Small-world phenomena in paper reference networks

Fang Xi

Shanghai Jiao Tong University
seefun@sjtu.edu.cn

Abstract

This report introduces the history and theory of traditional small-world networks in social science.And in the paper I design a expriment to find whether this phenomena exists in paper reference networks.This absolutely new research preliminarily represents a phenomena which is simlialr to "Six degrees of separationin"in some relatively small dataset of paper reference networks.But there also have many things to do to improve this expriment and find more intersting phenomena in parper reference networks.

I. Introduction

Small-world network is a type of mathematical graph in which most nodes are not neighbors of one another, but the neighbors of any given node are likely to be neighbors of each other and most nodes can be reached from every other node by a small number of hops or steps.

Small-world always appears in social science and management.It has two important properties: Low average hop count and High clustering coefficient.

Small-world properties are found in many real-world phenomena, including websites with navigation menus, food chains, electric power grids, networks of brain neurons, voter networks, telephone call graphs, and social influence networks.Cultural networks and word co-occurrence networks have also been shown to be small-world networks.

And the "six degrees of separation" is often used as a synonym for the phenomenon of " small world ". The concept of six degrees separation is that all things in the world are six or less steps away from each other, and a series of "friends" statements can connect up to two friends in any of the six steps. It was originally set out by Frigyes Karinthy in 1929 and popularized in an eponymous 1990 play written by John Guare.

In my report, I will use the small-world theory, and try to find whether "small-world" exists in paper reference networks or the "six degrees of separation" exists.I use the Microsoft Academic Graph(MAG) text networks dataset[1] and programming in python3 to do a experiment. In the expriment I get and visualize the path distribution between any two points in the reference graph(both directed graph and undirected graph), and record some properties.

II. Small-world theory

We can conclude two important properties of small world networks : low average hop count and high clustering coefficient.So,to quantify a small world,that two network measures can be used: average path length (L) and the clustering coefficient (CC). L measures the average number of intermediaries, that is, the degrees of separation, between any two actors in the network along their shortest path of intermediaries. The shorter the average path length, the closer people, resources, or ideas theoretically are to each other in the network. The CC measures how many of an actor's contacts

[1]Dataset: http://acemap.sjtu.edu.cn/acenap/index.php/datasets.html

are connected to each other. When many of an actor's contacts are connected to each other, the actor has a highly clustered network.

A small-world network requires L between two randomly chosen nodes (the number of steps required) grows proportionally to the logarithm of the number of nodes N in the network, that is:

$$L \propto logN$$

If the clustering coefficient is not small, in a big assemble or cluster, strangers may be linked by a short chain of acquaintances.A great example is " six degrees of separation ". Figure 1 shows a example of it.

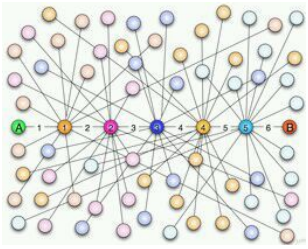

Figure 1: *Six degrees of separation*

We can express L as below : (L is a important value in our experiment shown next)

$$L = \sum_{i,j} \min L_{i,j} \ , \ L_{i,j} >= 1$$

And the clustering coefficient(CC) can be calculated by:

$$CC = \frac{closed}{closed + open}$$

That is: number of closed triplet / number of triplet(closed+open). A example is below:

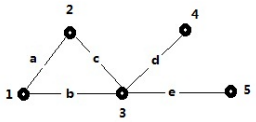

Figure 2: *clustering coefficient*

The number of closed triplet is 3 and the number of triplet is 8, so CC = 3/8 .

Apart from calculate these two traditional properties, we can find a method to get the distribution of path length and analyze it .

III. Experiment in paper network

In this experiment I use Microsoft Academic Graph text networks datasets. There are 32 networks file in this dataset. To start with, we use the first network in this datasets,which is papers of Internet privacy. This network has 749 nodes and 749 directed edges.

In the experiment, I'll get the distribution of hops or the path length between any two nodes in the network. And I will calculate the average shortest path length.

We do this thing in directed graph. Using python3 and the tool package "networkx" to bulid this network from MAG Internet privacy dataset. Firstly, I build the graph as a directed graph using networkx.For example, if paper A refer to paper B, I'll draw a directed edge from node A to node B.

After the directed graph of paper reference network has been constructed,traverse all pairs of nodes in the python program. Using depth-first search to find all the path between two nodes. Write done all the node pairs in a csv file if the path exists. Than we can calculate the number of paths, and calculate the average length or called average hops between one node pair. And write down all the solutions and the list of hops in the csv file. This is the first step of data mining.

