Beyond Citation Network: Analyzing Publication Data Set as a Multi-Layer Hypergraph

CS224W Reaction Paper

Jingyu Cui
Electrical Engineering
Stanford CA 94305
jycui@stanford.edu

Fan Wang
Electrical Engineering
Stanford CA 94305
fanw@stanford.edu

Jinjian Zhai
Computer Science
Stanford CA 94305
jameszjj@stanford.edu

ABSTRACT

Technical details of 4 papers related to the behavior on the citation network are presented, followed by discussion of their strengths and weakness. Several possible research topics and methods are proposed based on paper reviewing.

Keywords

Citation network, Impact metric, Six-degree of separation, Blog, Infection model.

1. INTRODUCTION

The characteristics of the citation network is of great interest to many researchers in the world. By analyzing the static and dynamic properties of the network, many insights can be obtained regarding research quality, collaboration behavior, even the evolution of science and technology. We focused on 4 papers on topics related to the citation network from various aspects.

In the first paper, *Planetary-Scale Views on an Instant-Messaging Network* [3], the authors studied one month of communication activities on the MSN instant message network all over the world in a period of 30 days. The paper emphasized on large numbers of people (240 million) and large dataset (4.5 TB text logs), forming an undirected communication graph with 180 million nodes and 1.3 billion edges. Each user was represented by a node and an edge was placed between users if they exchanged at least one message during the period of observation. A few characteristics were analyzed, such as the average path length among Messenger users (which is 6.6 and consistent with six degrees of separation), and the cluster coefficient. The authors also addressed "homophily", which is the tendency of communication between people of the same background, such as geography, age, gender, etc. These methodologies to characterize a network are helpful for our research topic.

In the second paper, *Citing for High Impact* [6], the authors used a database of publications and characterized the citation relationship between them. They investigated how patterns of citations varied between the scientific disciplines and how such patterns related to the impact of a paper. First, a citation projection graph was defined for each publication v_0, with nodes being all the papers that v_0 cites, connected according to their citation relationships. Several metrics and statistics were defined for this citation projection graph, such as density, clustering coefficient, connectivity, maximum betweenness, betweenness, and network constraint of v_0. These metrics were evaluated for each citation projection graph, and a normalized distribution of the metric values was obtained for each discipline for further analysis. Then a random graph was generated as a counterpart of each real citation projection graph, which had the same in- and out-degree sequences as the real one. By comparing these two graphs, it was concluded that the real graph was more clustered than the random one. Next, Impact of publications was defined, and how the graph metrics were changing with impact was also studied for different areas. Then all publications were classified into "high", "mid", and "low", according to their impact levels, and more detailed analysis was provided by using statistical hypothesis testing. Finally, how citation graph changed over time was analyzed by dividing the publications into "old" and "recent" groups. Those pre-defined metrics were evaluated for the citation projection graph corresponding to the two groups of publications. It was revealed that recent publications tend to have broader and more diverse citations.

In the third paper, *Cascading Behavior in Large Blog Graphs* [5], the author analyzed a data set containing 45,000 Blogs and 2.2 million posts, and provided several significant findings: Number of Blog posts was not bursty, but followed a weekly period with weekend drop-off; Popularity of posts dropped according to power law with parameter 1.5; a lot of properties of the graph followed power law, such as size of cascades, in and out degree distribution of Blog network, in and out degree distribution of post network, posts per Blog, etc.; Most popular cascade shape was the star shape; the cascade grew like a tree. In addition to the findings, the authors also proposed a susceptible-infected model which gave simulation results that were quite close to the real data.

In the fourth paper, *Cost-effective Outbreak Detection in Networks* [4], the authors tried to find the most important nodes to obtain information from the network. Since many important objective functions are submodular, greedy al-

gorithms such as CEF can obtain solution close enough to the optimal. The authors further proposed the cost-effective lazy forward selection algorithm (CELF) to further speed up the CEF by more than 700 folds.

In general, the four papers covered important statistical properties and methodologies to characterize a network from several different aspects. The importance of accurate modeling of the network was especially emphasized in the two papers [5] and [4].

