CSE 592: Social Networks

Identification of "Advisor – Advisee" Relationship in a Network

Group Details:
Tanvi Sirsat 110098393
Rutuja Sudam Marathe 109967750

ABSTRACT

In every community there exist several kinds of relationships. Similarly, the research community also has an underlying web of connections that depict the relationships among them. In order to understand the network structure, the relationships between the individuals have to be mined. They can be classified as co-author relationships, advisor-advisee relationships etc. In this project, we have focussed on identifying advisor-advisee relations in the research community network. In order to do so, we identified the parameters on which such relationships are dependent. Using these, a score for each researcher is calculated on the basis of which the probable pair of researchers in the relationship are identified. This helps in understanding the present network structure of the community.

INTRODUCTION

With the growth of internet, especially social networking sites, everyone is connected to one another in some well-defined relationship. The mining of these relationships has become a matter of interest in the recent years as they help us understand the structure of the modern network structure. It is well known that an individual is affected by the people he/she is in a relationship with and this affects the dynamics of social networks. If an individual is in a professional relationship with another person it is less likely that the individual will have personal ties with the person and he may not be that person's friends circle. However if the individual has personal ties with another person, they may have several more friends in common and pick on some habits of one another.

With regards to the advisor-advisee relationships, it can be seen that the advisee's research topics are greatly influenced by the advisor. Thus it has become essential to mine these relationships in order to understand the current social network. However it is a task that is easier said than done. The relationships are hidden in the data exposed to us. For example, relationships between family members are hidden in the friendships present on social networking sites like Facebook. Similarly the advisor-advisee relationships are hidden in the DBLP database in the form of co-authors. The motivation behind this project is that with the rapid growth of the research field, people are interested to know how the research community is evolving and what collaborations are playing a role in the shape of the community. The major contributing factor in this analysis is identifying advisor-advisee relationships between individual researchers.

The aim of the project is to assign sociological meaning to ties based on network and attribute information. We have used data from the DBLP dataset and performed analysis on it to identify advisor-advisee relationships in the network obtained from the DBLP database. The ground truth has been obtained from the Mathematics Genealogy Project site. It gives a list of individuals along with their respective advisors and advisees. The structural differences in the network formed between advisors and their respective advisees and the network between co-authors is also essential. It will help clearly demarcate the relationships and by examining the structure we can later predict if the individuals are indeed in an advisor-advisee relationship. We have considered some specific parameters to mine the advisor-advisee relationship from the DBLP database. The parameters considered by us are the year an individual started publishing his work, the number of years the individual has published and the number of publications done by the individual.

METHODOLOGY

The method followed to achieve desired results is as follows:

- Extract data from the DBLP database available on the internet as an input for the project which will contain information about collaboration between researchers on various publications. This data has following main attributes:
    - Name of paper
    - Name of publication
    - List of authors of each publication
    - Date of publication
- The extracted data is parsed in order to extract attributes of interest.
- Two files are created wherein one of them contains all pairs of co-authors and the other contains the attributes considered for each author.
- The score is calculated for each individual according to the parameters considered.
- In our analysis we have considered the following three parameters:
    - Number of publications by an individual.
    - Date when the individual started publishing.
    - Number of collaborations with other individuals(as one publication has more than one author so each collaboration is considered separately).
- According to the score of each person, each publication record retrieved from the DBLP database is analysed and the one with the greater score is deemed as the advisor and the other as an advisee.
- If the difference in the scores is below threshold set by us, the pair is not considered as that of an advisor advisee.
- The pairs are then compared against the ground truth data obtained from the Math Genealogy Project site.It mentions researchers and their advisors and it also contains data of people who don't have any advisor.

In our method, we have given weightage to different parameters after considering the following.
From the analysis of data obtained for each author, it is observed that the year an author started publishing plays an important role in deciding whether the relationship is an "advisor - advisee

relationship" or not. Hence it gets maximum weightage i.e. in our case, '100'. Number of papers published and number of collaborations with other authors are also important. Hence they get equal weightage but less than that of the parameter stating the year an author started publishing, labelled as "Y1". While comparing two authors, if the difference between "Y1" is greater than 10 then we have increased the score of that relationship by 2000. If difference between number of papers published of the two authors is more than 10 then we have increased the score of that relationship by 1000.

Another aspect of our project is to observe the structural differences between individuals in an advisor-advisee relationship and others who are just co-authors. In order to do so, the data was extracted from the DBLP database and was translated into a network graph. A similar graph was constructed from the Math Genealogy database. The structures of the two networks were found to be different.

ALGORITHM USED FOR IDENTIFYING REQUIRED PAIRS OF RESEARCHERS

CALCULATION OF SCORE OF EACH INDIVIDUAL

Weightage:
W(Number of papers published) =10
W(Year an author started publishing) =100
W(Number of collaboration with authors) =10
Current Year=2015

Individual author score = w1 * Number of papers published
+ w2 * (Current Year - Year an author started publishing)
+ w3 * Number of collaborations with other authors

The scores of two individuals in a collaboration are then compared in order to recognise the advisor and advisee in a collaboration.

