ENS 491/2 - Graduation Project
Progress Report II

Understanding the Academic World

Submitted by
Berkan Teber

Under the supervision of
Kamer Kaya
Hüsnü Yenigün

2017 - 2018

Sabancı University
Faculty of Engineering and Natural Sciences

Sabancı
Üniversitesi

1 Project Objectives

In this project, our goal is to have a better understanding of the academic world. To achieve this goal we will analyze papers, the building blocks of the academic world, and citations, the interaction between these papers. At the end of this project, we expect to have a method to evaluate authors with respect to different fields and the resulting evaluations. In addition, we expect to have a method to compare different fields and to find the most promising ones among them. To obtain the intended results and to achieve our goal, we have determined various objectives to fulfill throughout the project.

First objective was to preprocess the data we have. Since the data we have is huge, we needed some preprocessing before we work on them. This preprocessing step includes extraction of the relevant information and construction of a more accessible data structure for our future work. This step has been mostly completed in the period until Progress Report 1.

Second objective was to identify different fields. This identification has been performed by running a community detection algorithm such as the Louvain Method on our network and analyzing the comparison of the distribution of keywords within a community and the general distribution of keywords.

Third objective is to evaluate papers and authors with respect to the whole network, and then, with respect to different fields. While this evaluations can be performed in many ways, such as h-index which uses citation counts as the only criteria, we will use a Random Walk based model for our evaluations.

Fourth objective is to compare different fields with each other in various ways. As results of these comparisons, we expect to identify some attributes that a field behaves different than the other fields. In addition, we expect to identify the promising fields among all fields looking at these attributes that makes a field distinguishable.

Last objective is to represent the results obtained in previous steps such as the paper and author evaluations or the attributes of the fields that distinguishes one from each other or the list of promising fields with various visualization techniques.

2 Performed Tasks

2.1 Cartel Detection

We believe that the scientific society includes many cartels. In our definition, a cartel is the following: a group of people who don't work together that cites each other extensively to increase the citation number of each other.

To detect cartels we have performed community detection using the Louvain method [3] in the author cites author network.

2.1.1 Cartel Detection in the Original Network

As the first step, we have created the author cites author network with from the data we have. Since the data is huge, we have eliminated the edges with weights 1 and 2 (i.e. if an author cites another author 2 times only, we eliminated this edge). The idea behind this was that these edges are most likely to be outside of the community edges, so they were redundant.

As the second step, we have scored each community of size 3 to 20 with respect to the average ratio of the citations going inside of the community to all outgoing citations.

Below, you can find the pseudocode to score each community:

Algorithm 1 Scoring of a Community

```
procedure ScoreCommunity($V$, $E$, $C$)
    $totalscore \leftarrow 0$
    for $v \in C$ do
        $numerator \leftarrow 0$
        $denumerator \leftarrow 1$
        for $((v, u), w) \in E$ do
            if $u \in C$ then
                $numerator \leftarrow numerator + w$
            end if
            $denumerator \leftarrow denumerator + w$
        end for
        $totalscore \leftarrow totalscore + numerator/denumerator$
    end for
    return $totalscore/|C|$
end procedure
```

When we have looked at the selected communities from top-scored 100 communities, we have realized that the communities we have identified are not cartels but people working together.

For this reason we have decided to eliminate self citations while constructing the author cites author graph.

2.1.2 Cartel Detection in the Network without Self Citations

To distinguish between cartels and the working groups that we have identified previously, we have decided to create an author cites author network without self citations. In this context, self citations means the following:

Assume a paper A written by authors 1, 2 and 3, and another paper B written by 1, 2 and 4. In addition, assume that paper A cites paper B. Then, we have added only the edges $(3, 1)$, $(3, 2)$ and $(3, 4)$ to the network.

Additionally, since the data is huge, we have eliminated the edges with weight 1 (i.e. if an author cites another author only once, we eliminated this edge). The idea behind this was that these edges are most likely to be outside of the community edges, so they were redundant.

