The Analysis of A Citation Network Relating to Grid Computing

Xi He
Rochester Institute of Technology
Bldg 74, Lomb Memorial Drive, Rochester, NY 14623-5608
hexi111@gmail.com

***Abstract*—In this study, we analyze a Grid Computing related citation network and identify the appropriate literatures for researchers interested in Grid Computing. We first examine some network characteristics which are effectively used in our network analysis. Then we develop our methodology for network analysis, including modeling methods and data collection process. The experiments are conducted to validate our ideas and detailed interpretations are discussed. Our study shows that it is effective and efficient to find useful literatures for researchers by means of network analysis.**

***Index Terms*—Grid Computing, Citation Network, Average Path Length, Cluster Coefficient**

I. Introduction

In general, the term "network" means the interconnected system of people or things. A network is composed of a set of connections formed between the individuals in the system to achieve certain goals. For example, Tom has a lot of friends in the Facebook. If we model each of his friends as a node, and use an edge to connect every pair of his friends if they know each other, a network describing Tom's friends' relationship is formed (Figure 1). Some popular networks that are of interest to studies and researches include Computer Network, Communication Network, Financial Network, Social Network and so on. Networks exist almost everywhere in our world, and have a significant influence on every aspect of our lives.

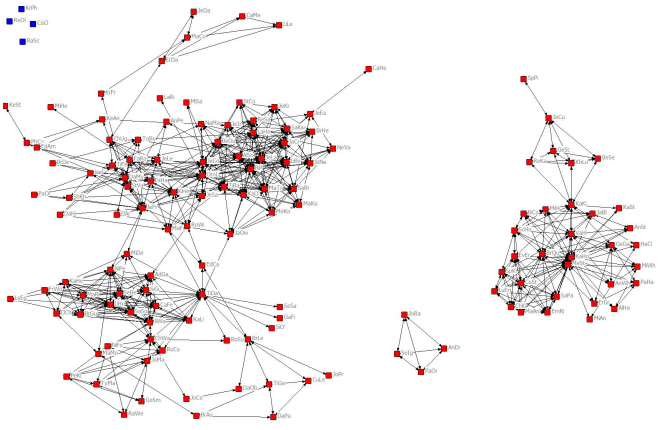

Fig. 1. A Facebook Network

In 1736, Swiss mathematician Leonhard Euler presented his solution to Seven Bridges of Königsberg problem in his paper *Seven Bridges of Königsberg* [1]. This event is widely regarded as the beginning of graph theory. After that, graph theory boomed with the contribution from mathematical giants such as Cauchy, Hamilton, Cayley and Kirchhoff. Two century after Leonhard Euler started graph theory, Paul Erdős and Alfréd Rényi introduced random network theory in their paper *On Random Graphs* [2]. Unlike graph theory which aims to discover and catalogue the properties of the various graphs, random network theory tries to answer such questions as how real networks form and what are the laws governing their appearance and structure. Paul Erdős and Alfréd Rényi viewed networks and the world they represented as fundamentally random. They believe that a network is obtained by starting with a set of n vertices and adding edges between them at random. In 1998, Duncan Watts and Steven Strogatz identified small world networks as a class of random networks [3]. Purely random graphs, built according to Paul Erdős and Alfréd Rényi's theory, exhibit a small average shortest path length along with a small clustering coefficient. Duncan Watts and Steven Strogatz measured that many real-world networks have a small average shortest path length, but also a clustering coefficient significantly higher than expected by random chance. They then proposed a novel network model, currently named *Watts and Strogatz model*. A year later, Albert-László Barabási and his colleagues found that some Web nodes, which they called "hubs", had many more connections than others and that the network as a whole had a power-law distribution of the number of links connecting to a node [4]. Albert-László Barabási and collaborators coined the term "scale-free network" to describe the class of networks that exhibit a power-law degree distribution.

Today network theory [5] is an area of computer science, network science and part of graph theory and applied in many disciplines including particle physics, computer science, biology, economics and sociology. Its topics include

- Network theorems: Max flow min cut theorem; Menger's theorem; Metcalfe's law.
- Network properties: Betweenness; Centrality; Closeness.
- Network theory application.
- Networks with certain properties: Complex network, Scale free network, small world network.

