VOL. 9, NO. 8, AUGUST 2014 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

ON UNDERSTANDING CENTRALITY IN DIRECTED CITATION GRAPH

Ismael A. Jannoud and Mohammad Z. Masoud
Al-Zaytoonah University of Jordan, Amman, Jordan
E-Mail: ismael.jannoud@zuj.edu.jo

ABSTRACT

Modeling complex networks as directed/undirected graphs is considered one of the most common methods in network science. Citation graph is a directed graph of scientific published papers. This graph has been studied massively in the past decade. Citation graph can be utilized to study relationships between authors and papers. It can be used to study the characteristics of citation network to demonstrate the growth model, graph type and to predicted hot new topics. In this paper, we attempt to study the relationship between popularity of a paper and the publication date. The purpose of this study is to demonstrate the relation between paper quality and hot topics. Betweenness metric has been used to measure the popularity of a published paper. Moreover, a comparison between betweenness and citation count (node degree) has been conducted to show that papers may have a small citation count, however, they may have a great impact in research field. We have generated a directed citation graph by crawling paper information from Citeseerx. Our study shows that date of publication is important to write a popular paper. However, high quality papers get opportunity to be popular regardless the date of publication.

Keywords: node degree, citation graph, graph metrics, cluster coefficient, betweenness.

INTRODUCTION

Citation is the process of quotation information, results and conclusions from book, paper, or websites in a research work. Scientific research papers (SRP) contain from few references to few hundreds. Citation is one of the metrics that used to measure the popularity and the importance of SRP. Moreover, citation may be used to find the roots of science fields. Citation process generates a complex directed graph, which is called citation graph or citation network. In this graph, nodes (papers) are connected with directed edges that begin from newly published papers and point to old papers.

Citation in these networks are the edges, papers are the nodes. This definition converted these networks into complex directed graphs. The emerged of network science, its applications and tools increased the possibilities in studying complex networks. Citation network is one of the complex networks that have been studied heavily in the past decade. Paper popularity, author ranking, author relations and community studies are some examples of the vast range of studies that have been conducted in this field.

Mapping graph parameters and properties into physical meanings that reflects useful meaning is the niche of network science. This mapping helped in predicting special phenomenon of citation networks, such as, power-law distribution, graph type and the expanding models of these networks. However, many graph parameters have not been studied and mapped in citation networks, such as, betweenness and closeness.

In this work, paper importance will be measured. Paper centrality in the citation network will be extracted. Timing of papers and their centrality will be examined. The purpose of this work is to demonstrate the important of hot topics time and the position of papers in citation networks. Graph betweenness will be utilized to measure the centrality of papers. To conduct our experiment massive amount of papers in a certain field must be harvested. To facilitate our experiment, a web crawler has been implemented to collect paper information from citeseerx website. To this end, we have generated a direct graph from the harvested data. We seek to answer the following questions using graph analysis:

- What is the meaning of centrality in citation network?
- What the relationship is between hot topics timing and paper centrality?
- Can researchers write a paper with high citation value without considering the hot topic?

The rest of this paper is organized as follows; this section ends with the related works that have been conducted in this area. Section 3 introduces the performance metrics. Section 4 provides a description of the experiment that has been implemented. This section ends with the results and discussion. Finally, section 6 concludes this work.

RELATED WORK

Studying real network as graphs inspired researchers over the years. They heavily studied social, information, technological, and biological graphs [1]. These studies gain researchers more insights of how these networks may evolve and how bugs, errors and diseases may separate.

Many types and kinds of networks have been studied as graphs. For example, in [2, 3] it has been reported that the World Wide Web (WWW) follows an exponential degree distribution with more than 269 thousands nodes and around 1 and half million edges. This study provided that the WWW is a small world graph. In the measurement works [4], authors attempted to generate Internet graph to study its properties. They have produced an undirected graph with 10 thousands nodes and 31 thousands edges. A 0.035 global cluster coefficient value has been computed. These values have been computed

over the years again to show the development of the Internet [5].

A third type of networks have been studied in [6], the author of this study generated a software-classes directed-graph. With 1377 nodes and 2213 edges, the author studied the properties of this small graph. The author reported a mean node-node distance with 1.5 hops. In addition, global clustering coefficient and degree correlation coefficient were computed. This study showed that software classes don't follow the small world phenomenon. Furthermore, in the works of [7, 8], authors generated a small undirected graph based on harvested data to simulate P2P network. A graph of 880 nodes and 1296 edges has been implemented and studied. The author found that the average shortest path in this graph is 4.2 hops and the global clustering coefficient is 0.012. Unfortunately, their implemented graph was too small to mimic a swarm of P2P networks. In [9], authors attempted to study human's neural networks as a directed graph. They have constructed a directed graph with 307 neural and 2359 edges between them. Subsequently, they studied the properties of this graph. They reported a global clustering coefficient of 0.18 and an average shortest path value of 3.97. These values demonstrated that human's neural network is a small world graph.

