ENS 491/2 - Graduation Project
Project Proposal

Understanding the Academic World

Submitted by
Berkan Teber

Under the supervision of
Kamer Kaya

2017 - 2018

Sabancı University
Faculty of Engineering and Natural Sciences

Abstract

Papers published can be considered as the building blocks of the academic world. With the technological and scientific development, a huge increase in the number of researches and research papers has been observed. Today, we have very large amounts of data about research papers and citation connections between them. In this project, we aim to construct a network from these papers and citations we have, and analyze this huge network. We believe that, an analysis of this network will lead us to understand the academic world better by helping us to answer some very important questions about it such as "Which papers are the most influential in their field?", "Which researchers are more successful?" and "Which scientific areas are the most promising?". Despite the fact that various evaluation methods have been proposed in the past to give a correct answer to these questions, most of these methods have their shortcomings. By applying different network analysis techniques, we intend to propose more realistic answers to these questions.

1 Introduction

1.1 Motivation

The academic world grows very rapidly in terms of both the number of new publications and the variety of scientific areas. With this rapid growth, it becomes much harder to understand and follow the academic world. This problem makes it necessary to develop some new methods to help us understand the academic world in a better way.

In the past, there have been proposed some methods to evaluate the quality of researches and the quality of researchers, however these evaluations were mainly based on evaluating the research and the researcher as its own. These evaluations that solely based on the quantitative data of the research or the researcher as its own have some problems to correctly evaluate researches or researchers. For example, having citation count as a measure of quality of a paper, doesn't take qualities of the citations into account. Similarly, h-index, an index to evaluate the researchers, doesn't take the field of the research into account.

In this project, we will try to understand the academic world not solely on numbers such as citation count but also according to the place of the research or the researcher in the whole scientific network. Throughout the project, we will first construct a network based on the available data. Then, we will use different network analysis techniques to analyze this network. Finally, we will develop a web application to present our results.

1.2 Problem Definition

In this project, our ultimate goal is to understand the academic world in a better way. To achieve this goal, we need to find solutions to a variety of subproblems. These subproblems will depend on each other. In other words, in each subproblem, we will use the solution proposed for the earlier subproblems. In the end, we aim to have a much better understanding of the academic world.

The first subproblem is about evaluating the quality of papers using the paper-citation network. There are already some measures such as citation count to measure the quality of papers. However, these measures have their own problems, and these problems are generally because of that they only take evaluated papers and their citations into account. In this project, we will evaluate papers according to their importance in the whole network, taking all papers into account. By doing so, we believe that we will have a more accurate evaluation of papers.

The next subproblem is about evaluating researchers. In the past, some indexes have been proposed for this purpose including h-index, the most commonly used index to evaluate researchers today. However, such indexes have a major problem: they generally take only the citation counts into account to evaluate papers. Therefore, they are problematic for the same reasons that measuring the quality of a paper solely on the citation counts are. In this project, we will evaluate researchers based on the evaluation of their papers according to the whole network, which we will get as a solution to the previous subproblem. Therefore, we believe that we will have a more realistic evaluation of researchers.

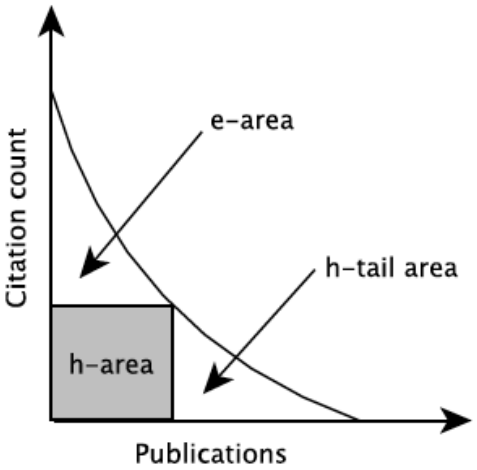
The last subproblem is about evaluating different fields. For this purpose, we will first divide the network into different fields using keywords and fields of the papers. Then, we will evaluate both papers and researchers again, according to these fields that we created. Lastly, we will try to develop a method to compare different fields. In the end, we will try to predict the promising fields based on these evaluations.

Finally, we will develop a web based system to demonstrate our solutions to each of the subproblems above. In this step, our objective will be to visualize the result of our projects in the best way.