This study focuses on analyzing the Grid Computing related citation network with the goal of identifying the most important literatures in Grid Computing. In Section II, some network characteristics adopted in our study are thoroughly

examined, followed by an introduction to the basic concept of Grid Computing and the motivation for this study in the Section III. Then we introduce the methodology for this study in Section IV. A discussion of the network modeling method and the network simulations adopted in this study is presented. In the next section, we show the result of the experiments and discuss their implication. Later the significance and limitations of the study is described in Section VI, followed by Section VII which is the conclusion.

II. Background

The section aims to provide the necessary background for this study. We introduce some network characteristics such as average path length and clustering coefficient, which are of importance to our study.

A. Average Path Length

Average path length is defined as the average number of steps along the shortest paths for all possible pairs of network nodes [6].

$$L = \frac{1}{\frac{1}{2}N(N+1)} \sum_{i>=j} d_{ik} \tag{1}$$

Where N is the number of nodes, d_{ik} is the distance between two nodes.

The average path length distinguishes an easily negotiable network from one which is complicated and inefficient, with a shorter average path length being more desirable. The famous six degrees of separation theory [7] says that if a person is one step away from each person they know and two steps away from each person who is known by one of the people they know, then everyone is at most six steps away from any other person on Earth(See Figure 2). In fact, six degrees of separation phenomenon does not limit to human network. For example, Albert-László Barabási and his colleagues predicted that the diameter of the Web is 18.59, close to 19. That is, a web page is on average only 19 clicks away from other web pages.

B. Clustering coefficient

The clustering coefficient of a vertex in a graph quantifies how close its neighbors are to being a clique [8].

$$C_i = \frac{2E_i}{k_i(k_i-1)} \tag{2}$$

Where E_i is the number of edges among the neighbors of node i and k_i is the node i's degree.

Let us take an example to explain the concept of clustering coefficient. Jack has four friends. If his friends are all friends with each other as well, each of them can be connected with a link. But chances are that some of his friends are not friends with each other. Let us say there are only three links between his friends, few than six links. So the clustering coefficient for Jack's friend cycle is 0.5. The clustering coefficient tells about how closely the cycle of Tom's friends is. If the clustering coefficient is 1, it means that all of his friends are good friends with each other. On the other hand, if the clustering coefficient is zero, Tom would the only person who hold his friends together.

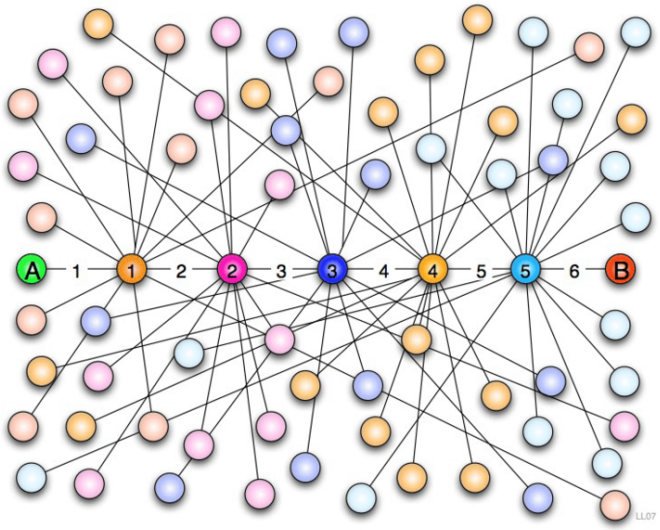

Fig. 2. Six degrees of separation [7]

C. Centrality measurement

Centrality of a vertex within a graph determines the relative importance of a vertex within the graph [9]. For example, how important a person is within a social network, or how important a room is within a building. Four kinds of centrality measurement is widely used in network analysis: degree centrality, betweenness, closeness and eigenvector centrality.

1) Degree centrality: Degree centrality is defined as the number of links incident upon a node. Degree is often interpreted in terms of the immediate risk of node for catching whatever is flowing through the network. If the network is directed, then we usually define two separate measures of degree centrality, namely indegree and outdegree. Indegree is a count of the number of ties directed to the node, and outdegree is the number of ties that the node directs to others.

