A DESCRIPTION ALGORITHM FOR COMMUNITY STRUCTURE

Lei Zhang, Zhixiong Zhao, Bin Wu, Juan Yang

Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia,
Beijing University of Posts and Telecommunications, Beijing, China
lzhang@tseg.org, zhaozhixiong@tseg.org, bwu@tseg.org, yangjuan@bupt.edu.cn

Abstract

In the last decade, a large number of graph mining algorithms have been proposed. But there are only a few descriptions about community structure. The communities in different network have different structure, and even in the same network the communities may have different community structure. If we can't describe the community structure reasonably, it is difficult to use the communities which are gotten from the community detection algorithms. Many community detection algorithms may have no meaning. In this paper, the community structure would be described from four different aspects. They are inside properties which describe the community in terms of the community itself, outside properties which describe the community in terms of relationship between communities, level properties which describe community in terms of relationship between the large community and the small communities which compose to the large community at different level, and dynamic properties which describe the evolution information of the communities in different time. Futher, a description algorithm based on the statistic is proposed. In this description algorithm, the community structure information can be descriped in detail and can be used for futher analysis. Also, the community structure can be described in different levels by choosing different statistic rules. A data structure is also proposed to save the community structure information for the purpose of searching it quickly.

Keywords: Community Structure; Complex Network

1 Introduction

The complex system can be seen as a graph in which nodes are elementary units and links are the interactions or relations. Community structure is an important feature for complex network [4]. Community is a group of vertices, with many edges joining vertices of the same community and comparatively few edges joining vertices of different communities.

Zachary used Kernighan-Lin algorithm to the Karate Club. The network was divided into two communities. In fact, Karate Club had already become two parts because the two leaders had different opinions about fee. The result is corresponse to the fact [7]. Pothen used matrices with eigenvectors to divide the Karate Club network. And the two communities were also gotten. Newman used GN algorithm to Computer-Generated Graphs, Zachary's Karate Club, and College Football. The result is almost accurate. By using GN method on Collaboration Network, it was found that each scientist coauthored articles with approximately five others. The community represented by circles is comprised of a group of scientists working on mathematical models in ecology and forms a fairly cohesive structure, as evidenced by the fact that the algorithm does not break it into smaller components to any significant extent [5] [9]. Palla used CP algorithm to the co-authorship network of the Los Alamos Condensed Matter archive and collected a few statistical properties of the network of communities [6]. Although a large number of graph mining algorithms have been proposed, in most cases, communities are algorithmically defined, i.e. they are just the final product of the algorithm, without a precise a priori definition [4].

Firstly, all the descriptions about community are just the descriptions about the result of community dectection algorithm. Different algorithms may result in different descriptions for the same community. Secondly, due to the large scale of the network, there may be thousands of communities. The average value of a community property just can describe the general characteristic of a community structure. The average values include the average shoretest path between any two nodes in the community, the average degree for any node in the community, the average out-degree for any node in the community and so on. On the one hand, only one average value can't show the community structure in detail. On the other hand, the value is obsolute, so it can't show the intrinsic property. For example, we can't say two communities are similar

978-1-61284-204-2/11/$26.00 ©2011 IEEE

if they have very different sizes because they will obsolutely have very different values. Moreover, it is impossible to analyze the community structure even futher by using only one average obsolute value.

We propose a description algorithm which results in a sequence. One sequence can show one community property which belongs to inside properties, outside properties, level properties or dynamic properties. In this description algorithm, we get the statistic of property first, and then the statistic rule is applied on the statistic of property. Because the time cost here is much more than just calculating the average value, it is much better to save the result than to calculate it every time.To make it easy to search for analyzing, The data about communities structure properties is save in a data structure like BigTable.

This article is structured as follows. In Section 2, we present previous related work. In Section 3, we show our description algorithm based on the both distribution and statistic rule, and data structure to save the result. In Section 4, we run detailed experiments on a telecommunication network. Section 5 discusses our experiences and concludes this paper.

2 Related works

Jure Leskovec [2] proposed NCP plot (network community profile plot) to measure the quality of network communities at different size scales. This kind of description describes one property of all the communities in the network. For each community, it has one value. This value can be gotten from the structure of community and the structure of network.

Andrea Lancichinetti [3] analyzed many properties of a variety of communities. For the choosen property in the network, one value is response to one community. So it has a distribution for each property, such as the distribution of the community size, the distribution of degree and so on. The result is shown in Figure 1. For the Information network, we can see that most of the communities have the size between 100 and 1000. But for the Communication network, it is not the same. It has almost the same number of communities at any community size. From this method, what we know is how different communities are distributed in the network for one property (community size here). But it says little about the structure of the community itself. It also has only one value for one community, such as community size here. It shows more about the similarities of all the communities in the network.

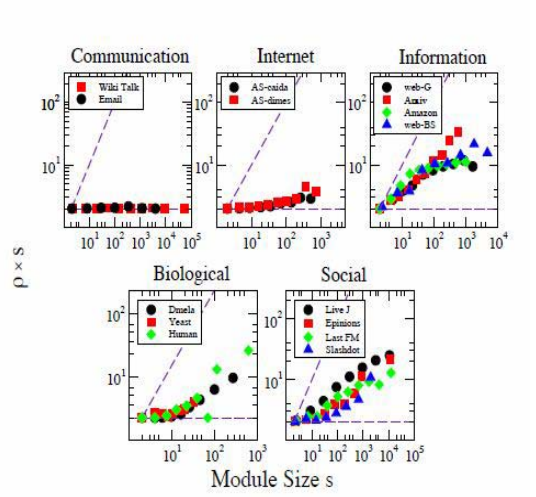

Figure 1

BigTable is used by more than sixty Google products and projects, including Google Analytics, Google Finance, Orkut, Personalized Search, Writely, and Google Earth. The map is indexed by a row key, column key, and a timestamp; each value in the map is an uninterpreted array of bytes.