CALCULATION OF SCORE OF EACH RELATIONSHIP

Difference between starting year of an author (year_diff)
= Starting year of author1 - starting year of author2
Difference between papers published by an author (paper_diff)
= Number of papers published by author1 - Number of papers published by author2
Score of each relationship : score

```
if(10<Math.abs(year_diff))
{
        if(year_diff) score + = 2000;
        else score - = 2000;
}
```

```
if(10<Math.abs(paper_diff))
{
        if(paper_diff) score + = 1000;
        else score - = 1000;
}

if(score of author1 <  score of author2)
{
        score = (score / (score of author1 + score of author2))*100;
}
else
{
        score = (-score / (score of author1 + score of author2))*100;
}
```

IDENTIFYING ADVISOR-ADVISEE RELATIONSHIP

If absolute value of the score of the relationship is in range of 9 to 50 then the pair is probably that of an advisor- advisee. Otherwise they are just authors collaborating on some project.

ASSUMPTIONS MADE

We have assumed that the advisor-advisee relations are dependent only on the parameters considered by us and not other parameters like institute of the individual, field of the individual or number of different publications done by the individual(e.g. IEEE or ACM). Secondly, we have assumed that the difference in the year the advisee started publishing from that of the advisor is greater than or equal to ten. Also, for a pair to be in the advisor-advisee relation,the difference in the number of collaborations between an advisor and advisee require to be greater than or equal to 10. We observed the trends occurring in the DBLP dataset and based our calculations on them.

RESULTS

The outcome is a set of advisor-advisee relationships between different academic researchers. These relationships are then compared with the ground truth data obtained. The ground truth data was obtained from the Math Genealogy project database.
The second part of our project revolves around assigning sociological meaning to the ties in the network graph derived from initial DBLP dataset. The structural differences between the individuals in a relation and those who are not part of any relation are visually represented. Also a graph is constructed from the math genealogy database. The structures of the two graphs are analysed and differences are noted down.

GRAPHS

1. Extract of Network obtained from DBLP dataset

Fig 1:Extract of network obtained from DBLP dataset

For full graph use following link:
https://www.googledrive.com/host/0Bx0l4tCanmstbDI2VVRGZGNiZDg/Chart 1.html

2. Extract of Network obtained from Math Genealogy Project dataset

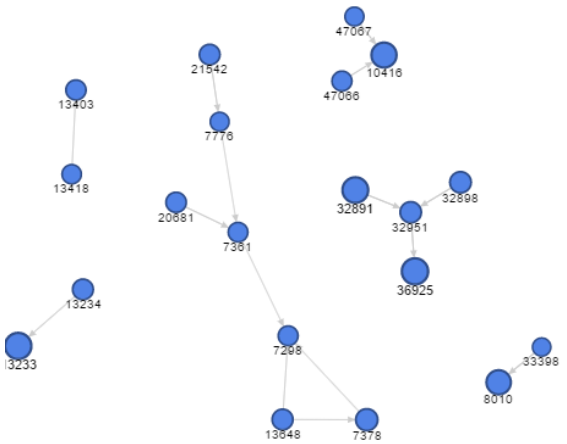

Fig 2:Extract of network obtained from Math Genealogy dataset
For full graph use following link:

https://www.googledrive.com/host/0Bx0l4tCanmstd2xTR29pbzg5eWs/Chart 1.html

Structural Analysis: Test Cases

1) Graphs of C.A.R. Hoare

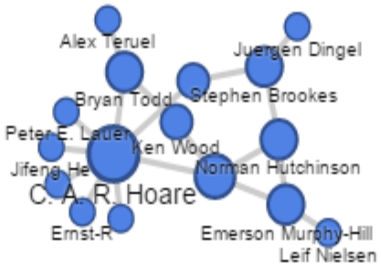

Fig3: DBLP graph

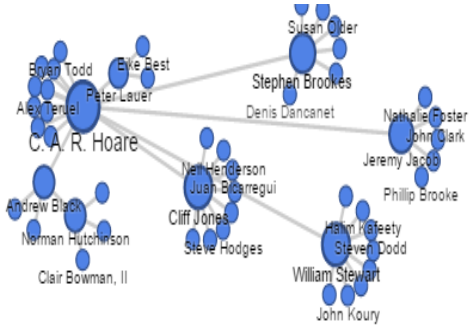

Fig4: Math Genealogy graph

2) Graphs of Wei Zhang

Fig5: DBLP graph

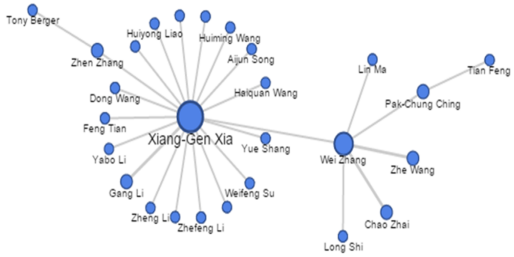

Fig6: math Genealogy graph

RESULT ANALYSIS

1) Analysis from Algorithm

The algorithm was executed on half a million records obtained from the DBLP database. Out of this, only authors of 3000 pairs were found to be present in the Math Genealogy dataset. In order to compare our analysis with the ground truth data obtained from Math Genealogy dataset we took 60 pairs of authors. The data has records of co-authors as well as advisor-advisees.