After the extract step, we store the more intuitive and more useful message of the network in a csv file. Than we will do further analysis. Program and calculate the number of pairs between which have paths. The result is that the number of pairs existing roads is 6108. We can also get the average length/hops of the existing road is 5.925671, the expected number of hops between two nodes is 3.330856, the expectation of the minimum number of hops between two nodes of any existence path is 2.671415 .That means the L in this dataset is 2.671415.

$$\bar{L_{01}} = 5.925671$$
$$E(L_{01}) = 3.330856$$

L in this dataset:

$$L_{01} = 2.671415$$

We can know from the result that in the filed of Internet privacy(the graph dataset we use now), we are very likely to find the association between two papers within three hops. This result is correct in some relatively small dataset like Internet privacy graph. If the dataset is more bigger,we can also use this method and program, and the L is bigger and ti appears many clusters, which we can see from acemap. But the time of calculate will increase sharply, which will be analyzed in part four of this paper.

After get these results, we can also use python with matplotlib and seaborn (two third-part python packages) to visualize the path distribution between any two nodes with paths. The visulization result of my program is that(Fig.3):

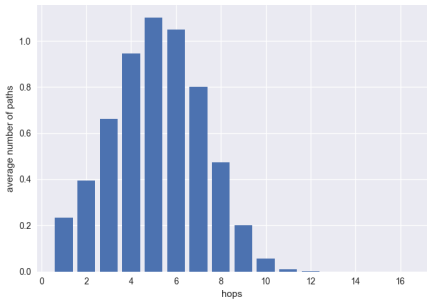

Figure 3: *distribution of hops(directed graph)*

We can visually understand the hops distribution from the above figure. Such a distribution curve may be a new characteristic of small-world .The path length is centered between 3 and 7 hops. The paths of 5 hops have the largest number.

And we can also visualize the probability distribution of hops of shortest path between any pair of nodes which have paths to join them. Through this probability distribution, we can easily calculate the L ,which is approximately between 2 and 3. We have get from program that $L = 2.671415$. (Fig.4)

How about bulid this network graph as a undirected graph? For example, if paper A refer to paper B, I'll draw a undirected edge between node A a node B.Using the same method

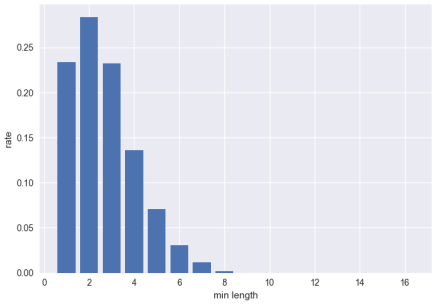

Figure 4: *distribution of min length (directed graph)*

as above,we can get the result which have a little difference. The number of pairs existing roads is 36002(with in 3 hops), which is apparently bigger than this of directed graph. We can also get the expectation of the minimum number of hops between two nodes of any existence path is smaller than 2.617 .That means the L smaller than L in directed graph but not apparently.

We can find the number of path is increase very sharply when I change the directed graph to undirected graph. And the using time of python program also have a substantial increase, although the graph is not so big. So a two-way connection makes the connection graph more dense, and the small-world clustering is more apparent.

Limited by the computing power of my laptop, the statement above is only the result within three hops. And we can visualize the result similarly.

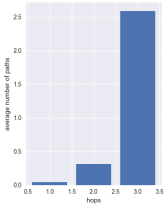

Figure 5: *distribution of hops (undirected graph,within 3 hops)*

The number of paths increasing sharply when the hops increase.

IV. Future works

First,The bigger dataset in MAG datasets can also use this method . Limited by the computing power, we only use one small dataset: 01-MAG-Internet-privacy. And the algorithms used now to find all paths between two nodes is a modified depth-first search. A single path can be find in $O(V + E)$ time, but the number of single paths in a graph can be very large, e.g. $O(n!)$ in the complete graph of order n. Total time the program used may up to $O(V! * V^2)$ .Once the network is large enough, the computational complexity will be too large to imagine. This is the foundation reason of the small-world phenomena, and it is also difficult for us to research large datasets because of it. Maybe we should use more powerful computer and improve our algorithms in the future.

Second, clustering coefficient(CC) is another important property of small-world network. this argument will be programmed to calculate in the next step. A big paper reference network may have many small clusters,and we will do further research in this aspect. And it is also a interesting work to find difference of hops distribution ,L and CC between different size of networks.

Third, we can unearth more interesting information using the datasets and the method. We can do much more things in the future work.

V. Conclusion

We can conclude from the experiment that, in paper reference network, any two paper have great possibility connecting with each other by many path.And minimum length of path between two nodes most possibly very small. And the average path length L is also small. Hops distribution between arbitrary two paper may similar to Figure 3. The small-word phenomena do exist in paper reference network in some aspect.

Through this course project, I am able to grasp the skill of python programming and using some visualization tool and many algorithms about graph. And I also read some papers and master many theory and algorithms in practice.

It's my honor to accomplish this paper in prof. Wang's course. I should thank all the teachers for their guidance and help.

References

[1] Small-world networks and management science research: a review (2007).
Brian Uzzi, Luis AN Amaral, Felix Reed-Tsochas

[2] How small is the center of science? Short crossdisciplinary cycles in co-authorship graphs (2015).
Chris Fields

[3] A small world of citations? The influence of collaboration networks on citation practices.
Matthew L. Wallace, Vincent LariviÃĺre and Yves Gingras

[4] Wikipedia: *en.wikipedia.org*