Finally, Citation networks have not been forgotten by researchers. In [5], the author studied the citation network as a directed-graph. The author computed node number and vertexes. However, the cluster coefficient and degree correlation were not computed. These results made it hard to predict node distribution in citation network. In [10] a citation graph has been generated to study the impact of co-authors on the popularity of papers. In addition, co-authors are connected if they have shard work. Our work in the citation network differs from these works in two points. First, we utilized a new graph parameter to study the centrality and popularity of a paper in a certain field. Second, time has been utilized to demonstrate the relation between time and popularity.

PERFORMANCE METRIC

This section introduces the metrics that have been utilized from graph theory. Graph centrality through betweeenness and node degree will be utilized. Graph centrality is a metric to evaluate the importance of any node in a graph. Many methods are utilized to measure the centrality of a node. Node degree, eigenvector, closeness and betweenness are the most common methods to calculate the centrality of nodes. In citation graph, centrality metric may be used to extract the core nodes of the graph. These core nodes may be introduced as the root of a scientific field. Any research in a scientific area starts from an idea in the past. However, there is a time when this field emerges as a hot topic. Core nodes are the nodes that convert a topic into a hot topic.

In this work, we have utilized the betweenness and node degree of the nodes in the citation graph to measure graph centrality. In the next section node degree and betweenness will be defined.

Betweenness and Node Degree Centrality

Betweenness of a node is defined as total number of paths that flow through this node [11]. Equation (1) shows the mathematical method of calculating betweeness of a node. In a directed graph such as citation graph, betweenness may be used to demonstrate the importance of a scientific paper.

$$G(v) = \sum Q_{si}(v) / Q_{si} \quad (1)$$

Node degree is defined as the total number of incoming and outgoing connections that connects a node with other nodes.

Node degree may be used as centrality metric. However, some nodes may be important and they may have less degree than other nodes. This fact is emerged in betweenness centrality more than node degree centrality. We believe that betweenness centrality results are more accurate for two reasons. First, in scientific research timing is an important factor. The important of a paper depends on the time. Old popular paper will not get more citation after a long time. Rich will be richer if they are not too old unlike other graphs, such as, the AS graph. Second, an important paper may be cited by little number of papers that may become more and more popular. These two facts can be found crystal clear in the betweenness centrality. Figure-1 shows an example of the differences between node degree and node betweenness. Nodes 1-7 have node degree more than node 8. However, node 8 is more important since it connects between this node and nodes: 9, 10.

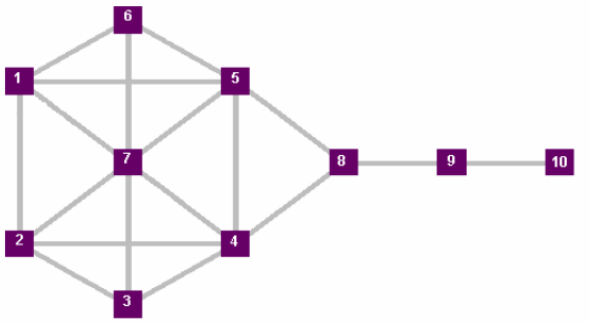

Figure-1. Example of the differences between Node degree and Node Betweenness Centrality.

THE EXPERIMENT

A web-crawler has been implemented to harvest scientific paper information from CiteSeerx website. CiteSeerx is a scientific literature digital library. It focus primary on scientific research papers in the fields of computer and information science. It has many functionalities and methods that allow it to index postscript and PDF articles.

Our web-crawler starts by retrieving the search results of a specific topic. We have used ad hoc as our main topic to generate our graph. The maximum allowed

retrieved results are 500 papers. Subsequently, the crawler uses the retrieved 500 paper names, which are the seed list of the crawler, to retrieve their cited paper. Each cited retrieved paper is added into the list to retrieve its citations and so on. The harvesting process started in 15th of January and continued for 10 days. The crawling process stopped when encountered papers that have no information in the CiteSeerx web-site.

The harvested data of the conducted experiment has been utilized to generate a directed citation graph. Figure-2 shows an example of the generated direct citation graph.

