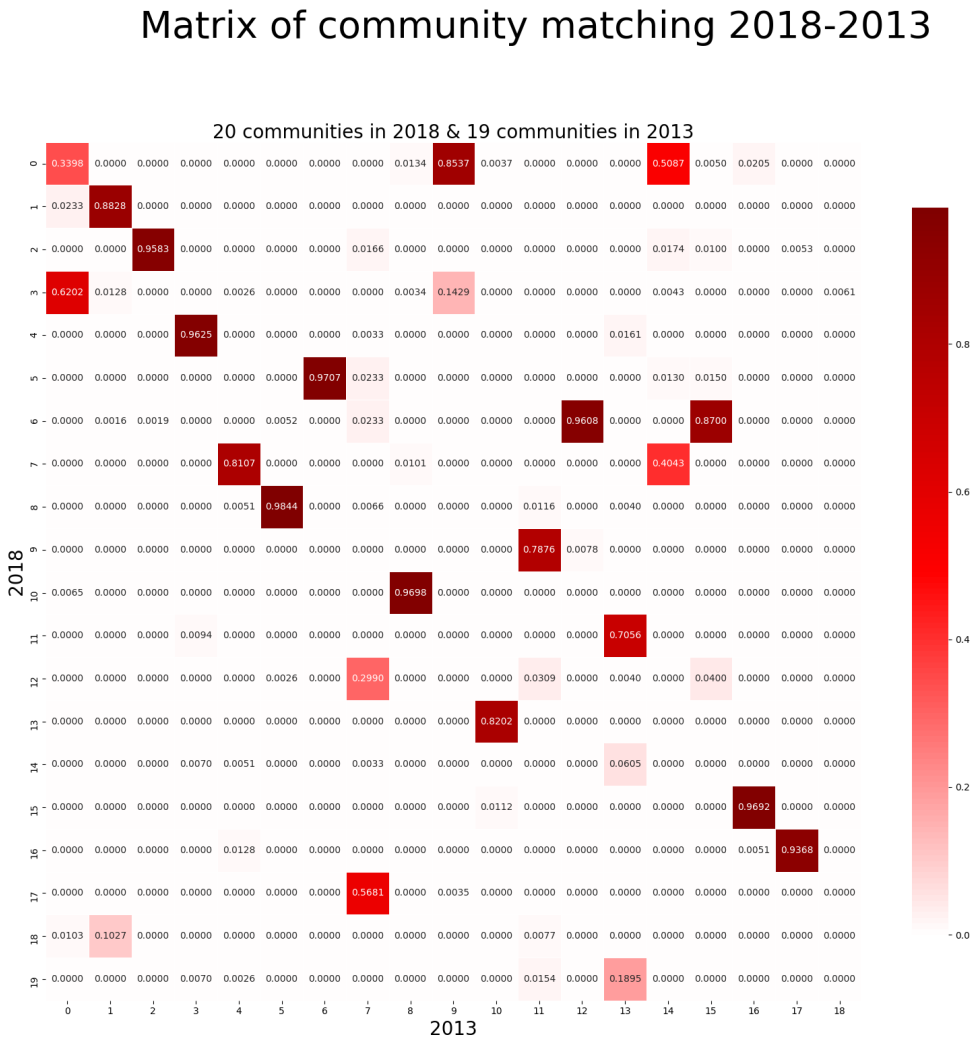



Figure 3.7: Matrix of community matching 2018/2013.

Thus, for each community from 2013 we try to match it with the community from 2018 for which the highest value of matching coefficient was obtained.

The main problem with this approach is that there is no one to

one matching between all communities. According to the matrix of community matchings, there are three scenarios of evolution of communities. Either there is solid matching between 2 communities with more than 0.9 matching coefficient which is the perfect scenario, or, the most difficult cases to process, when communities from 2013 merged into one community, or even split among several communities in 2018.

Nonetheless, if we look deeper into underlying data of this matrix, we can see the following: if we count the number of nodes that "stayed" together through all years and did not move from its "colleagues" to another community we can see that from all 6395 members from 2013, 5658 stayed together in 2018, which is about 89% of all members of the network.

Thence, due to this special behaviour of community members, it was decided that we can use one community partition of 2018 through all years. The most important feature of this approach is that even if communities are merged together or split among several communities in the future, we can be sure that most of the associates of each member are the same.

## 3.3 Modified Stochastic Block Model

Inspired by Stochastic Block Model approach for generating random graphs with communities firstly defined by Holland et al. [10] and other applications of this model [3, 11] it was decided to adapt this method for detecting changes in community structure over time.

To study changes in community collaborations before and after communities got PhD students with Labex/IdEx funding we use non-symmetric matrix $W \in \mathbb{R}^{k \times k}$, where $k$ – corresponds to the number of communities and $w_{i,j} \in \mathbb{R}$, where $i, j \in \{1, \ldots, k\}$ is an estimator of

probability to make a publication between community $i$ and $j$. Probability to make a publication for each ordered pair of communities can be computed as follows:

$$w_{i,j} = \frac{\text{number of common publications between } c_i \text{ and } c_j}{\text{total number of publications of community } c_i}$$

note that

$$\sum_{j=1}^{k} w_{i,j} = 1, \text{ where } i \in \{1, \ldots, k\}$$

This way we define matrix $W$, that determines how often community $i$ collaborates with community $j$, which also can be referred as *productivity matrix*.

**Remark.** *Since we are interested only in collaborations between communities, all publications inside the community will not be counted. Thus, all values in the diagonal of each matrix will be zero.*

For example, there are 4 communities $[C_1, C_2, C_3, C_4]$. Then the matrix of number of publications defined as follows (see Algorithm 3):

Algorithm 3: Matrix of number of publications

**Input:** list of publications $P$
**Output:** $N_{k\times k}$ – matrix of number of publications

1: $N[k \times k]$, where $k$ – number of communities
2: $C \leftarrow$ empty set;
3: **for all** publication $p$ in $P$ **do**
4: $C \leftarrow$ community of each author of $p$;
5: **for all** unordered pairs of communities $\{C_i, C_j\}$ of $C$, where $C_i \neq C_j$ **do**
6: $N_{C_i,C_j} \leftarrow N_{C_i,C_j} + 1$ ;
7: $N_{C_j,C_i} \leftarrow N_{C_j,C_i} + 1$ ;
8: **end for**
9: **end for**

In the example on the left in Figure 3.8 we can see the matrix of number of publications for communities $[C_1, C_2, C_3, C_4]$. Then matrix $W$ is computed by dividing each value in the row by the sum of all elements in this raw. Finally, the resulting matrix has the form as on the example on the right in Figure 3.8.

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| C1 | 0 | 4 | 3 | 1 |
| C2 | 4 | 0 | 2 | 0 |
| C3 | 3 | 2 | 0 | 1 |
| C4 | 1 | 0 | 1 | 0 |

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| C1 | 0 | 0.5 | 0.375 | 0.125 |
| C2 | 0.67 | 0 | 0.33 | 0 |
| C3 | 0.5 | 0.33 | 0 | 0.17 |
| C4 | 0.5 | 0 | 0.5 | 0 |

Figure 3.8: Example of matrix of probabilities to make publication between each pair of communities.

Since we suppose that PhD work that was supervised by two researchers from different communities can boost future collaborations of these communities, we hope that it will be reflected in probabilities to make publication between them.

To find differences in productivities of communities, one more metric was defined. Let $M \in \mathbb{R}^{k \times k}$ be the matrix where $m_{i,j}$ corresponds to gain in probability to make publication between communities $i$ and $j$ and $m_{i,j}$ is computed as follows:

$$m_{i,j} = \frac{w_{i,j}(2018) - w_{i,j}(2013)}{w_{i,j}(2018)} \text{ , where } i, j \in \{1, \ldots, k\}$$

Thus, $m_{i,j}$ can fall into $(-\infty, 1]$ range, where

- $m_{i,j} < 0$ – corresponds to decrease in intensity of collaboration of communities $i$ and $j$;
- $m_{i,j} = 0$ – indicates that the intensity of cooperation stayed approximately at the same level since 2018;
- $m_{i,j} \in (0, 1]$ – implies that intensity of cooperation increased and $m_{i,j}$ corresponds to the gain coefficient.