2. DISCUSSION OF RELATED WORK

2.1 Relationship with course content

The contents of these papers have a lot of overlap with the content in our course. The small world phenomena and six degrees of separation of real world social network was discussed in [3]. Many metrics and statistics for analyzing graph from [6], such as the graph density, clustering coefficients, connectivity, etc., have been covered in our course as well. The random evolution of network [6] was also discussed in our course. The other two papers [5, 4] mainly talked about the cascade behavior in the network, which will be covered in our course later this quarter. For example, Leskovec et. al. [5] built a model for network cascades, and the nodes which had the maximum influence on the network were found.

All in all, these papers we've selected are highly related to our course content, and the techniques involved in them cover the topics from basic statistics to the characteristic behavior of a network. With the help of these techniques, we can make better analysis on the citation network which we intend to investigate.

2.2 Weakness Discussion

The papers we discussed are of great technical value. However, there are several general aspects that we think could be improved, which are stated as follows:

First, we believe that it would be meaningful to investigate the network during different time periods. For example, in [3], the authors investigated a data set containing one month of communication activities in the MSN instant message network, and made the conclusion of six degree of separation. We are also curious about properties of the network in a longer period, such as one year. It would also be interesting if we could investigate the evolution of the network over a longer time period. Does the homophily vary in different seasons of the year? Does it behave differently in weekdays and weekends? As we usually add more friends and seldom remove old friends, the network should become denser over time, so is the degree of separation still six in 2010, when it is already four years after the dataset in [3] was originally collected?

Second, both global and local structures of a network are interesting to us. For example, the citation projection graph in [6] was only a part of the whole citation graph, i.e., only the local structure of the whole citation graph was investigated in [6], while the analysis of the global structure was missing somewhat. We think the global citation graph in different scientific disciplines can be also analyzed using the similar metrics, and it's possible to get some new valuable conclusions.

Third, the cascades behavior in Blog graphs will generate a tree structure of the network in [5] and [4], i.e., the nodes at the same level will not have links between them. We think this might be a limitation of the analysis. This cascade structure will prevent a node citing other nodes which are on the same level as itself. We think this might be a result of the discrete time delay. When the time is discretized, some posts that are actually posted on different time points will be put onto the same time slot, and the links between them might be ignored. To solve this problem and make the network more reasonable, we can try to make the time delay a continuous variable.

There are also several possible improvement that can be made to the specific papers:

In [6], the impact of each paper might be defined more precisely. It was stated that the number of citation each publication received would vary according to the publication year. We think this needs to be verified by investigating the average citation number per paper changing over time, and maybe a model can be built to capture the general characteristics of the citation pattern.

The publications are classified into groups of "old" and "recent" in [6] to investigate the citation pattern changing over time. It might be more interesting to investigate the citation pattern changing year by year, that is, the division of publication time can be more subtle. Although it is generally believed that the citation is becoming more diverse in recent years, it might be less diverse in some of the recent years. It should be useful to show whether the diversity is increasing, or oscillating, over years.

In [5], how the information propagated over the network was simulated by a simple but powerful model called Susceptible-Infected-Susceptible model, in which each node spread its information to its neighborhood nodes with probability β. The choice of β is very critical here. However, it is somewhat limiting to have one value of β to describe the behavior of all nodes. When a Blogger is referring to other Blogs for information, they are not making decisions based on a fixed parameter. Many important factors will have influence in this process, such as whether or not the source Blog is famous, how much the Blogger trusts the source, etc.

Besides, when we want to transfer the idea of cascade analysis onto the citation network, we should be aware that each researcher is not citing papers randomly either, even when all the related publications are available. They might have biases or preferences about the authors, institutions, overall journal/conference quality, etc. It should also be noted that sometimes the authors are more willing to cite publications of their own.