2) Betweenness centrality: Betweenness is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have higher betweenness than those that do not.

3) Closeness centrality: Closeness centrality represents how far from all other vertices in the network. Closeness centrality is based on the concept of network paths. Vertices that tend to have short geodesic distances to other vertices with in the graph have higher closeness.

4) Eigenvector centrality: Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

III. Motivation

Grid Computing (see Figure 3) is a form of distributed computing paradigm originated in the early 1990's in Ian Foster's and Carl Kesselman's seminal work: "The Grid: Blueprint for a new computing infrastructure". Ever since then, Grid Computing has been regarded as one of the hottest research fields in the world. The basic idea of Grid Computing is the combination of computer resources from multiple administrative domains, and applied the powerful computation ability to the scientific, technical or business problems that requires a great number of computer processing cycles or the need to process large amounts of data.

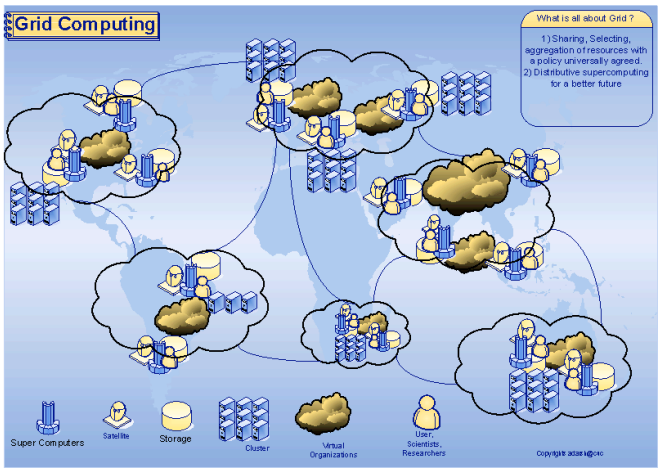

Fig. 3. Grid Computing

Through a decade's development, Grid Computing is mature and well developed. There exist thousands of literatures which talk about Grid Computing. It is not easy to select the right Grid Computing related literatures from millions academic literatures across various disciplines. Some papers with the title of "Grid" might actually talk about the power grid in the electricity industry. What is more, not all of the literatures are worthy of reading. Some of literatures are good and useful. Unfortunately, much more literatures might not be as good as expected. Due to the enormous number of the literatures, it is also not realistic to read through all of the literatures in Grid Computing and then identify which ones are good or which ones are bad. As a result, how to find the worthwhile literatures becomes a critical problem posed in front of the novices with little experience in Grid Computing.

The purpose of this study is to provide a way or method for students or researchers interested in Grid Computing to locate the classic literatures in a relatively shorter period. With an exhaustive research and study on the Grid Computing related citation network, we can study every paper's role and impact in the network of academic papers, thus obtaining the reasonable clues determining how to select the appropriate papers. Actually, the study of citation networks has been conducted for a while. In [10], the author analyze motif of Journal citation networks, citation pattern and develop trend. In [11], the research analyze the small world phenomenon in patent citation networks. But these papers analyze the citation networks from other aspects and are targeted on a different goal from our goal.

IV. Methodology

In this section, we describe our methodology that was adopted in the study of Grid Computing related citation network. The objective of the study is to analyze the structure and characteristic of the citation network and identify a few papers that have significant impact on the development of Grid Computing and thus researchers in the field of Grid Computing need to pay attention to and look into. We also want to extract the main path out of the citation network so that we can trace the development path of Grid Computing and continue to advance the research in Grid Computing. two steps are involved in our study:

- Network modeling
- Network simulation

Below we will describe in detail the process of each step.

A. Network modeling

It is obvious that a citation network has a graph-like structure. Here we define the Grid Computing related citation network as follows: each paper is regarded as a node, and the citation relationship between papers is represented using arcs. For example, if a paper A cited another paper B, we add an arc starting from B and ending at A (See Figure 4). So the number of arc starting from A represents the number of papers that A cited and the number of arc ending at A stands for the number of papers in which A is cited. In addition, it is almost impossible that a cycle is formed in a citation network. A paper can be cited only when it is published and thus its publishing date is earlier than those citing it. So it can not cite other papers that already cited it so that a cycle is not suppose to appear in the citation network. In fact, a citation network is modeled as a DAG (Directed Acyclic Graph).