Thusly, the following plan was defined:

1. Build two collaboration networks. One that corresponds to state of the network till 2013 (Nice-Sophia-2013) and the second one – till 2018 (Nice-Sophia-2018).
2. Detect communities on Nice-Sophia-2018 network.
3. Compute matrices of number of publications between communities using partitions from Nice-Sophia-2018 network. For matrix for Nice-Sophia-2013, use all publications on computer science till 2013, where at least one of the authors is affiliated to Nice or Sophia Antipolis. Use similar approach for publications matrix for Nice-Sophia-2018 network.
4. Compute $W_{2013}$ and $W_{2018}$ matrices using above-mentioned matrices.
5. Compute matrix $M$.
6. Find and compare changes in collaborations between communities that supervised a common PhD work, and see if there really was some significant growth, also compare changes in collaborations between communities without PhD work.

## 3.4 First Results on Nice-Sophia Network

When all steps and metrics were defined, experiment on Nice-Sophia network has been conducted.

As a reminder, for Nice-Sophia investigation there are two key-dates: 2013 – is the year, when first PhD student got LabEx scholarship, and 2018 – is the latest papers from Scopus database we have, thus this is the most recent network state we could get by building the network based on all available data. For more information see Section 2 and Section 3.1.

Recall that Nice-Sophia network till 2018 is:

- Graph on all data: 14 967 nodes and 116 684 edges. (10 821 publications involved.)
- Largest component: 12 901 nodes and 109 636 edges.
- Among 45 communities, only 20 of them make about 80% of the network.
- 20 largest communities contain 10 018 authors.

In total, we have 25 PhD works with LabEx scholarship and all of them were involved in the collaboration network till 2018. However, there are only 13 distinct pairs of communities that supervised 14 works and other 11 works were made inside a single community.

**Remark.** *We compute distinct pairs as all combinations of communities involved in PhD work: communities of supervisors and community of PhD student.*

The number of "LabEx" values is two times more because if you

see definition of matrix $W$ in Section 3.3 you can notice that $w_{i,j} \neq w_{j,i}$.

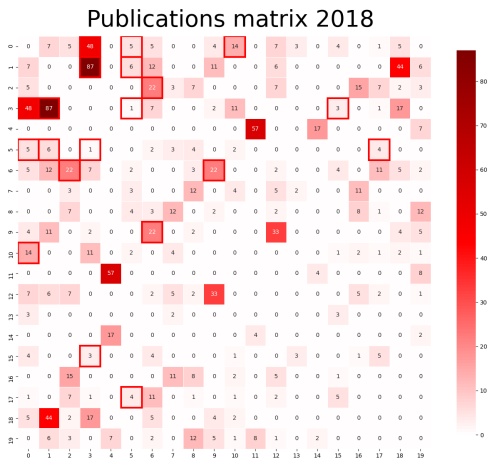



Figure 3.9: Matrix of number of publications till 2018.

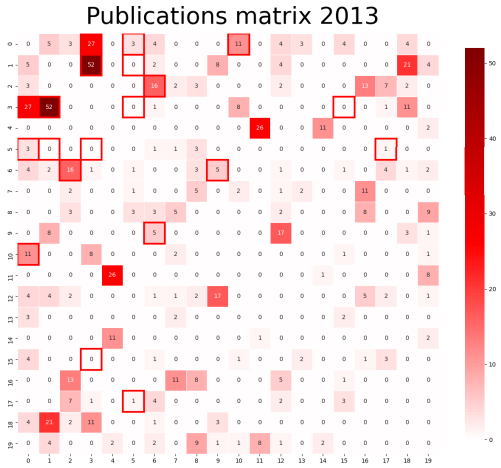



Figure 3.10: Matrix of number of publications till 2013.

Firstly, when communities were detected and largest of them were picked for investigation, matrices of number of publications were built. Figures 3.9 and 3.10 represent publications matrices till 2018 and 2013 correspondingly. Values on diagonal are zeros because we do not consider productivities inside a single community. Also, for convenience heat map visualization technique was used to highlight places with the large number of publications, as well as "LabEx" pairs were emphasized by red squares.

Secondly, matrices $W_{2018}$ and $W_{2013}$ were computed as well as matrix $M$ of differences between productivities of communities in 2013 and 5 years after when the first PhD work with LabEx grant has started. For more information on how these matrices were build see Section 3.3. Figure 3.11 represents the resulting matrix.

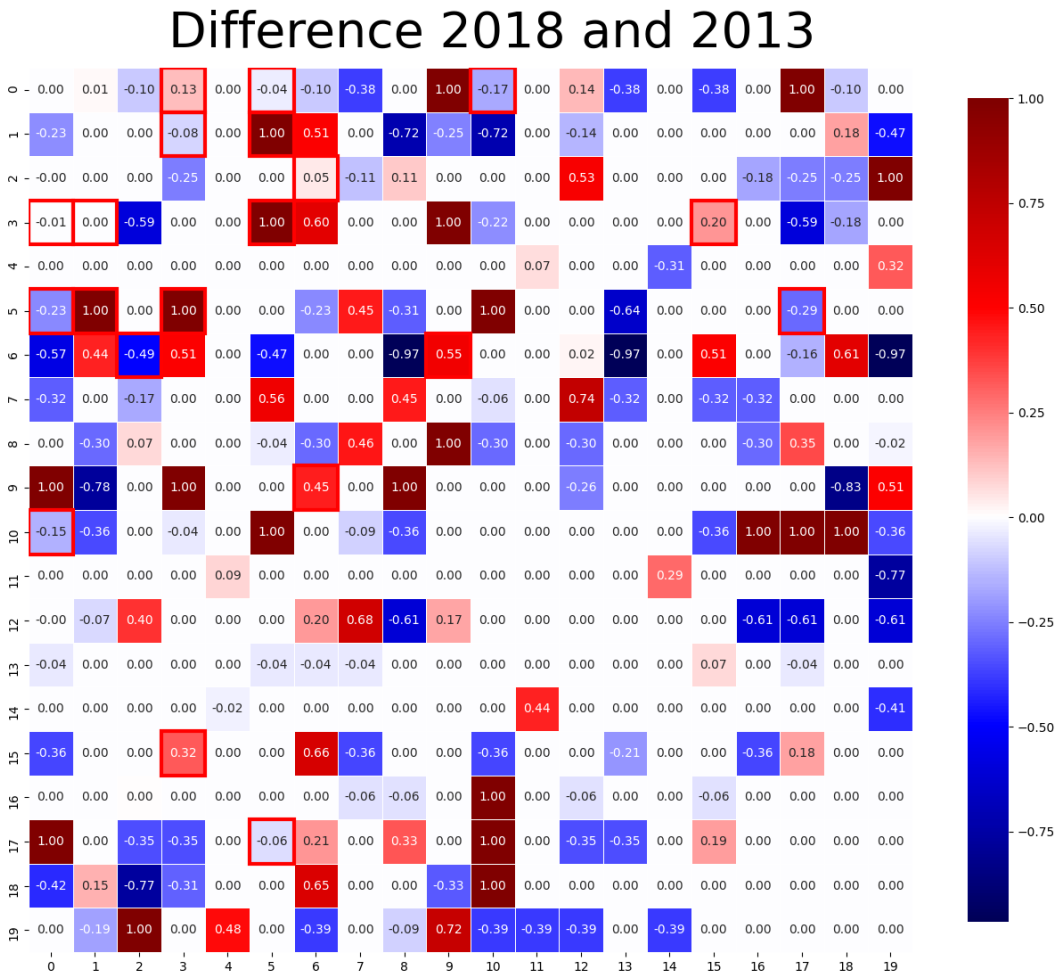



Figure 3.11: Difference matrix $M$. Values in blue colour indicate decrease of probability to make publication between two communities, and red colours – increase of it.

Then, difference values were extracted from difference matrix $M$ (Figure 3.11) and organized into two disjoint groups:

- *"Labex" group* – a group of values that correspond to communities that supervised PhD works with LabEx grant,
- and *"All" group* – a group that consists of all other productivity values, except values on diagonal of the matrix.