As the second step, we have scored each community of size 3 to 20 with respect to the average ratio of the citations going inside of the community to all outgoing citations.

Scoring the communities have been performed in the same way in the original network and the network without self citations.

In the latter approach, we have seen that the citation counts were too little to form a cartel.

An exemplary community which has been the top-scored community in the scoring process can be seen below:

```
10068631  0.573129  7

David R Mekala
Gary William Schukar  3  IN
John Russell Mlinar  5  IN
Scott Alan Ripley  2  IN

Dennis E Ferguson
Gary William Schukar  3  IN
John Russell Mlinar  5  IN
Scott Alan Ripley  2  IN

Donald G Peterson
Gary William Schukar  3  IN
John Russell Mlinar  5  IN
Scott Alan Ripley  2  IN

Duane D Fansler
Alexander M. Klibanov  2  OUT
Dean S. Milbrath  3  OUT
Gary William Schukar  3  IN
Jerald K. Rasmussen  2  OUT
John Russell Mlinar  5  IN
Scott Alan Ripley  2  IN

Gary William Schukar
David R Mekala  2  IN
Dennis E Ferguson  2  IN
Donald G Peterson  2  IN
Duane D Fansler  2  IN
Gareth Monkman  3  OUT

John Russell Mlinar
David R Mekala  3  IN
Dennis E Ferguson  3  IN
Donald G Peterson  3  IN
Duane D Fansler  3  IN
Gareth Monkman  3  OUT

Scott Alan Ripley
```

2.2 Word Clouds for Fields

In the previous report, we have already identified the top 5 communities as follows:

1. Social Sciences
2. Computer Science
3. Physics & Chemistry
4. Biology
5. Medicine

To clarify and support this identification, we have created word clouds using the top 30 keywords and fields of study for each of the communities.

Below you can find these word clouds for each communities.

2.2.1 Social Sciences

According to the word clouds we have generated the 1st largest community is the social sciences community.

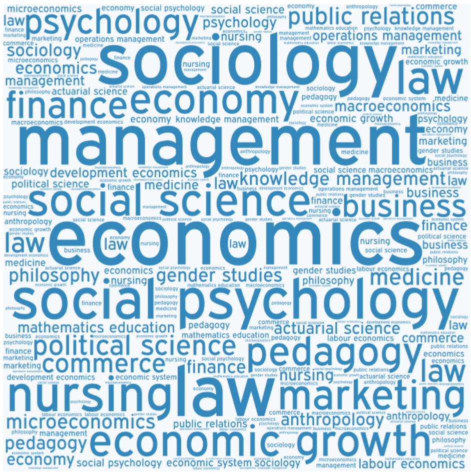

Figure 1: Fields of Study

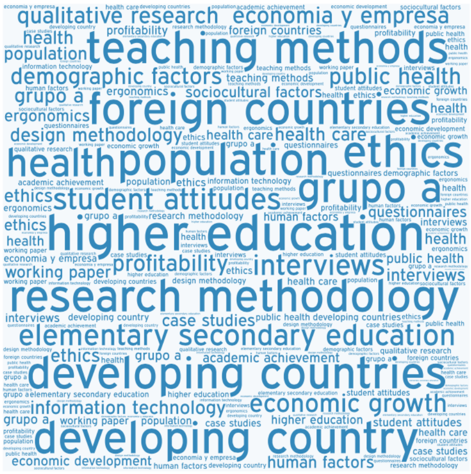

Figure 2: Keywords

2.2.2 Computer Science

According to the word clouds we have generated the 2nd largest community is the computer science community.