Table-1 lists the properties of this graph. We can notes from the graph that the clustering coefficient of this graph is too small comparing with other graphs, such as, the AS graph. In addition, the average path length is longer since it's a directed graph.

Table-1. The Constructed Citation Graphs Properties.

Property	Citation Graph
Number of Nodes	91211
Number of Links	221677
Average Node Degree	2.43
Cluster Coefficient	0.067
Average Path Length	7.4
Graph Diameter	24

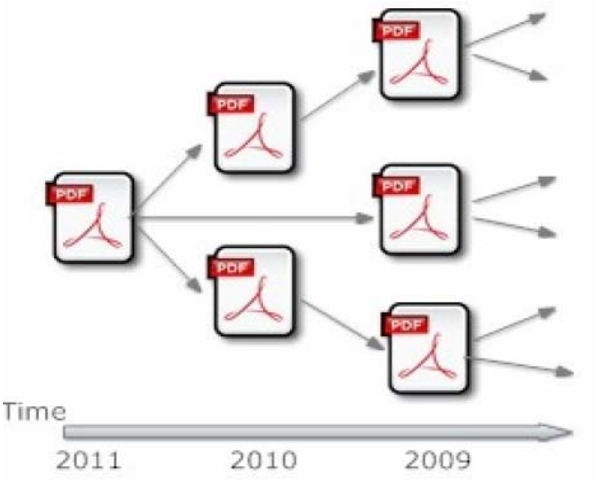

Figure-2. Example of the Generated Direct Citation graph.

RESULTS

Figure-3 shows the CDF of node betweenness value. From this figure we can observe that more than 88% of the nodes have a betweenness value less than 1. In addition we can observe that less than 4% of the nodes have a betweeness value less than 100. More over we can observe that a tiny number of nodes (0.02%) have a vast betweeness value. These nodes are the central of our citation graph.

Figure-4 shows a relation between publication year and the betweenness value. To compute this figure, 10K papers have been excluded from the result. These papers have no publication year information in CiteSeerx web-site. The rest of the papers (more than 8K) have been used to generate this relation. We can observe that the figure is a random figure. We notice that for each year of publication, the betweenness value varies. In addition, we observed that the highest betweenness value was in 2005. However, the papers in ad-hoc started before this year. This figure proves that high quality papers we be in the core of any field regardless of the year of publication.

Figure-5 shows the relation between betweenness value and node degree or citation count. This figure shows that the relation between betweenness and citation count is fuzzy. With the increased number of citation count, betweenness value increases. However, some node with high citation value has small betweenness. In addition, we can observe that the node with highest betweenness value is not the node with the highest citation count. The figure has a fuzzy slop to the right. This figure shows the differences between betweenness centrality and node degree centrality. This figure demonstrates that some nodes with less citation count may be more important than others. Betweenness may be used to show this fact.

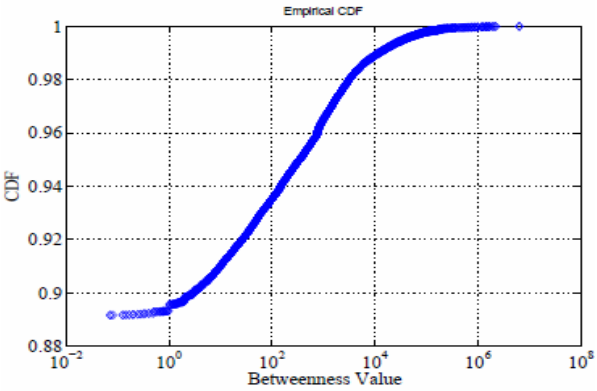

Figure-3. The CDF of Node Betweenness Value.

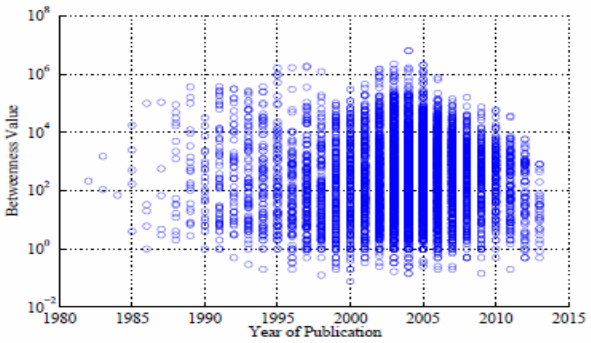

Figure-4. Relation between Publication Year and the Betweenness Value.