Finally, several metrics were computed for each group.

| | Median | Average |
|---|---|---|
| LabEx | 0.178 | 0.277 |
| All | 0.000 | -0.00021 |

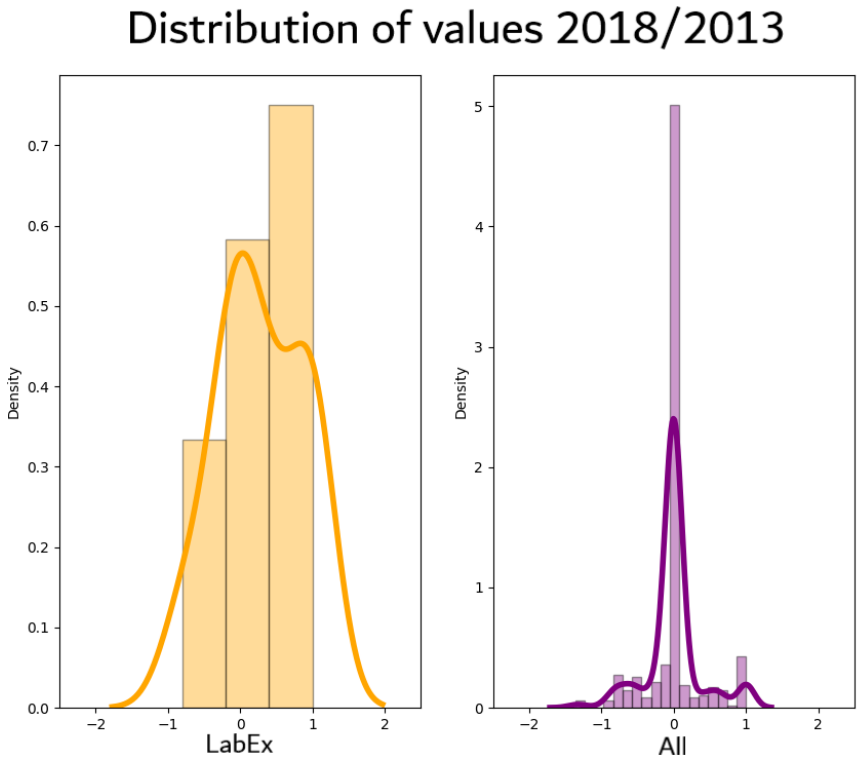



Figure 3.12: Distribution of values of matrix $M$.

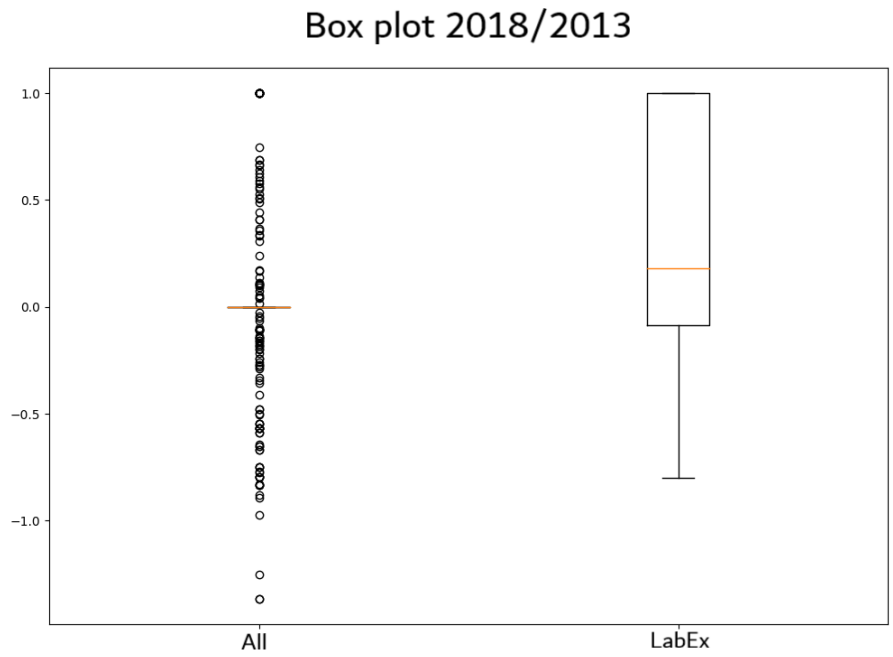



Figure 3.13: Corresponding box plot for distribution of values of matrix $M$.

Thusly, judging by the median and average values of each group and the shape of the data in Figures 3.12, and 3.13 we can say that "LabEx" communities have greater gain in productivity than other communities. Median gain in "LabEx" communities is about 17.8%, while in other communities there is almost no changes in their productivity.

However, these results are based on a very small number of "LabEx" samples. Among 25 PhD works there are 13 distinct pairs of communities with LabEx PhD students and 11 other LabEx PhD works were distributed among 9 different communities but supervised only by a single community: 11 PhD works assigned to one of 9 communities.

If we return to distribution of community sizes, they ranged from 261 till 915 members. We can expect that for the network with 10 018 nodes, 20 communities with such size range are too big and real changes in productivities are hidden inside too big communities.

## 3.5 Reducing Community Size

In order to test our suggestion that too large communities can hide some important changes in the network, it was decided to try the following method:

1. Take only largest communities that make about 80% of the largest connected component as it was done in first steps.
2. Then, for each of the largest communities, treat them as an independent network, and run the Leiden algorithm on each of these "networks".

In this way, the original partition is preserved and partitioning of each large community is not affected by their neighbouring communities from original partition (see Figures 3.14 and 3.15).

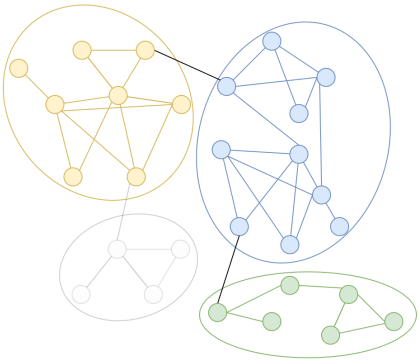


Figure 3.14: From the original partition, only largest communities are kept.

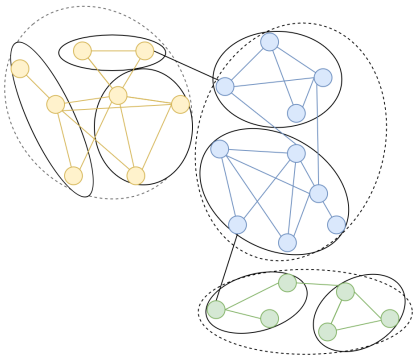


Figure 3.15: Run Leiden on each of the partitions.

After community size reduction, 20 communities were split into 227 smaller communities. Now, community sizes range from 2 to 243 members. Figure 3.16 shows the distribution of new community sizes.

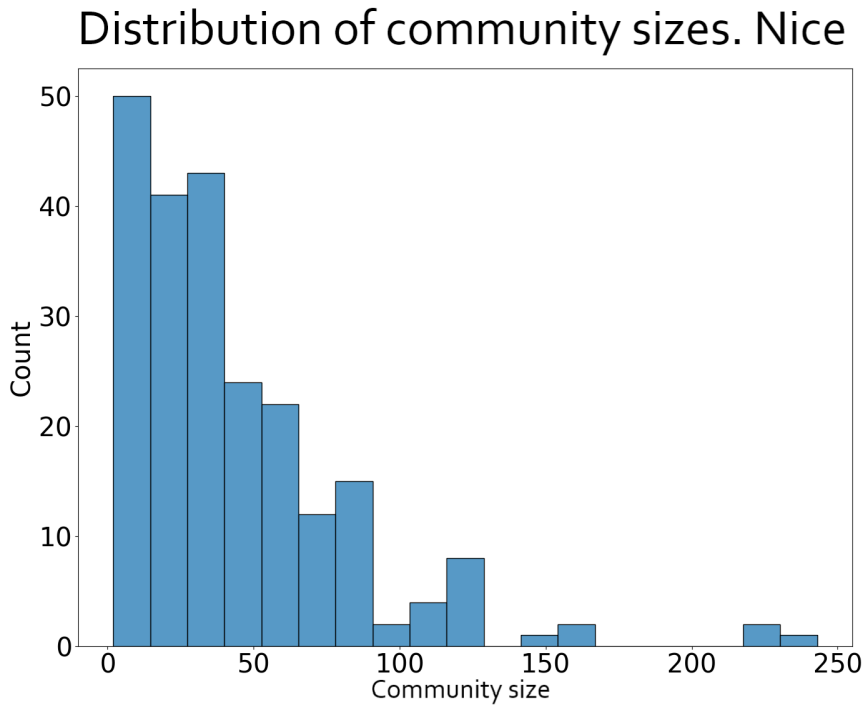



Figure 3.16: Distribution of community sizes after splitting largest communities.