1.3 Literature Survey

In the past, some methods have been proposed to evaluate papers and researchers. Today, the most common methods generally use citation counts as the only criteria to evaluate papers and uses different approaches to evaluate researchers based on these evaluations of papers.

Below graph shows the distribution of papers of a researcher with respect to their citation counts:



Hirsch's h-index [6] is the most common and widely used metric to evaluate a researcher in today's world. A researcher has a h-index h if h of his papers have at least h citations and all of the remaining papers of his have less than h citations. However, this index considers only the h-area from the above graph. Therefore, some alternatives of h-index have been proposed.

An alternative for the h-index is the g-index [7]. A researcher has a g-index g if top g papers have at least g^2 citations in total and top $g+1$ papers have less than $(g+1)^2$ citations when papers are sorted in a descending order according to thier citation counts. We can see from the formula that this index uses the e-area from the above graph in a better way.

Another alternative for the h index is the h' index [8]. A researcher has an h'-index $h' = eh/t$ where h is his h-index, and e and t are the square roots of the e-area and h-tail area from the above graph, respectively. We can say that this index uses the above graph better than the h-index and the g-index. One can say that, for the same h, $e > t$ represents a perfectionist, while $t > e$ represents a mass producer, and this index favors a perfectionist more than a mass producer.

Although these 3 metrics to evaluate researchers uses different approaches, they have one thing in common: they use only the citation counts to evaluate papers. However, we believe that although citation counts are a good metric to evaluate papers, there are some better ways to do the same thing. In our project, we will use network analysis techniques such as Random Walk for this purpose.

In graph teory, Random Walk models a random walker who uses the edges in a graph to walk between the vertices. In our citation graph, the edges represent the citations and the vertices represents the papers. It is fair to assume that a random walker on o paper will probably end up in one of the references of that paper. However, since this may not be the case all the time, we add a teleportation step to our model, which makes our new model Random walk with Restarts. In this new model, with a probability d, the damping factor, the random walker follows an edge; and with a probability $1-d$, it teleports to a random vertex. To represent this new model, Random Walk with Restarts, we use the following formula:

$$P(u) = (1-d) + d \sum_{(v,u)\in E} P(v)/N(v), \forall u \in V$$

where $N(v)$ represents the number of outgoing edges from the vertex v, and.

It can be seen that the Random Walk with Restarts algorithm is the same as Google's PageRank algorithm [9].

2 Proposed Solution and Methods

This project consists of 3 subproblems: paper, researcher and field evaluation. Since each subproblem depends on the solution to the previous subproblem, each subproblem will form a different phase. In addition, there will be a data preparation phase in the beginning, and a demonstration phase at the end. In brief, this project consists of the following phases:

1. Data Preparation
2. Paper Evaluation
3. Researcher Evaluation
4. Field Evaluation
5. Visualization

2.1 Data Preperation

This phase is about gathering and preparing the data to be worked on in later phases. We have 2 different datasets for this problem: AMiner dataset and Microsoft Academic Graph. These 2 datasets include 154,771,162 and 166,192,182 papers respectively, and there are also 64,398,608 one-to-one matching between these 2 datasets.

In this phase, firstly, we will construct a graph for each of the datasets. These graphs will have papers as vertices with their years, authors and keywords, and citation relationships as edges.

As the second step, we will merge these 2 graphs with respect to one-to-one relations between them. As a result, we will have one combined graph representing the papers and their years, authors and keywords along with the citation relationships between papers.

Then we will perform some data processing on this combined graph. This will include the construction of the co-authorship network and performing Author Name Disambiguation (AND).

2.2 Paper Evaluation

In this phase, we will evaluate the papers. We will perform this evaluation not based on the citation counts as some other measures but according to the places and importances of the papers in the whole network. In this evaluation process, we will use a Random Walk with Restarts, a Random Walk variant.

2.3 Researcher Evaluation

In this phase, we will evaluate the researchers. We will perform this evaluation according to both the quantity and the quality of the papers of the researchers. In addition to that, we will also take time factor into account to be able to determine the promising or popular authors.