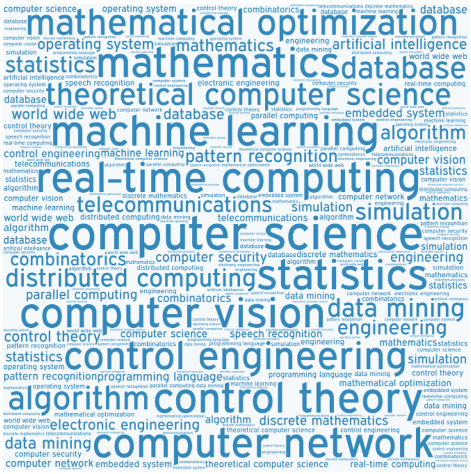

Figure 3: Fields of Study

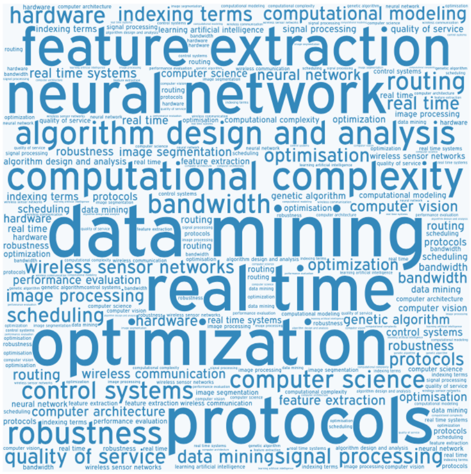

Figure 4: Keywords

2.2.3 Physics & Chemistry

According to the word clouds we have generated the 3rd largest community is the physics and chemistry community.

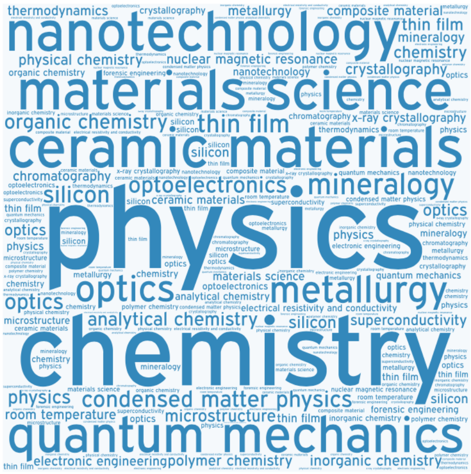

Figure 5: Fields of Study

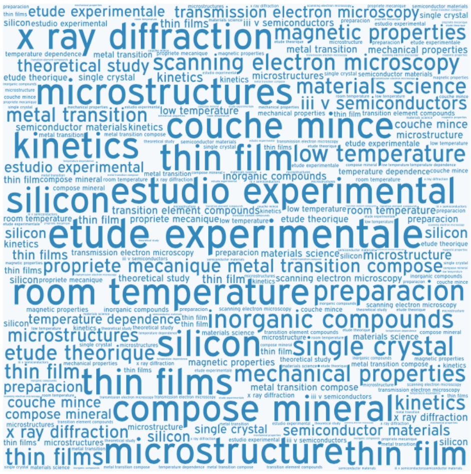

Figure 6: Keywords

2.2.4 Biology

According to the word clouds we have generated the 4th largest community is the biology community.

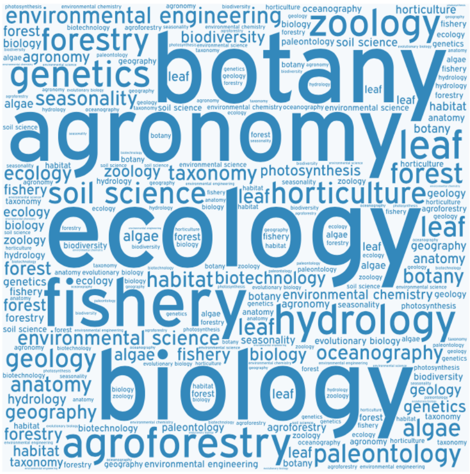

Figure 7: Fields of Study

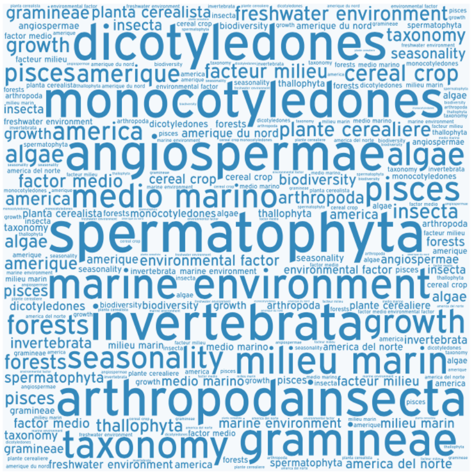

Figure 8: Keywords

2.2.5 Medicine

According to the word clouds I have generated the 5th largest community is the medicine community.