Also, community stability had to be checked. Same approach as in Section 3.2 was used, to find how consistent newly defined communities are if current partitioning is applied on Nice-Sophia-2013 network. Matrix of community matchings (Figure 3.17) shows that there is still high number of similar communities in both networks (Nice-Sophia-2018 and Nice-Sophia-2013). Among all members of Nice-Sophia-2013 network, about 81% of them stayed together in the same community in 2018 (5200 out of 6395 authors).

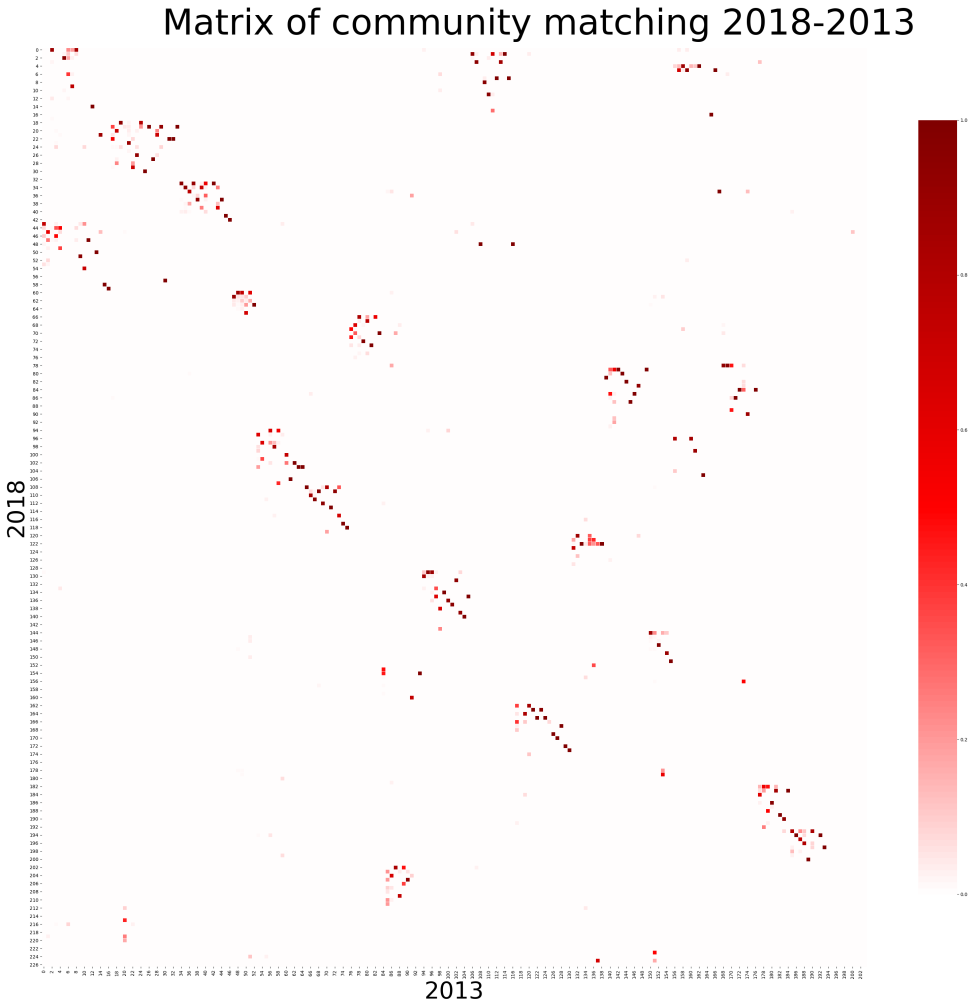



Figure 3.17: Matrix of community matchings after splitting largest communities.

Among 227 communities, there are 32 distinct pairs of communities that supervised PhD works. Recall that we compute distinct pairs as all combinations of communities involved in PhD work. Thus, if PhD student and their two supervisors were assigned to 3 different communities, there will be 3 distinct community-pairs that were involved with this PhD. These 32 distinct pairs of communities involve 21 of 25 PhDs with LabEx scholarship. This means that such com-

munity sizes are much closer to real life data, since LabEx funding promotes PhDs that are supervised by at least two different research teams or universities, etc.

When difference matrix was computed, we got the following results: (see Figure 3.18 and Figure 3.19)

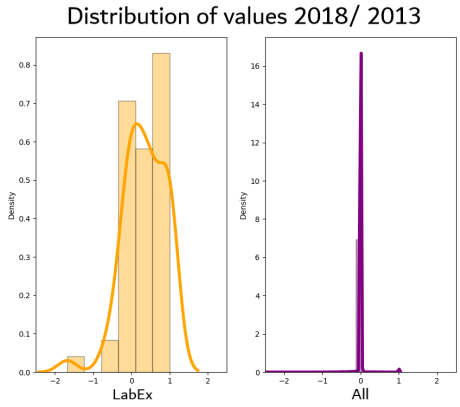



Figure 3.18: Distribution of values of difference matrix $M$.

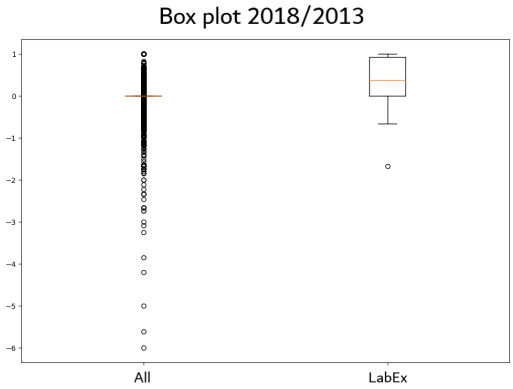



Figure 3.19: Corresponding box plot for distribution of values of difference matrix $M$.

| | Median | Average |
|---|---|---|
| LabEx | 0.375 | 0.343 |
| All | 0.000 | 0.000142 |

Despite the fact that we still have a small number of observable communities with LabEx PhD students, now with smaller communities, we can better see changes in productivities of these communities. In this way, we can suppose that such approach with community size reduction will help us in further investigation of France Computer Science network.

# 4 Impact of PhD Works in France Computer Science Network

## 4.1 France Computer Science Network

*France Computer Science network* is built similarly to Nice-Sophia network. It was built based on Scopus data on articles published between 1990 and 2018, where at least one of the authors was affiliated to France. In total, 239 414 publications were used to build the France Computer Science network. This network has 258 145 nodes and 1 591 382 edges, and its largest component consists of 229 322 nodes and 1 533 435 edges.

On such large network, more data was needed to be used to investigate the impact of PhD theses on pairs of communities that supervised these PhD works. The data on PhD works were retrieved from Theses.fr [19] online database. For more information, you can refer to Section 2.1.2.

After preprocessing of the information on PhD works it was identified that before 2010 were only 35 PhD works with 2 supervisors, which were present in the France Computer Science network (more

could be excluded because some supervisors could not be uniquely identified in the network). In total, 5595 PhD theses were used for this investigation.

## 4.2 Large Community Problem

Leiden algorithm for community detection discovered 160 communities in the France Computer Science network.

Modularity of this partition was 0.861, and sizes of these communities ranged from 15 to 15 171 members.

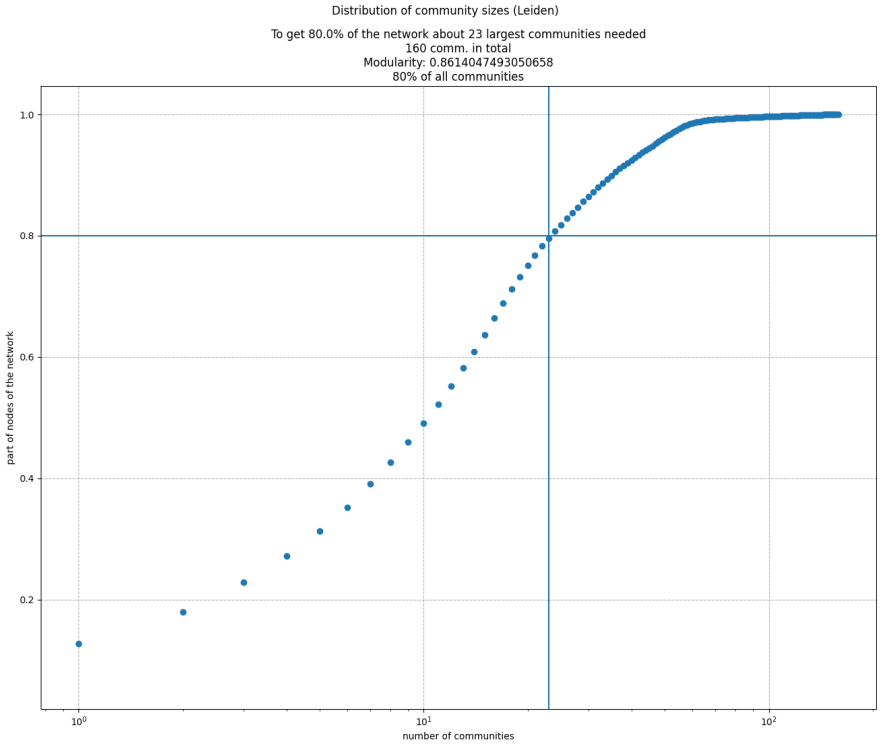



Figure 4.1: Distribution of sizes of the largest communities in France Computer Science network. It shows what number of the largest communities should be taken to get 80% of the network.