To evaluate researchers, we may use similar methods to the current methods. However, we will use the evaluations from the paper evaluation phase while evaluating the researchers. Since we expect these evaluations to differ from the evaluations with the current methods, we expect that the evaluation of the researchers to also differ.

2.4 Field Evaluation

In this phase, we will evaluate the fields. To do so, firstly, we will create different fields; secondly, we will evaluate the papers and the researchers according to these newly created fields; and thirdly, we will compare the fields based on these evaluations.

First step will be creating different fields. In this step, we will use the keywords and the fields of each of the papers along with the citation network. We will use clustering methods on the keywords and the fields of study, to come up with different fields.

Second step will be evaluating the papers and the researchers again, according to each of these newly created fields. By doing so, we will have different evaluations of the papers and the researchers in each of the fields. For example, a researcher may have different evaluations in different fields, and the goal is to produce such evaluations.

Third step will be comparing different fields. In this step, we will develop a method to be able to compare different fields based on previous evaluations. At the end of this step, we aim to predict the promising fields in the academic world.

2.5 Visualization

This phase is about visualizing the results of our project. To do so, we will develop a query based web application. In addition to developing the web application, we will extract the most interesting findings from our evaluations by trying different queries.

3 Realistic Constraints

This project has various realistic constraints including the following:

1. Economic constraints
2. Manufacturability constraints
3. Sustainability constraints
4. Social constraints

3.1 Economic Constraints

Since this is a data driven project, the resources of this project is the data we have. Therefore, the constraints on the data we have can be considered as economic constraints.

First constraint on the data we have is that although the datasets are huge, they are not complete. This means that there may be papers or researchers that are not included in our data as well as the citations. This may lead to a huge problem, however we believe that the incompleteness in the data is homogeneous and therefore it will not affect the evaluations much.

Second constraint on the data we have is that it is not perfect, and therefore needs preprocessing. For example, we have different names for some researchers in the data, therefore we need to perform Author Name Disambiguation.

3.2 Manufacturability Constraints

The manufacturability constraints are about the processing power we have. We use Gandalf, a server of the school, to keep and process the data. However, although the server is enough to keep and process the data, since the data is huge, each operation takes a lot of time, generally hours. This limits the speed of the project.

3.3 Sustainability Constraint

The data we have is up to June 2017. Therefore, our results will be based on the data up to that date, and the most recent data will not be included in our project. For sustainability, the data needs to be updated regularly.

3.4 Social Constraints

In this project, we will try to evaluate researchers and their works. Although we believe that our evaluation will reflect the data we have in the most objective way, since we try to evaluate the quality, a subjective measure, the result will be still subjective. For that reason, at the end of the project, our results may conflict with current measures or opinions.

In addition to that, since our data consists of real people and their works, some people may be unhappy about the results of this project.

4 Legal / Ethical Consequences

There is no legal consequence for this project. However, there are some ethical consequences related to the social constraints.

The ethical consequence of this project is because of the fact that we will work the data about real people and their works. As explained in the "Social Constraints" section, this project may lead to some results that make some people unhappy. However, we can only assure that, in our project, we will use only the data we have and produce results according to it, and not use any non-data related opinions.

5 Complexity of the Problem

First of all, this problem has no obvious solution. We can conclude this fact by looking at the variety of the proposed solutions. In addition to that, we can also say that, because this problem consists of many different subproblems such as paper evaluation, author evaluation, field evaluation, each again has no obvious solutions.

Secondly, in this project we will work on some networks. Because of that, first, we should know what this network represents and how it is produced, to construct it. Therefore, we need some insight about how the academic world works. Second, we need some research-based knowledge to be able to analyze the constructed networks.

Thirdly, this problem is about a real-life problem as stated in the motivation part, and it has ethical consequences as stated in the previous section. A solution to this problem will result us to understand the academic world much better, therefore it is closedly related to real-life.

And lastly, this problem has a variety of realistic constraints including economic, manufacturability, sustainability and social constraints, as explained in the "Realistic Constraints" section.

6 Risks and Requirements Changes

The risks of the project that will lead some changes in requirements are mostly data related and time related.