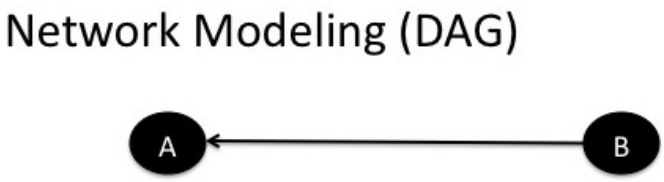

Fig. 4. A network model

B. Network simulation

Based on the network model, we can now collect real data in the Internet, select a useful tool and start to analyze the data.

1) Data Collection:
$CiteSeer^X$ (*http://citeseerx.ist.psu.edu/*) is a famous scientific literature library and search engine that focuses primarily on the literature in computer and information science. One of its feature is that it provides a lot of metadata and related services for scientific literature, such as citation statistics and reference linking. So $CiteSeer^X$ is an appropriate candidate as the data source for our study. Grid Computing starts in 1990's and has been one of the hottest research topics in the academic community. As a result, there exists thousands of Grid Computing related papers in $CiteSeer^X$. Due to the limitation of time and resources, our study can only extract a little part of these papers and analyze their citation network. In the process of data collection, we make two assumption. First, we assume that most of the grid related papers has the keyword "grid" in their title. Second, we believe that the more an article is cited, the more important the article is to the development of Grid Computing. Following our assumption, we collect a sample of 52 papers which is representative of the whole set of Grid Computing related paper. The following is the steps of data collecting:

- Search for the papers with the title containing the keyword *grid* in $CiteSeer^X$.
- Sort the search result in the descending order of number of citations that each paper has.
- Pick out the papers that obviously do not belong to Grid Computing.

Table I lists the title, label and the number of citation of 10 most cited papers in our data collection.

TABLE I
10 MOST CITED PAPERS IN GRID COMPUTING

Title	Label	Citation
The Anatomy of the Grid: Enabling Scalable Virtual Organizations	001	1327
The physiology of the grid: An open grid services architecture for distributed systems integration	002	811
Grid Information Services for Distributed Resource Sharing	003	425
A Security Architecture for Computational Grids	004	343
Condor-G: A Computation Management Agent for Multi-Institutional Grids	005	309
The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets	006	297
Nimrod/G: An architecture for a resource management and scheduling system in a global computational Grid	007	212
High Performance Parametric Modeling with Nimrod/G: Killer Application for the Global Grid	008	176
The Grid: Blueprint for a Future Computing Infrastructure	051	1076
Globus: A Metacomputing Infrastructure Toolkit	052	1240

2) Network Analysis tool: Pajek:
Pajek [12], [13] is a program for analysis and visualization of large networks. It has widely used as an efficient analysis tool in all kinds of networks, such as social networks, Internet networks and ISP networks. Pajek provides a wide range of powerful functionalities for network analysis while it is very easy to install and use. Also it is freely available for noncommercial use. All these are the reasons why we choose Pajek as our analysis tool.

In Pajek, six types of objects are used and each type of objects has its input file format. These objects and their corresponding input file extension is as follows:

- Network: Vertices and lines. (.net)
- Partition: nominal or ordinal properties of vertices. (.clu)
- Vector: Numerical properties of vertices. (.vec)
- Cluster: Subset of vertices. (.cls)
- Permutation: Reordering of vertices. (.per)
- Hierarchy: General tree structure on vertices. (.hie)

Pajek also provides a large set of functionalities that facilitate and simplified the network analysis. The citation network we are studying belongs to Pajek's network object. A list of functionality aiming at network object is as follows:

- Manipulate the network. Add, delete and modify the vertices and lines in the network;
- Retrieve general information about the network. Such as the number of vertices, the number of arcs, edges and loops, density of lines, average degree and so on;
- Path. Shortest paths, all paths between two vertices;
- Reordering. Topological ordering, Richchards's numbering, depth/breadth first search;
- Flows. Maximum flow between two vertices;
- Critical paths;
- Visualization;

With the assistance of Pajek, we can analysis the citation network from different perspectives with a little bit learning curve. Otherwise, we would have to program to implement the functionality we need, which is much more time-consuming. The following steps show how we use Pajek to facilitate our study.

a) Installation: For this study, we use Microsoft Windows XP Professional platform. Download pajek125.exe from *http://pajek.imfm.si/doku.php?id=download* and run the installation program.

b) Data input: Pajek requires the input file to conform to the special format. We have to make up the input file manually.

c) Visualization: Draw a picture representing the citation network (See Figure 7).