To get about 80% of the network, only 23 communities are needed (see Figure 4.1). These 23 communities consist of 179 606 authors and

now the smallest community size is 3564 (see Figure 4.2).

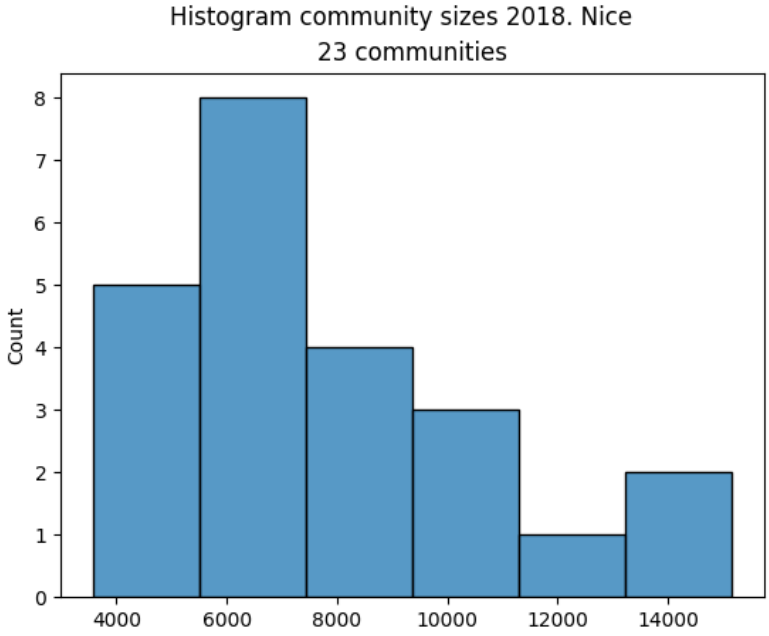
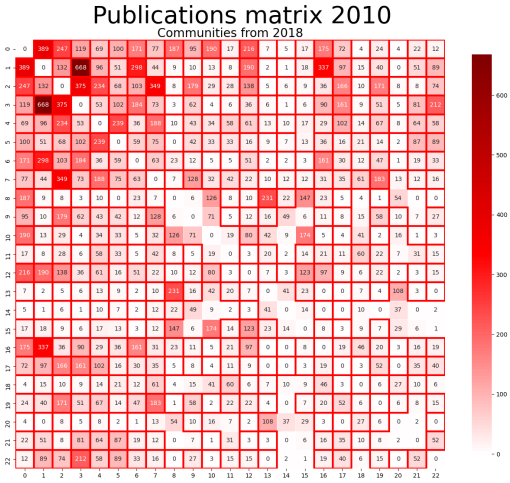
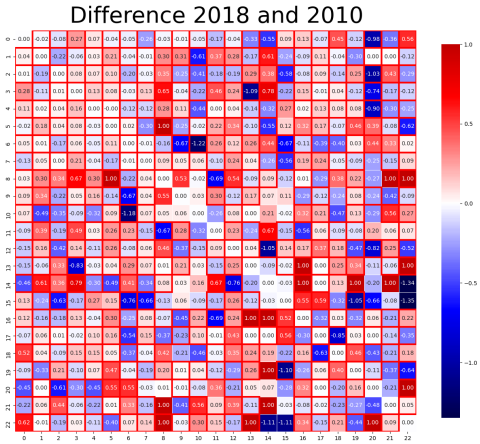



Figure 4.2: Distribution of sizes of the largest communities.

As expected, too big communities lead to a problem that community productivity is hidden inside a large cluster.




(a) Matrix of number of publications.

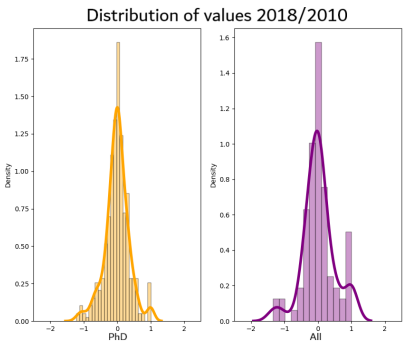



(b) Difference matrix on largest communities.

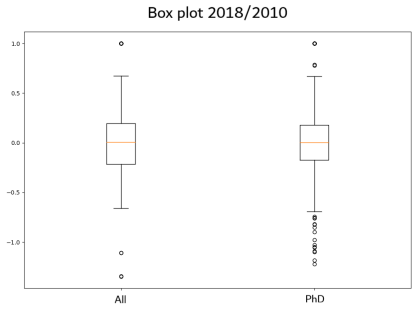
Figure 4.3: Almost every pair of communities has PhD work. Cells emphasized with red square correspond to communities with PhD students.

As a result, in our case, almost every pair of communities has supervised at least one PhD thesis (see Figure 4.3). The key metrics in such case show that pairs of communities that supervised PhD had a decrease in intensity of their collaborations as well as other pairs of communities without PhD (see Figure 4.4).

| | Median | Average |
|---|---|---|
| PhD | 0.000 | -0.0129 |
| All | 0.039 | -0.002 |




(a) Distribution of values of difference matrix $M$.




(b) Corresponding box plot of values of difference matrix $M$ for both groups.

Figure 4.4: Inside of difference matrix $M$

However, we suppose that this is not the real case for 2 reasons.

Firstly, such large communities do not represent the real state of the researchers' network. For example, most of the universities have smaller academic staff than the minimum community size that we got with the Leiden algorithm with the current network configuration (see Section 2.2).

Also, the problem of the optimal size of human communities are discussed in multiple works [7, 9, 18], where different real-world net-

works were investigated, and it was shown that usually community sizes range between 5 and 1500 members.

Secondly, when we looked into the distribution of the number of PhD students and the number of publications that they co-authored (see Figure 4.5), we found that 20 812 publications were co-authored by these PhD students. We expect that such productivity of PhD students must have some influence on community collaborations, and it should be reflected in the difference matrix $M$.

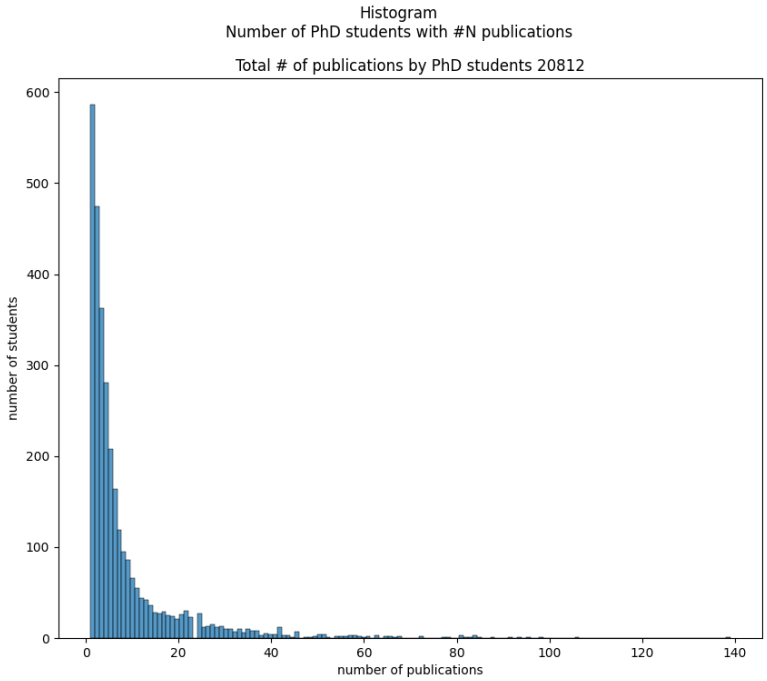



Figure 4.5: Distribution of number of PhD students who co-authored a given number of publications.