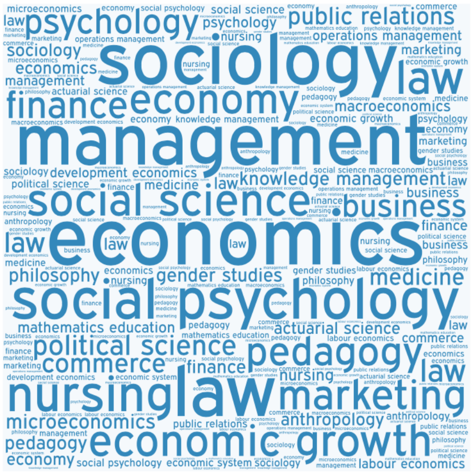

Figure 9: Fields of Study

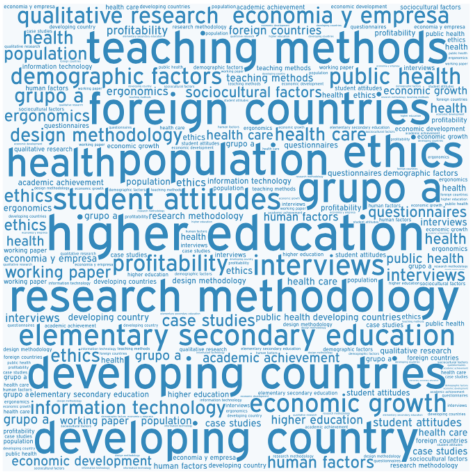

Figure 10: Keywords

2.3 Isolating Communities

In this taski we have constructed an isolated network for top 5 communities. When isolating the network, we have used the following method: take only the vertices belonging to that community and take only the edges from a vertex inside of the community to a vertex inside a community.

3 Changes in the Goals

One of our tasks was the Author Name Disambiguation. However, since we haven't designed an algorithm for such a task yet, and also even we had designed such an algorithm, to test the correctness of this algorithm without a control group is hard, we will most likely proceed without any disambiguation/unification regarding author names.

Everything else goes as expected, and there is no other changes in the objectives or the plans.

4 Future Tasks

4.1 Computing Different Indexes

There are several indexes proposed to evaluate an author. We have already identified some of these indexes in the Proposal. Also, we have already computed h-index for each author. As a next step, we will compute some more indexes such as g-index and h'-index.

4.2 Computing Random Walk Based Indexes

Currently, another group of the same project has been working on to compute Random Walk based index for each author. As a next step, we will modify this computation such that we take the time factor into account.

4.3 Computing Indexes for Different Fields

We have already identified different fields and isolated our graph for each of these fields. As a next step, we will repeat our computation with both h-index and similar indexes, and random walk based index on these isolated graphs.

4.4 Demonstrating Results

The main objective of this project is understanding the academic world better. After all the computations, evaluations and comparisons we expect to have a better understanding about the scientific society. However, we also need to demonstrate our findings to other people. For this reason, as the last step, we will develop a system to visualize our results.

References

[1] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. (2008). *ArnetMiner: Extraction and Mining of Academic Social Networks.* In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: p: 990-998.

[2] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. (2015). *An Overview of Microsoft Academic Service (MAS) and Applications.* In Proceedings of the 24th International Conference on World Wide Web: p. 243-246

[3] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. (2008). *Fast Unfolding of Communities in Large Networks.* In Journal of Statistical Mechanics: Theory and Experiment 2008 (10): P10008 (12pp).