Percentage of advisor-advisee relations correctly identified
= (advisor-advisee relations correctly identified/total number of advisor-advisee relations)*100
= (21/32)*100
= 65.625%

Percentage of co-author relationships identified as advisor-advisee wrongly
= (advisor-advisee relations identified wrongly/total number of co-author relations)*100
= (13/28)*100
= 46.428%

Percentage of relationships identified correctly
=(advisor-advisee relations correctly identified+co-author relationships identified)/(total number of records considered)*100
= ((21+15)/60)*100
= 60%
We have included this analysis in the file "ResultAnalysis.xlsx"

2) Structural Analysis

From the graphs obtained Fig1 and Fig2, it can be seen that when individuals are in an advisor-advisee relationship in Fig2, the structure of the resulting graph is tree-like where every node has one or more ancestors and respective descendents. Such is not the case in the graph constructed from the DBLP dataset as seen in Fig1. Here the network is more clustered wherein there exist several co-authors of an author. One author may collaborate with several individuals hence, the network is centric around such authors. Since the DBLP dataset also contains hidden advisor-advisee relations, it can be seen that for nodes in such a relation, the structure of those nodes only is tree-like.
We did two case studies in order to analyse the structure of the graphs obtained. Thus, as observed from Fig3 and Fig5, the graph structure is clustered as there are links between co-authors and the interconnections are higher in number. In Fig4 and Fig6 it can be seen that the structure is tree-like where there is a root and its descendents. Thus if this structure can be identified in the graph obtained from the DBLP dataset, the pairs of individuals in advisor-advisee relation can be extracted from the set of co-authors.

CHALLENGES FACED

There were several challenges we faced in this project. The main challenge was obtaining the data as it was not available in a ready format for us. We had to use some specific tools to obtain it. In order to obtain the Math Genealogy database we used SQLite and extracted this data onto Excel sheets as needed for computation purpose. The DBLP dataset was obtained using Java Parser. As the DBLP dataset is huge we could not parse it completely due to resource restrictions. However, we managed to parse five hundred thousand records in order to complete the analysis.
Another challenge was that the records in the DBLP dataset did not match with those of the Math Geneology dataset as the former has data of mainly computer science researchers and the latter has data

about mathematicians.(Thus, we had to look toward external sources to compare ground truth.) The number of common researchers is extremely less and thus ground truth verification could be done only for those.

CONCLUSION

The project focusses on assigning sociological meaning to ties in a network.The hidden advisor-advisee relationships are extracted from the network. The factors taken into consideration are as follows:

- The year a researcher started publishing his work.
- Number of collaborations he is a part of.
- Number of publications done by the researcher.

The maximum weightage is given to the year a researcher started publishing his work as it is found to be the factor on which the relationship is dependent upon the most.
Results obtained from our analysis were 65.62% relationships correct.

The relationships can also be mined from any dataset using the results obtained in the structural analysis. The difference of the structures is identifiable and can be easily recognized in the network obtained from the dataset.

FUTURE SCOPE

As can be seen from the results, the false positives are great in number. In our analysis we have considered only a few parameters and ignored the rest. The project can be extended by considering these parameters as well such as the number of collaborations between two particular individuals, the area of research of the individuals and the university an individual belongs to.

DESCRIPTION OF THE SOURCE CODE

1. dblp.xml:- DBLP database downloaded in xml format from DBLP website that contains all bibliographic records.
2. dblp.dtd:- Document type definition file needed to validate above xml file.
3. parse.java:- : parses xml file and retrives and stores data in txt file.
4. filename.txt:- text file with contains records retrieved using above parser. It contains information of all papers published in different publications. For each paper, name of authors, year of publication, title of paper, name of journal in which it is published and number of issue are stored.
5. ID.java :- Takes filename.txt as input finds out name of all authors and assigns ID number to them.
6. Author_data.java: - For each author, finds out how many papers that author has published.
7. Year_data.java:- For each author, finds out what year that author published his first paper.
8. End_year_data.java:- For each author, finds out what year that author published his last paper and how many years he was active.

9. No_of_partnerships.java:-For each author, finds out with how many author he has collabrated.
10. Coauthor.java:-Find all possible relationships.
11. Score.java:- Assign scores to individual author.
12. Final_score.java :- Calculate score for each pair and determines whether that pair is possible advisor advisee pair or not.
13. ResultAnalysis.xlsx :- Analysis of algorithm applied to data.
14. External API's used:- Apache.poi

REFERENCES

- https://sites.google.com/site/cse592spring14/
- DBLP dataset :http://dblp.uni-trier.de/
- Math Genealogy dataset :http://genealogy.math.ndsu.nodak.edu/search.php
- Extract Math Genealogy database: https://code.google.com/p/math-genealogy-db/
- Extract DBLP database:http://dblp.uni-trier.de/faq/Extracting+data+from+dblp