## 4.3 Results With Reduced Community Size

To reduce community size the same method described in Section 3.5 was used. After that, each of 23 communities was split into $\approx$ 41 smaller communities. As a result, from 23 large community 952 were produced, with sizes ranged from 4 to 1370 members (see Figure 4.6).

**Note.** *Such community sizes appear in above-mentioned scientific works on investigation of communities in Human Networks [7, 9, 18].*

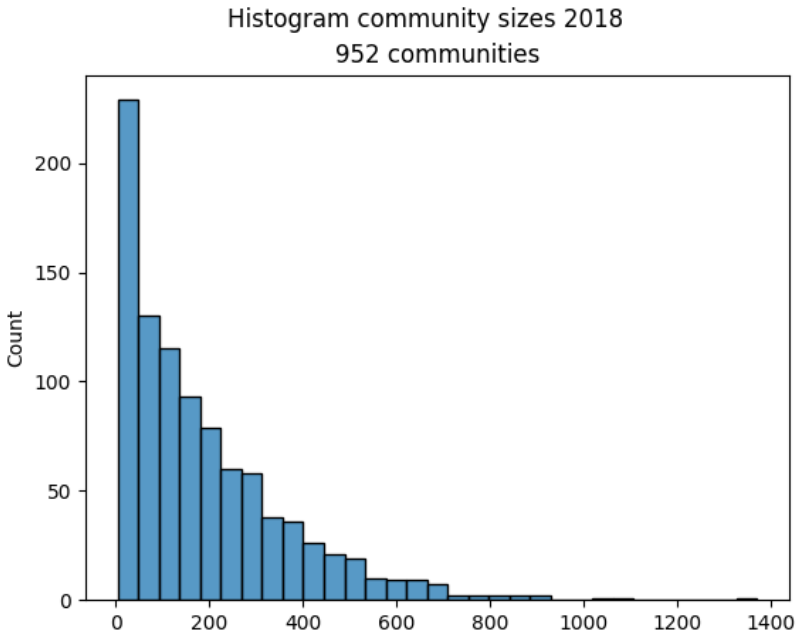



Figure 4.6: Distribution of sizes of the largest communities.

### 4.3.1 Smaller communities — better picture

On almost a thousand smaller communities, we got 3499 distinct community pairs that supervised PhD theses, which is 6998 PhD values in difference matrix $M$.

Since the difference matrix $M$ is too big to be presented here, we proceed to analyse its contents. The table below shows the key metrics of both groups of $m_{i,j}$ values, which were also organized in box plot and histograms in Figure 4.7 and Figure 4.8.

| | Median | Average |
|---|---|---|
| PhD | 0.000 | 0.0187 |
| All | 0.000 | -0.00141 |

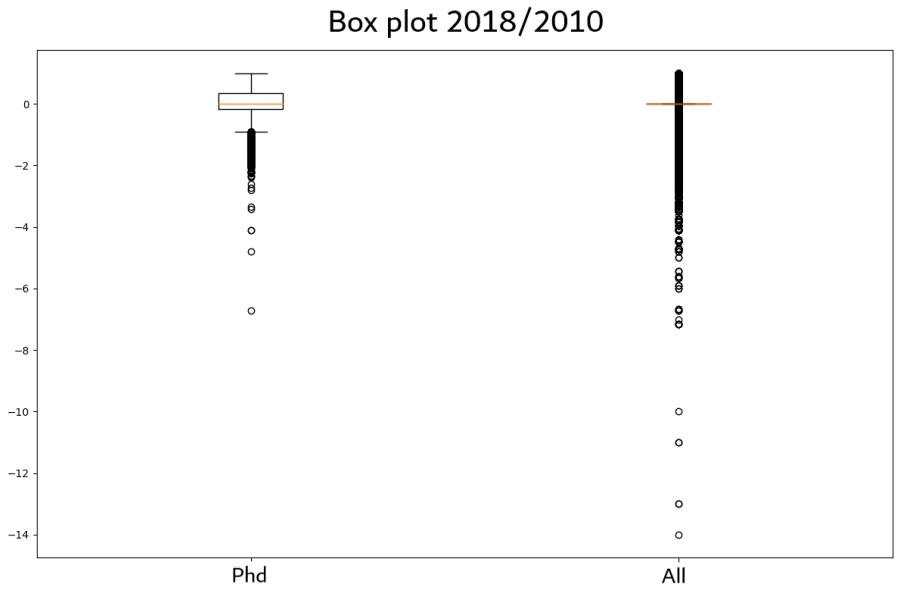



Figure 4.7: Box plot of distribution of $m_{i,j}$ values.

As you can see, that smaller communities indeed give a much clearer picture than large ones. From this data, we can see how different is the shape of the "progress" of the "PhD" group and All other groups without PhD, in spite of the fact that the key metrics are very low.

However, these picks on zeros and ones on the histogram in Figure 4.8 drew our attention. When we analysed obtained values of each group, we got the following numbers of occurrences of 0 and 1 in each group.

| | 0 | 1 | In total |
|---|---|---|---|
| PhD | 2513 | 1050 | 6998 |
| All | 855 424 | 16 032 | 898 354 |

In the table above one may notice that the fraction of 1's (which indicate the appearance of new collaboration after 2010) in case of the "PhD" group is significantly higher than in the "All" group.

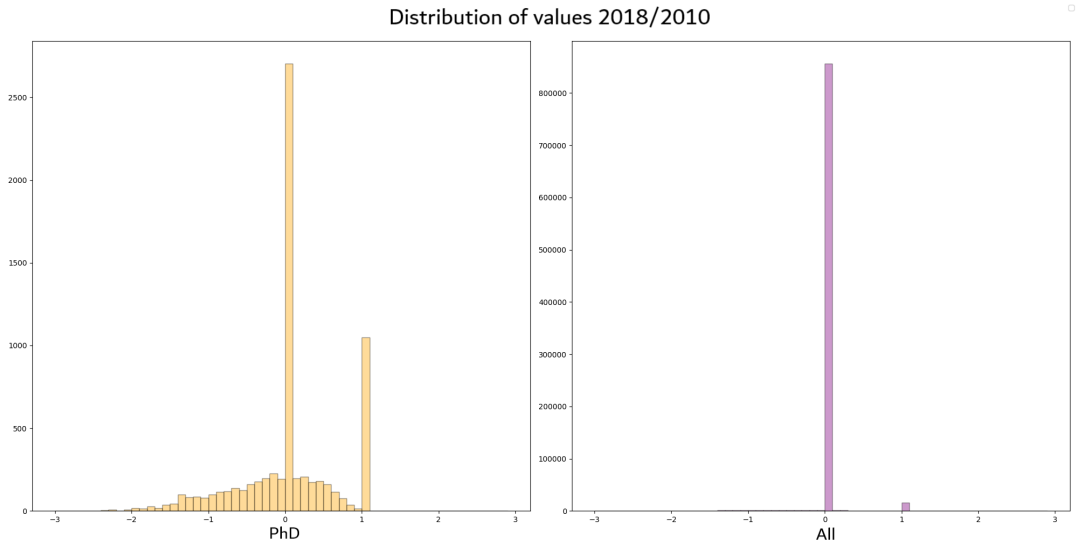



Figure 4.8: Distribution of values of difference matrix $M$.

## 4.4 New Metric

Smaller communities indeed allow us to see small changes in community collaborations. However, it is not enough to draw some conclusions with current results.

Therefore, we try another method of computing the difference matrix $M$. Even though we know that this method carries the same information as the previous one, it will be another way of representing the same concept, but it may give us a new insight into the problem.

According to current approach, it is possible that the difference between probabilities to make common publication between two communities can take very small negative value.