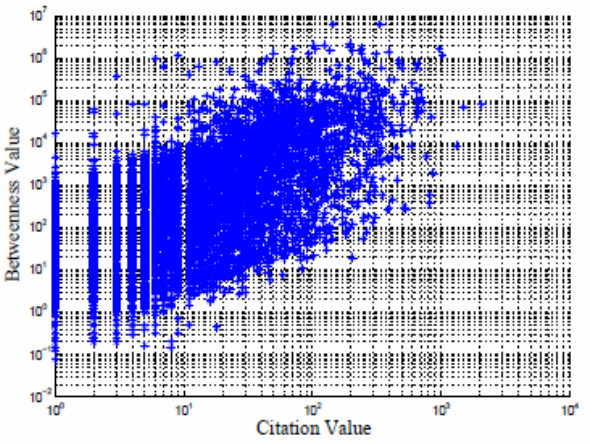

Figure-5. The relationship between Betweenness Value and Node Degree.

Finally, Figure-6 shows the CDF of node citation count (node degree). We can observe the similarity between the CDF of betweenness value in figure and this figure. However, unlike the betweenness CDF that has less than 10% of the nodes with significant betweenness value, we can observe that more than 33% of the nodes have significant node degree (more than 10). This fact demonstrate how betweenness centrality can be used to filter a massive number of nodes to obtain the most central and important once in a graph.

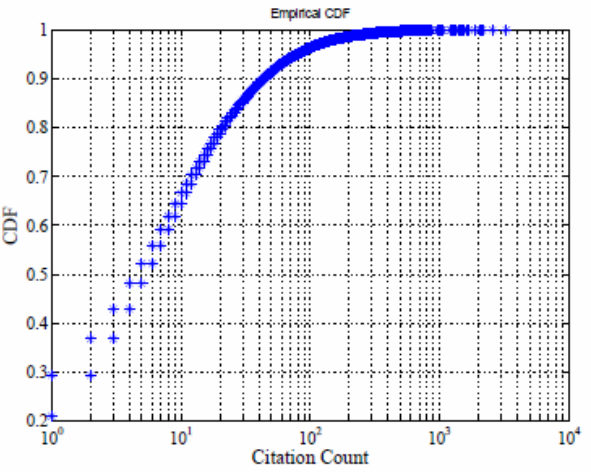

Figure-6. The CDF of Node Citation Count.

CONCLUSIONS

Researches utilized the citation graph to study and introduce the publication networks for various reasons. The last decade witnessed a heavy study of the properties of this graph. In this work, we have studied the impact of the publication date on the popularity of published papers. Three main graph properties have been utilized to study the relationship between popularity and hot topic (date of publication). Betweenness, global and local cluster coefficients and node degree distributions has been computed. To this end we have generated a citation graph by crawling paper information of ad hoc (research field) from Citeseerx web-site. Our results demonstrated three main points. First, the probability of getting high citation count increases with the time or the date of publication. Second, there is a probability that a high quality paper will get a good citation count even if it is late in the field. Finally, the data shows a fuzzy slope in the graphs. This uncertainty requires more experiments.

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks," SIAM REVIEW, vol. 45, pp. 167 256, 2003.

[2] A.-L. Barabasi, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the world-wide web," PHYSICA A, 2000.

[3] R. Albert, H. Jeong, and A.-L. Barabasi, "Diameter of the world-wide web," Nature, pp. 130-131, 1999.

[4] S. N. Dorogovtsev and J. F. F. Mendes, "Language as an evolving word web," Proc. Roy.Soc. London Ser. B, 268, pp. 2603-2606, 1999.

[5] M. Masoud, X. Hei, and W. Cheng, "A graph-theoretic study of the flattening internet as topology," In: IEEE ICON, December 2013.

[6] S. Valverde, R. F. Cancho, and R. V. Sole, "Scale-free networks from optimal design," In: Europhys Lett., 2002.

[7] M. Ripeanu, I. Foster, and A. Iamnitchi, "Mapping the gnutella network: Properties of large-scale peer to-peer systems and implications for system design," In: IEEE Internet Comput., 2002.

[8] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman, "Search in power-law networks," in Phys. Rev. E,, 2001.

[9] J. G. White, J. N. T. E. Southgate, and S. Brenner, "The structure of the nervous system of the nematode," In: C. Elegans, Philos. Trans. Roy. Soc. London, 1986.

[10] K. Healy, "A co-citation network for philosophy," June 2013.

[11] T. G.Lewis, "Network science theory and practice," In: Wiley.