V. EXPERIMENT AND RESULT

A. General Information

As shown in Table II, our citation network has 52 papers, and the number of citation between these papers is 128. The network diameter is 4, which means that there are at most 4 nodes at the shortest path of every pair of nodes. The average path length is 1.69, which is small and as a result we think the citation network is a small world network.

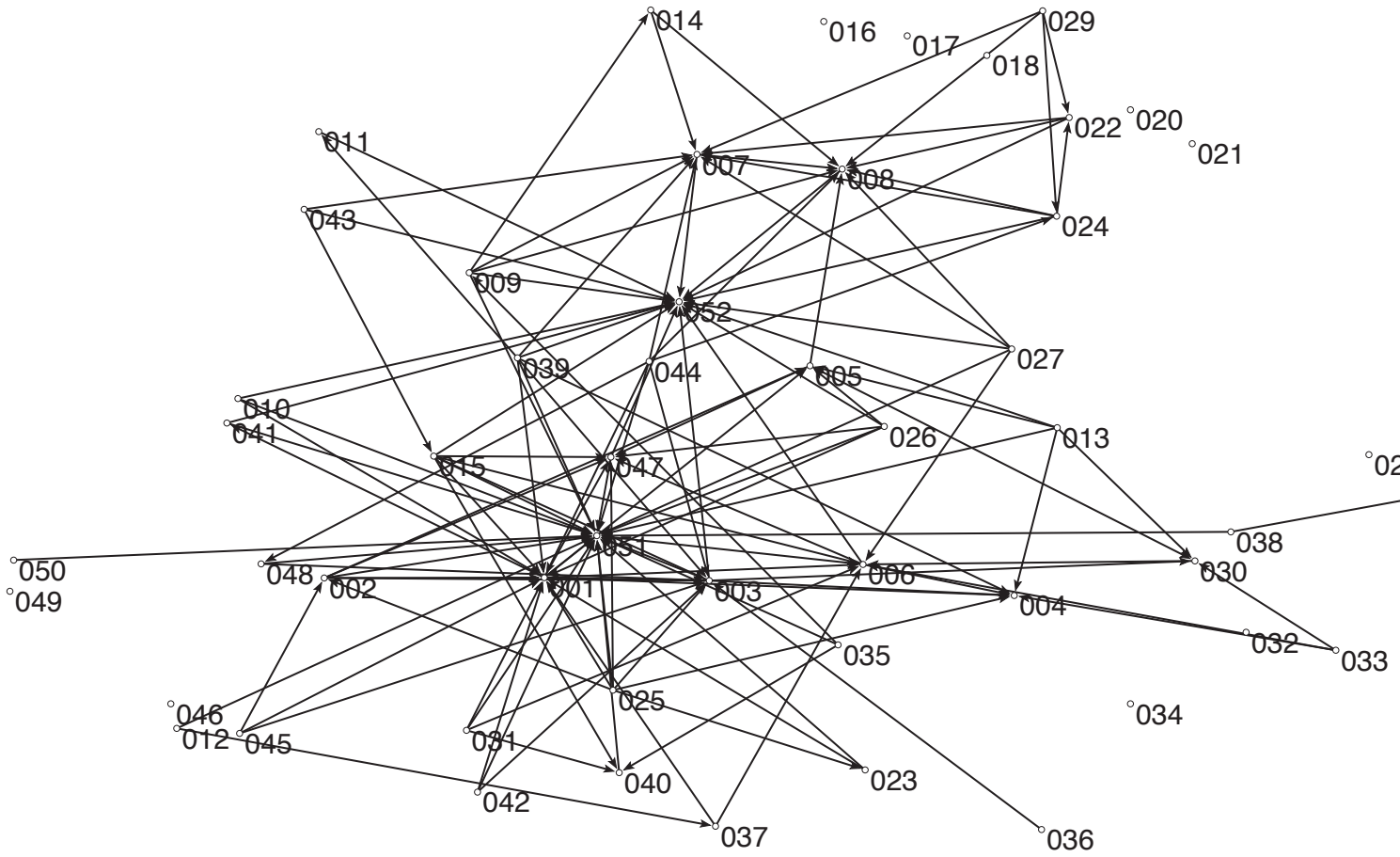

Fig. 5. Visualization of the citation network

TABLE II
GENERAL INFORMATION

Node number	52
Link Number	128
Radio L/N	2.46
Maximum in-degree	24
Maximum out-degree	7
Network diameter	4
Average path lengh	1.69

B. Weakly Connected Subgraphs

In the data collection process, we use the key word "grid" to find the Grid Computing related papers. But that does not guarantee that all of the selected papers are related to Grid Computing. As another way to validate our data collection, We conducted a depth first search on the citation network and try to exclude unrelated papers. We assume that if the papers are all related to Grid Computing, they should be in a weakly connected graph. The result shows that there are 10 weakly connected subgraphs. As shown in Table III, 9 0f these subgraphs have only one vertex while one subgraph contains 33 vertices. The subgraph with 33 vertices represents the main research progress in Grid Computing since a few classic papers [14], [15], [16] are all in this subgraph. Among the papers in the other subgraph, some have nothing to do with Grid Computing. Some are related to Grid Computing, but primarily focus on other research topics. So researchers in Grid Computing can pay less attention to or just skip these papers

TABLE III
WEAKLY CONNECTED SUBGRAPHS IN THE CITATION NETWORK

Subgraph	**Papers in the subgraph**
1	016
2	017
3	018
4	020
5	021
6	028
7	034
8	046
9	049
10	001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 019 022 023 024 025 026 027 029 030 031 032 033 035 036 037 038 039 040 041 042 043 044 045 047 048 050 051 052

C. Directed Acyclic Graph