For example, in Figure 4.9 there are three communities $C1, C2, C3$. For community $C1$, the probability to make publication with community $C2$ in 2013 is 0.5. However, in 5 years community $C1$ collaborated with community $C3$ very actively, but never with community $C2$ and in 2018 the probability to make common publication of communities $C1$ and $C2$ dropped to 0.09. Thus,

$$m_{C1,C2} = \frac{w_{C1,C2}(2018) - w_{C1,C2}(2013)}{w_{C1,C2}(2018)} = \frac{0.09 - 0.5}{0.09} = -4.55$$

2013

| | C1 | C2 | C3 |
|---|---|---|---|
| C1 | 0 | 1 | 1 |
| C2 | 1 | 0 | 0 |
| C3 | 1 | 0 | 0 |

| | C1 | C2 | C3 |
|---|---|---|---|
| C1 | 0 | 0.5 | 0.5 |
| C2 | 1 | 0 | 0 |
| C3 | 1 | 0 | 0 |

2018

| | C1 | C2 | C3 |
|---|---|---|---|
| C1 | 0 | 1 | 10 |
| C2 | 1 | 0 | 0 |
| C3 | 10 | 0 | 0 |

| | C1 | C2 | C3 |
|---|---|---|---|
| C1 | 0 | 0.09 | 0.91 |
| C2 | 1 | 0 | 0 |
| C3 | 1 | 0 | 0 |

Figure 4.9: Matrices of number of publications and corresponding matrices $W$ of probability to make publication between communities.

In order to work with positive numbers, a new difference matrix is introduced below.

$M_{new} = (m_{i,j})^{k \times k}$, where:

$$m_{i,j} = \begin{cases} -1 & \text{if } w_{i,j}(2018) = 0 \text{ and } w_{i,j}(2010) = 0, \\ K + \delta & \text{if } w_{i,j}(2018) \neq 0 \text{ and } w_{i,j}(2010) = 0, \\ \dfrac{w_{i,j}(2018)}{w_{i,j}(2010)} & \text{otherwise,} \end{cases}$$

$\forall i, j \in \{1, \ldots, k\}$

Where K is $\max(\frac{w_{i,j}(2018)}{w_{i,j}(2010)})$, for all $i, j \in \{1, \ldots, k\}$, such that $w_{i,j}(2010) \neq 0$.

The first condition covers the case when collaboration never existed.

The second condition covers the case when a new collaboration appeared in 8 years. Here we take the largest value in the matrix $M_{new}$, which indicates the highest gain in probability to make publication and add some $\delta > 0$ to distinguish this case, from high gain values.

And finally, the last $m_{i,j}$ will show the gain coefficient in productivity of communities $i$ and $j$, when $w_{i,j}(2018) \neq 0$ and $w_{i,j}(2010) \neq 0$.

The value of $m_{i,j}$ in this case may be roughly interpreted as follows:

- $m_{i,j} \in [0, 1)$ – the intensity of collaboration has dropped since 2010;
- $m_{i,j} = 1$ – the intensity of collaboration stayed more or less at the same level since 2010;
- $m_{i,j} \in (1, +\infty)$ – the intensity of collaboration increased since 2010 - the probability for common publication between communities $c_i$ and $c_j$ (when we take as a base the number of publications of community $c_i$) increased $m_{i,j}$ times.

### 4.4.1 Results With New Matrix $M_{new}$

When a difference matrix $M_{new}$ was computed according to method described in Section 4.4, we got the following results:

| | Median | Average |
|---|---|---|
| PhD | 0.622 | 3.716 |
| All | 0.000 | 0.398 |

According to key metrics, the intensity of collaboration increased significantly in *"PhD" group*[1]. For communities that supervised PhD thesis, the probability to publish in 2018 is 3.7 times higher than in 2010. While for communities without PhD students, we can interpret that the intensity of collaboration (understood as the probability to publish together) dropped to the 40% of the level from 2010. In figures below you can see the shape of $m_{i,j}$ values of both groups (see Figures 4.10 and 4.12).

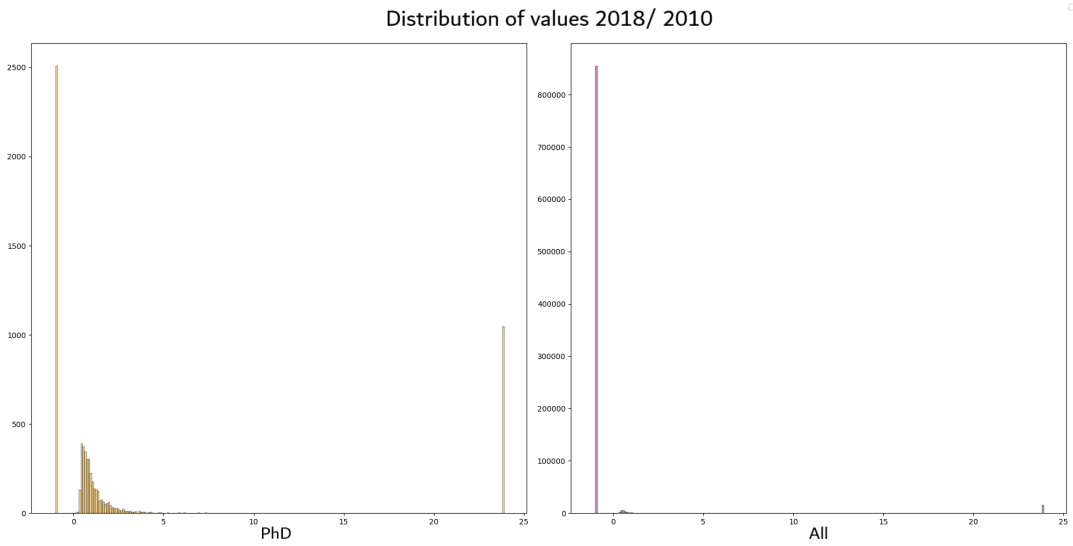



Figure 4.10: Distribution of values of difference matrix $M_{new}$.

The Figure 4.11 is zoomed part of Figure 4.10. It shows the distribution of values of difference matrix $M_{new}$ that fall into $[0, 2]$.

[1] "PhD" group is a group of values that correspond to communities that supervised PhD works

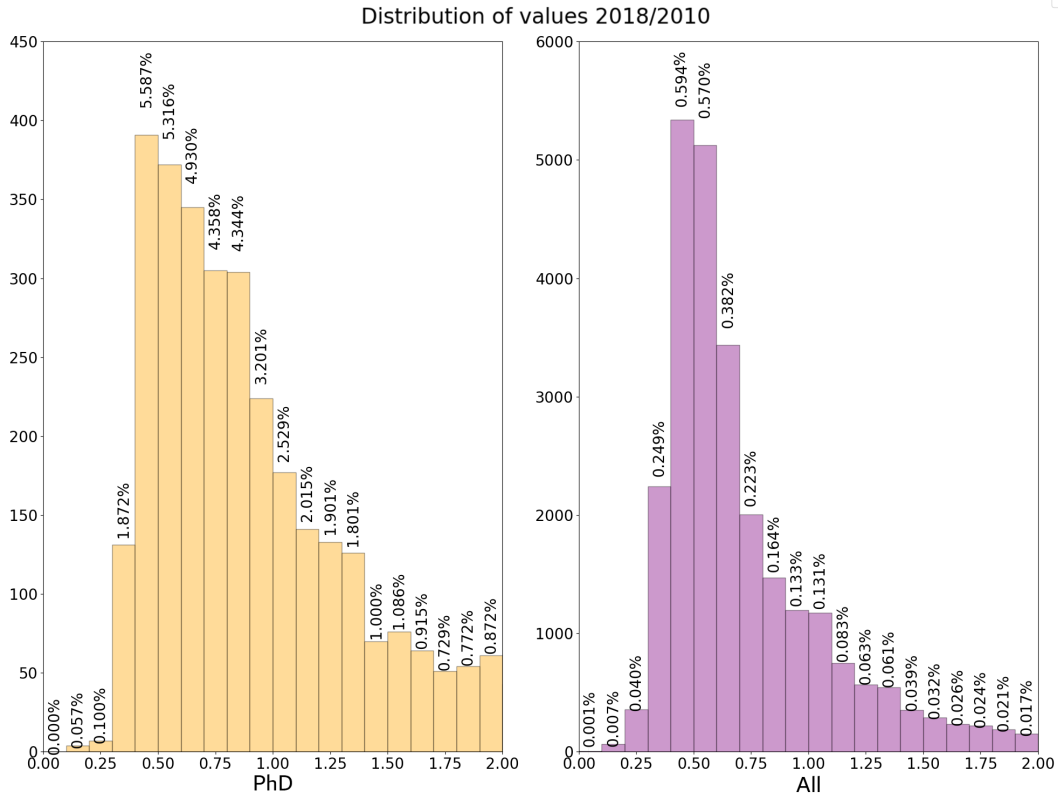



Figure 4.11: Zoom of interval $[0, 2]$ of Figure 4.10. Distribution of values of difference matrix $M_{new}$ with percent to all other values in a group.