3. FUTURE DIRECTIONS

In order to extend the work in the papers mentioned in Section 1, and analyze the citation network in a more accurate and complete manner, we propose to analyze the citation network from the following different aspects:

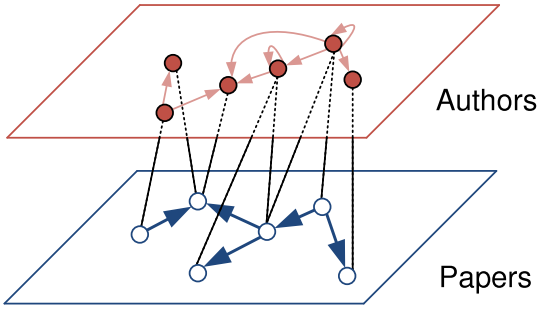

Figure 1: Example of a two layer network involving the paper layer and the author layer. Note the authorship edges (black), paper citation edges (blue in the papers layer), and the induced author citation edges (pink in the authors layer).

3.1 Multi-Layer Hypergraph

The citation network can be more accurately and completely modeled as a multi-layer graph which consists of entities in various levels (Figure 1), from top to bottom, being research fields, research groups, authors, and papers. Each layer itself is an interconnected network defined by reference relationship, but nodes between consecutive layers can also be connected to each other. For example, an author can write multiple papers, and a paper typically has multiple authors. The edges in a higher layer are induced by edges in the layer right below it, together with the edges connecting the two layers.

When grouping the properties of the graph, we can do finer grain quantization instead of grouping into two or three discrete groups (recent and old, high impact and low impact, etc.). Moreover, a continuous function can be fit onto the discrete data to gain more insight in the trend of the data, as well as to do plausible extrapolations.

In addition, we can model the time varying properties of the graph. More specifically, we can look at the evolution of the citation for a specific paper over time, or look at how new scientific discoveries (e. g., SVM, Compressive Sensing, FFT, etc.) propagate in the academic network, especially across different fields. Or inversely, new trends in the citation network can be discovered by mining the information flow in the network. We also expect to find that many properties of the graph are strong functions of time, whereas others are not.

Moreover, we can try to answer time evolution related questions such as: 1. As papers are more and more available in digital forms, does the citation graph change? We expect that the popularity of digital papers makes it easier for researchers to access papers from a different field and make citations. 2. Does six degree separation still hold in citation networks in all layers? How does the separation change over time as digital papers become more and more available? How far away are two different research fields? As research becomes more and more specialized, is it becoming easier or harder to collaborate with someone from another field? 3. Does the pattern of publication change over time as one evolves in his academic career? We are interested in looking at change of publication number, impact, author order, etc., across the lifetime of researchers.

To create a model that captures the citation behavior between the papers, similar to the way Blog network is modeled in [5], we can use the susceptible-infected mechanism, but give papers more flexibility in the probability of being cited, for example according to some initial estimation of impact based on other factors. Then we can see if the model fits well with the real data. This model might still fail since other important factors that affects citation behavior are not modeled, such as tendency of self-citation, citation bias towards a certain subset of papers, etc.

3.2 Domain Specific Subgraph Analysis

Citation graph properties might vary across different domains. For example, citation pattern for high impact papers in computer science might be different from that in physics. Putting all papers together obscures this difference. We could analyze domain specific subgraphs to reveal this variation.

Similar to finding the important nodes in the Blog network, we can find key publications for specific fields, and automatically generate a reading list for people who want to do a quick survey of a research direction.

3.3 Impact Metrics

We are also interested in how to integrate different metrics of the network to get an effective metric to measure impact of an entity, e.g., author impact, paper impact, research group impact, etc. There are existing works to measure author impact, such as citation number normalized by average citation in the same field in [6], h-index [1], g-index [2], etc. Some of the metrics can be generalized from measuring author impact to measuring paper impact, or research group impact.

An interesting topic to look at is the graph composed of papers published by a specific author or research group. By analyzing self-citation patterns and its correlation with other commonly known impact metrics, we could gain interesting insights on how researchers differ in self-citation behaviors.

Algorithms similar to page rank can be also applied to different layers of the hypergraph to determine the impact of a paper, an author, or even a research group.

4. REFERENCES

[1] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69:131–152, 2006.

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

[3] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.

[4] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak

detection in networks. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429, New York, NY, USA, 2007. ACM.

[5] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *In SDM*, 2007.

[6] X. Shi, J. Leskovec, and D. A. McFarland. Citing for high impact. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 49–58, New York, NY, USA, 2010. ACM.