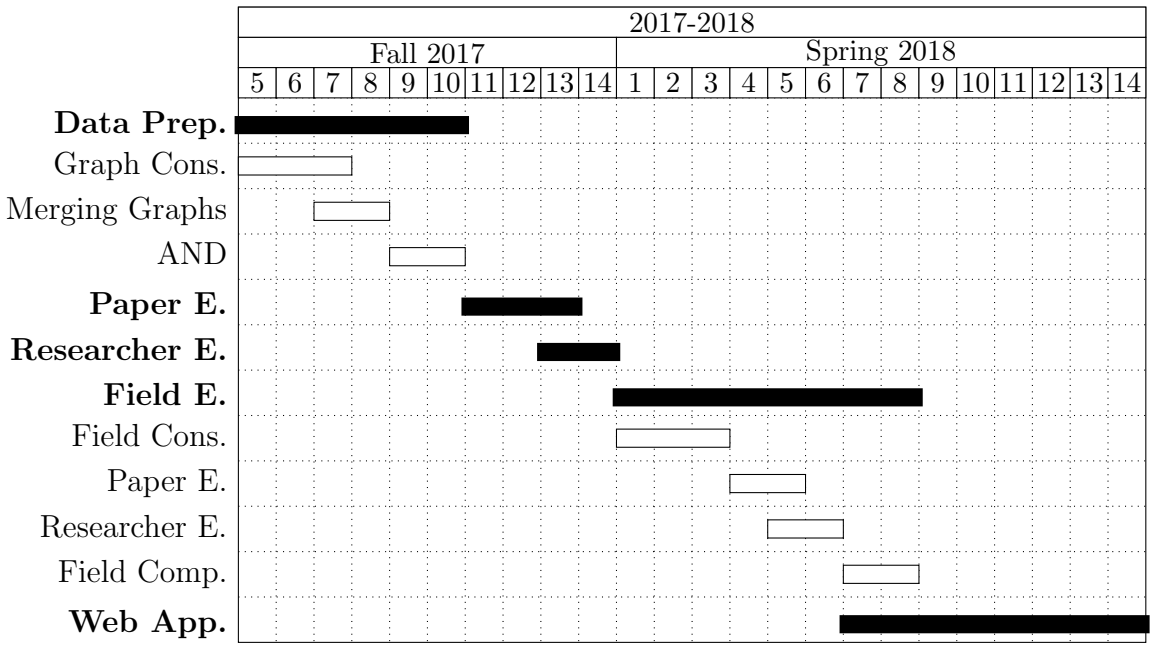
The first risk is about processing the data. We may result in a situation where our processing is not enough to proceed. In that case, we may think about pruning data and getting rid of the "bad" data.

The second risk is abour time management. Since the operations take a lot of time as explained in the "Manufacturability Constraints". Therefore, we may need to revise our project schedule.

The last risk is about the results. We believe that our results will differ from the current approaches. However, this decreases the measurability of the correctness. Therefore, we may need to check the correctness differently.

7 Project Schedule

The project consists of the following phases as explained in the previous section: (1) data preparation, (2) paper evaluation, (3) researcher evaluation, (4) field evaluation, and (5) visualization.



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] Kucuktunc, O., Kaya, K., Saule, E., & Catalyurek, U.V. (2012). *Direction Awareness in Citation Recommendation.* In Proceedings of the 6th International Workshop on Ranking in Databases: p. 6.

[4] Kucuktunc, O., Kaya, K., Saule, E., & Catalyurek, U.V. (2013). *Towards a personalized, scalable, and exploratory academic recommendation service.* In Advances in Social Networks Analysis and Mining: p. 636-641.

[5] Kucuktunc, O., Kaya, K., Saule, E., & Catalyurek, U.V. (2014). *Diversifying Citation Recommendations.* In ACM Transactions on Intelligent Systems and Technology, January 2015. 5(4): p. 55:1-55:21.

[6] Hirsch, J.E. (2005). *An index to quantify an individual's scientific research output.* In Proceedings of the National Academy of Sciences of the United States of America. 102(46): p. 16569-16572.

[7] Egghe, L. (2006). *Theory and practise of the g-index.* In Scientometrics. 69(1): p. 131-152.

[8] Zhang, C. T. (2013). *The h'-Index, Effectively Improving the h-Index Based on the Citation Distribution..* In PLoS ONE. 8(4): p. e59912.

[9] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web.* Technical report, Stanford InfoLab.