In general, a citation network should be a directed acyclic graph. But in case of wrong data or other reason, a loop can still be formed in the citation network. We conduct a topological sorting on the citation network to check if there are some loops in the network. Surprisingly, we do find a loop in the citation network. The paper 001 [14] and the paper 005 [17] are cited with each other. According to DBLP(*http://dblp.mpi-inf.mpg.de/dblp-mirror/index.php*), [14] published in three different conferences or journals while [17] published in two different conferences(See Table IV). We also compare different versions of [14]'s. The first version did not cite [17] while the later version did. So this can explain why we can find a loop in the citation network.

TABLE IV
TWO PAPERS CITED WITH EACH OTHER

Author	Title	Publisher
Ian T. Foster	The Anatomy of the Grid: Enabling Scalable Virtual Organizations	CCGRID 2001
Ian T. Foster	The Anatomy of the Grid: Enabling Scalable Virtual Organizations	CoRR 2001
Ian T. Foster	The Anatomy of the Grid: Enabling Scalable Virtual Organizations	Euro-Par 2001
James Frey	Condor-G: A Computation Management Agent for Multi-Institutional Grids	Cluster Computing 2001
James Frey	Condor-G: A Computation Management Agent for Multi-Institutional Grids	HPDC 2001

D. In-degree and Out-degree

In our citation network, a node's in-degree is the number of the papers who cited it, and a node's out-degree is the number of the papers it cited. Table V, VI lists in-degree and out-degree distribution of our citation network. In Table V, we noticed that some papers are much more cited than others. For example, paper 051 is cited 24 times and paper 052 is cited 17. Generally, the more a paper is cited, the more contribution it gives to academic community, the more possibly it is a worthwhile paper. Figure 6 shows the in-degree distribution of the network and Table VII lists the important literatures with high in-degree in the network.

E. Topological Sorting

As discussed above, we can identify a few classic papers with the extremely high citations. However, sometimes researchers would be more interesting in an overview of the research progress in Grid Computing, for instant, how many research topics are there in Grid Computing, how does the research in Grid Computing advance. To answer these questions, we conduct a breadth first search on the citation network. As shown in Figure 7, researchers can easily know the development path of the research in Grid Computing.