In Figure 4.10 we can see that the part of community pairs that have increased their collaboration productivity is higher in "PhD" group than in "All" group. The table below shows what fraction of all $m_{i,j}$ values fall into 4 intervals.

| | -1 | [0,1] | (1, max($M_{new}$)) | max($M_{new}$) |
|---|---|---|---|---|
| PhD | 35.867% | 29.765% | 19.362% | 15.006% |
| All | 95.205% | 2.364% | 0.645% | 1.786% |

Let us present the above results, splitting the set of pairs of communities into two subsets: pairs of communities that did not collaborate before 2010 and pairs that collaborated before 2010.

| Fraction of new collaborations | |
|---|---|
| PhD | 0.295 |
| All | 0.018 |

Table 4.1: Pairs of communities which did not collaborate before 2010.

| Fraction with increase intencity of collaboration ($m_{i,j} > 1$) | |
|---|---|
| PhD | 0.394 |
| All | 0.214 |

Table 4.2: Pairs of communities which collaborated before 2010.

From the set of pairs of communities that did not collaborate before 2010 (Table 4.1) we can see that among community pairs that supervised a common PhD work there is a larger fraction of newly appeared collaborations than in other community pairs. Moreover, Table 4.2 shows that"PhD" groups are also leading in the proportions of community pairs that have increased their intensity of collaboration between 2018 and 2010.

Thus, we can say that communities that supervised PhD work became more productive within 8 year, unlike pairs of communities that did not.

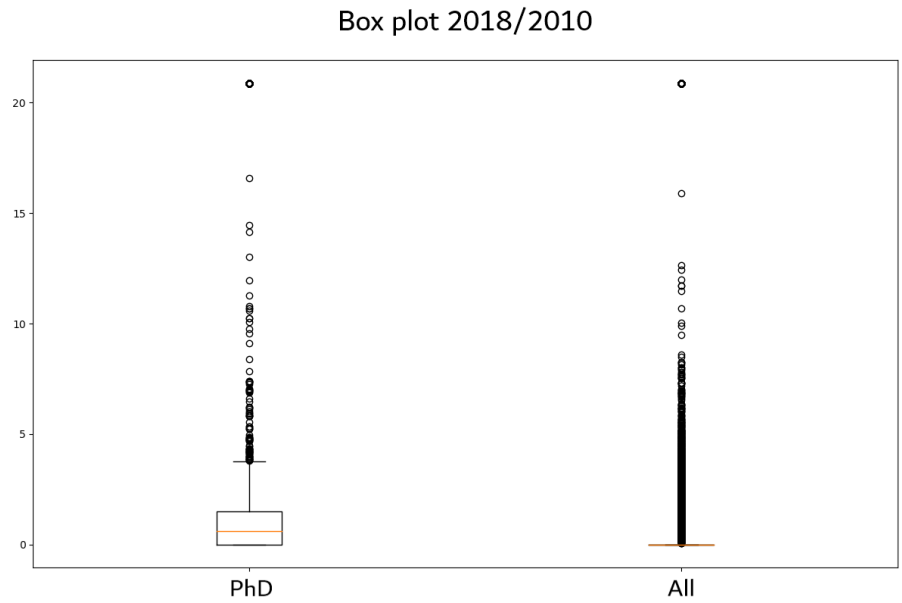



Figure 4.12: Corresponding box plot of values of difference matrix $M$ for both groups.

With such approach of investigation of community productivity, we can see that communities that supervised PhD works continue to collaborate more actively.

# 5 Conclusion

Within this Master thesis, several approaches for measuring the evolution of community productivities were developed.

The first results on the Nice Sophia Antipolis Computer Science network showed that LabEx PhD theses had a positive influence on future collaborations between communities that supervised these PhD works (Section 3.4). After that, community sizes were reduced by running Leiden algorithm [20] for community detection on each community after first partitioning (Section 3.5). This approach allowed us to work with communities of size from 2 to 243 members instead of communities from 261 to 915 members. Finally, we saw that on smaller communities the impact is more visible than on the large ones.

Then, when we tried to reproduce the results from Nice Sophia network on almost 18 times larger network, we encountered a problem with large communities (from 3564 to 15 171 members), which made us to use the same method for community size reduction. This approach slightly improved the results, however, to draw some conclusions this was not enough.

Finally, to improve the approach for comparing changes in productivity of communities, another method for computing the difference matrix was used (Section 4.4). This computation method showed that the intensity of cooperation between communities that supervised a common PhD thesis increased 3.7 times, while in other communities

we can see the decrease in intensity of their cooperation to 40% of its level at the beginning of investigated timespan.

Hence, according to the proposed method for studying the impact of PhD theses on collaboration networks, a PhD thesis supervised by several communities can increase the future collaborations between these communities.

## 5.1 Further Work

For further work, the following problems could be considered:

1. This investigation was conducted on Computer Science network, considering only *interdisciplinary* collaborations. It also would be interesting to study *multidisciplinary* collaborations, and to see whether multidisciplinary funding programs have the similar impact on intensity of future collaborations.

2. Develop a model which would allow making predictions about the impact of funding on future collaborations.

# References


[1] LabEx. https://univ-cotedazur.eu/ucajedi-idex-of-universite-cote-dazur/labex. Accessed: 2021-08-30.

[2] Polytech Nice Sophia. https://polytech.univ-cotedazur.fr/. Accessed: 2021-08-30.

[3] Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018. URL: http://jmlr.org/papers/v18/16-480.html.

[4] CDlib is a Python software package that allows to extract, compare and evaluate communities from complex networks. https://cdlib.readthedocs.io/. Accessed: 2021-08-30.

[5] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), Dec 2004. URL: http://dx.doi.org/10.1103/PhysRevE.70.066111, doi:10.1103/physreve.70.066111.

[6] Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data. https://cytoscape.org/. Accessed: 2021-08-30.

[7] Robin I.M. Dunbar and Richard Sosis. Optimising human community sizes. *Evolution and Human Behavior*, 39(1):106–111, 2018. URL: https://www.sciencedirect.com/science/article/pii/S109051381730209X, doi:https://doi.org/10.1016/j.evolhumbehav.2017.11.001.

[8] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, and et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi:10.1038/s41586-020-2649-2.

[9] Russell Hill and Robin I.M. Dunbar. Social network size in humans. *Human Nature*, 14:53–72, 2003. URL: https://pubmed.ncbi.nlm.nih.gov/22506741/, doi:https://doi.org/10.1007/s12110-003-1016-y.

[10] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. URL: https://www.sciencedirect.com/science/article/pii/0378873383900217, doi:https://doi.org/10.1016/0378-8733(83)90021-7.

[11] Clement Lee and Darren J. Wilkinson. A review of stochastic block models and extensions for graph clustering. *Appl Netw Sci 4, 122*, 2019. doi:10.1007/s41109-019-0232-2.

[12] Matplotlib: a comprehensive library for creating static, animated, and interactive visualizations in Python. https://matplotlib.org/. Accessed: 2021-08-24.

[13] NetworkX: a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks. https://networkx.org/. Accessed: 2021-08-24.

[14] Mark Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), Feb 2004. URL: http://dx.doi.org/10.1103/PhysRevE.69.026113, doi:10.1103/physreve.69.026113.

[15] Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. https://pandas.pydata.org/. Accessed: 2021-08-30.

[16] Scopus: Elsevier's abstract and citation database. https://www.scopus.com/. Accessed: 2021-08-24.

[17] Scientific networks and IdEx funding. https://ds4h.univ-cotedazur.eu/research-and-labs/funded-projects-and-calls/project-snif. Accessed: 2021-01-24.

[18] Alistair Sutcliffe, Robin Dunbar, Jens Binder, and Holly Arrow. Relationships and the social brain: integrating psychological and evolutionary perspectives. *British journal of psychology (London, England : 1953)*, 103(2):149–68, 2012. URL: https://pubmed.ncbi.nlm.nih.gov/22506741/, doi:https://doi.org/10.1111/j.2044-8295.2011.02061.x.

[19] theses.fr is a search engine to find French doctoral theses. http://www.theses.fr/. Accessed: 2021-08-24.

[20] Vincent Traag, Ludo Waltman, and Nees Jan van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9, March 2019. doi:10.1038/s41598-019-41695-z.

[21] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi:10.21105/joss.03021.