TABLE V
THE IN-DEGREE DISTRIBUTION OF THE CITATION NETWORK

Label	In-degree	Label	In-degree
001	14	002	2
003	9	004	7
005	4	006	7
007	9	008	9
009	1	011	1
014	1	015	1
019	1	022	2
023	1	024	2
030	4	037	1
040	3	041	1
047	7	048	1
051	24	052	17

TABLE VI
THE OUT-DEGREE DISTRIBUTION OF THE CITATION NETWORK

Label	Out-degree	Label	Out-degree
001	14	002	2
003	9	004	7
005	4	006	7
007	9	008	9
009	1	011	1
014	1	015	1
019	1	022	2
023	1	024	2
030	4	037	1
040	3	041	1
047	7	048	1
051	24	052	17

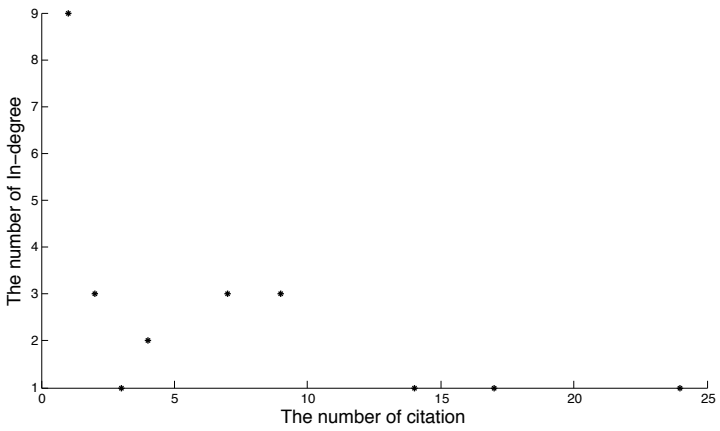

Fig. 6. In-degree distribution

VI. SIGNIFICANCE AND LIMITATION

In this work we aim to find the most useful and classic literatures via the analysis of the citation network. We first exclude literatures unrelated to our theme by means of dividing the network into a number of weakly connected subgraphs and excluding the subgraphs with little vertices. We also conduct a topological sorting to detect if there exist some cycles

TABLE VII
PAPERS WITH HIGH IN-DEGREE

In-degree	Title	Citation
24	The Grid: Blueprint for a Future Computing Infrastructure	1327
17	Globus: A Metacomputing Infrastructure Toolkit	1076
14	The Anatomy of the Grid: Enabling Scalable Virtual Organizations	1329

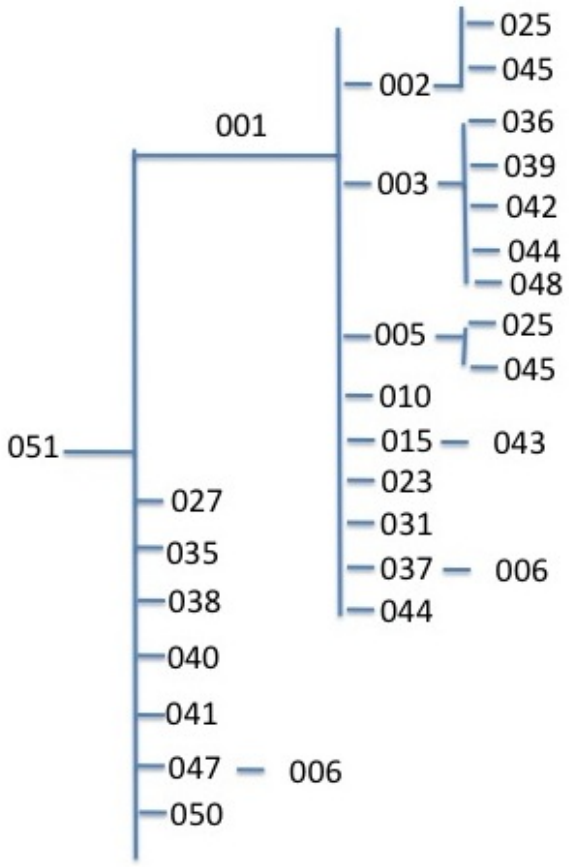

Fig. 7. Development path of Grid Computing

in the graph. Then we analyze the in-degree distribution of the network and find a few vertices with extremely high in-degree which we think represent significant literatures. The significance of this work lies in the following aspects:

- We drew a simple but useful citation network model that can be applied to other types of citation networks. Besides, the analysis techniques used in this study can be also applied to other cases of citation network analysis.
- The paper illustrates a comprehensive introduction of the data collection process we adopted in our study. For example, how to collect data in the website, how to select tools and use the appropriate tool to deal with the data.
- Our work has a very innovative idea. It leverages the analysis result of the citation network to select the more useful literatures from much more literatures. Although there are already a lot of works on the analysis of the citation network, these works are targeted on other goals.

However, there also exist a few limitations in this paper, including

- Due to the limitation of time and resources, the data we collected is limited. We think that the study based on more data would be more significant.
- As is known in academic community, self-citation exists widely and these self-citations have little academic implication. We think that our work would be improved if we take the self-citations into consideration in our study.
- Other than in-degree, there are also other metrics we can use to judge the nodes' centrality and thus identifying the more important nodes in the networks. Our work would be improved by combining other metrics in our study.

VII. CONCLUSION

In this study, we conduct a study on a Grid Computing related citation network with the goal of identifying appropriate literatures for researchers interested in Grid Computing. We develop our methodology for network analysis, including modeling methods and data collection process. Based on the methodology, we first exclude literatures unrelated to our theme by means of dividing the network into a number of weakly connected subgraphs and excluding the subgraphs with little vertices. We also conduct a topological sorting to detect if there exist two or more literatures citing with each other. Then we analyze the in-degree distribution of the network and find a few vertices with extremely high in-degree which we think represent significant literatures. Our result shows that our methods is effective and efficient to find useful literatures for researchers.

REFERENCES

[1] "Graph Theory." [Online]. Available: http://en.wikipedia.org/wiki/Graph_theory
[2] "Random Graph." [Online]. Available: http://en.wikipedia.org/wiki/Random_graph
[3] "Random Graph." [Online]. Available: http://en.wikipedia.org/wiki/Small-world_network
[4] "Random Graph." [Online]. Available: http://en.wikipedia.org/wiki/Scale-free_network
[5] "Network Theory." [Online]. Available: http://en.wikipedia.org/wiki/Network_theory
[6] "Average path length." [Online]. Available: http://en.wikipedia.org/wiki/Average_path_length
[7] "Six Degrees of Separation." [Online]. Available: http://en.wikipedia.org/wiki/Six_degrees_of_separation
[8] "Clustering Coefficient." [Online]. Available: http://en.wikipedia.org/wiki/Clustering_coefficient
[9] "Centrality." [Online]. Available: http://en.wikipedia.org/wiki/Centrality
[10] W. Wu, Y. Han, and D. Li, "The Topology and Motif Analysis of Journal Citation Networks," in *International Conference on Computer Science and Software Engineering*, 2008.
[11] S. Hung and A. Wang, "A small world in the patent citation network," in *IEEE International Conference on Industrial Engineering and Engineering Management, 2008. IEEM 2008*, 2008, pp. 1–5.
[12] V. Batagelj and A. Mrvar, "Pajek-program for large network analysis," *Connections*, vol. 21, no. 2, pp. 47–57, 1998.
[13] ——, "Pajek-analysis and visualization of large networks," *Lecture notes in computer science*, pp. 477–478, 2002.

[14] I. Foster, "The anatomy of the grid: Enabling scalable virtual organizations," *INTERNATIONAL JOURNAL OF SUPERCOMPUTER APPLICATIONS*, vol. 15, no. 3, p. 2001, 2001.
[15] I. Foster and C. Kesselman, *The grid: blueprint for a new computing infrastructure*. Morgan Kaufmann, 2004.
[16] I. Foster, "The physiology of the grid: An open grid services architecture for distributed systems integration," 2002.
[17] J. Frey, T. Tannenbaum, M. Livny, I. Foster, and S. Tuecke, "Condor-g: A computation management agent for multi-institutional grids," in *Cluster Computing*, 2001, pp. 237–246.