Figure 2 is a slice of an example table that stores Web pages. The row name is a reversed URL. The contents column family contains the page contents, and the anchor column family contains the text of any anchors that reference the page. CNN's home page is referenced by both the Sports Illustrated and the MY-look home pages, so the row contains columns named anchor:cnnsi.com and anchor:my.look.ca. Each anchor cell has one version; the contents column has three versions, at timestamps t3, t5, and t6 [1].

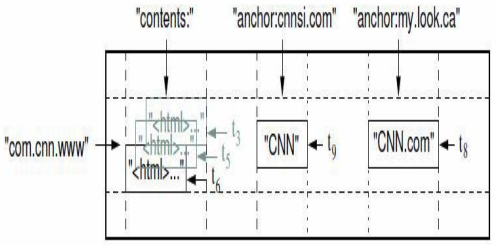

Figure 2

3 Description algorithm and data structure

3.1 Description algorithm

The algorithm is shown in Figure 3. Below is the detail in every step.

Step 1 Calculate the statistic of one choosen property for one community
Step 2 Normalize the statistic of property
Step 3 Choose the percentage statistic rule
Step 4 Calculate number statistic rule
Step 5 Calculate the final statistic result
Step 6 Save the final statistic result in the data structure above
Step 7 Choose another property and continue from step1 until all the properties are calculated.
Step 8 Choose another community and continue from step 1 until all the communities in the network are calculated

Figure 3

Step 1 Calculate the statistic of one property for one community
The statistic is for the units of the community, such as the nodes in the community. The statistic for this property is a sequence. The position of one element in the sequence is response to the "property value". The value of the element in the suquence is the count of units which have the response property value. The result is the statistic of property. For example, let the property be in-degree and the units are nodes in the community, so the "property value" is the in-degree of one node. The statistic of property is a sequence. The position i in the suquence means that the in-degree of node is i. The number m in position i means that there are m nodes that have in-degree i. The size of the sequence depends on the largest in-degree of node in the community.

Step 2 Normalize the statistic of property
In this step, the sequence gotten from step1 will be processed by normalizing it. The counts are changed to be percentages. The result is the normalized statistic. For example, if the statistic of property is a b c d, the normalized staticstic can be gotten: a/(a+b+c+d) b/(a+b+c+d) c/(a+b+c+d) d/(a+b+c+d)

Step 3 Choose the percentage statistic rule
The percentage statistic rule is a sequence. Any element in the sequence is percentage. The sum of all the elements in the sequence is 1. The size of the sequence determinates the final statistic result parts. The value m at position i in percentage statistic rule means the ith value in the final statistic result is the summation of m (m is a perceantage) elements in the normalized staticstic. For example, if the normalized staticstic is a/(a+b+c+d) b/(a+b+c+d) c/(a+b+c+d) d/(a+b+c+d), percentage statistic rule is 0.5 0.5, and final statistic result is r1 r2. We can know that r1 is the summation of the first 50% elements in normalized staticstic; r2 is the summation of the next 50% elements in normalized staticstic. But how can we get percentage statistic rule? It is very important. The percentage statistic rule should be made considring both the community structure and the network feature. The samplest percentage statistic rule is all elements in percentage statistic rule are all the same. For example, deviding 1 by n and n is the count of parts that is contained in final result. In this percentage statistic rule, every part has the same percentage elements. So the percentage statistic rule is 1/n 1/n 1/n 1/n …1/n and there are n elements in total.

Step 4 Calculate number statistic rule
The number statistic rule is also a sequence. It replaces the percentages by numbers in percentage statistic rule. To get it, we need percentage statistic rule and the size of the normalized staticstic. After multipel percentage statistic rule by the size of the normalized staticstic, the number statistic rule is gotten. The sum of all the elements in the number statistic rule is the size of the normalized staticstic. The value k at position i in number statistic rule means the ith value in the final statistic result is the summation of k elements in the normalized staticstic. The number in every part in the number statistic rule is rounding when the part is not the last one. The number in the last part is the result of the size of the normalized staticstic subtracting the sum of numbers in other parts. For example, statistic rule is 0.25 0.25 0.25 0.25 and the size of the normalized staticstic is 11. The number in every part is 3 3 3 2 respectly.

Step 5 Calculate the final statistic result
To get the final statistic result, normalized statistic and number statistic rule are needed. For the value ni at position i in number statistic rule, it means the ith element in final statistic result is the summation of ni elements in normalized staticstic. The start position of the ni elements is the next element of last element in the ni-1 elements. For example, if the normalized staticstic is a/(a+b+c+d) b/(a+b+c+d) c/(a+b+c+d) d/(a+b+c+d), number statistic rule is 2 2, and final statistic result is r1 r2. We can know that r1 is the summation of the first 2 elements in normalized staticstic; r2 is the summation of the next 2 elements in normalized staticstic.

Step 6 Save the final statistic result in the data structure in 3.2

Step 7 Choose another property and continue from step1 until all the properties are calculated.

Step 8 Choose another community and continue from step 1 until all the communities in the network are calculated.

3.2 Data structure

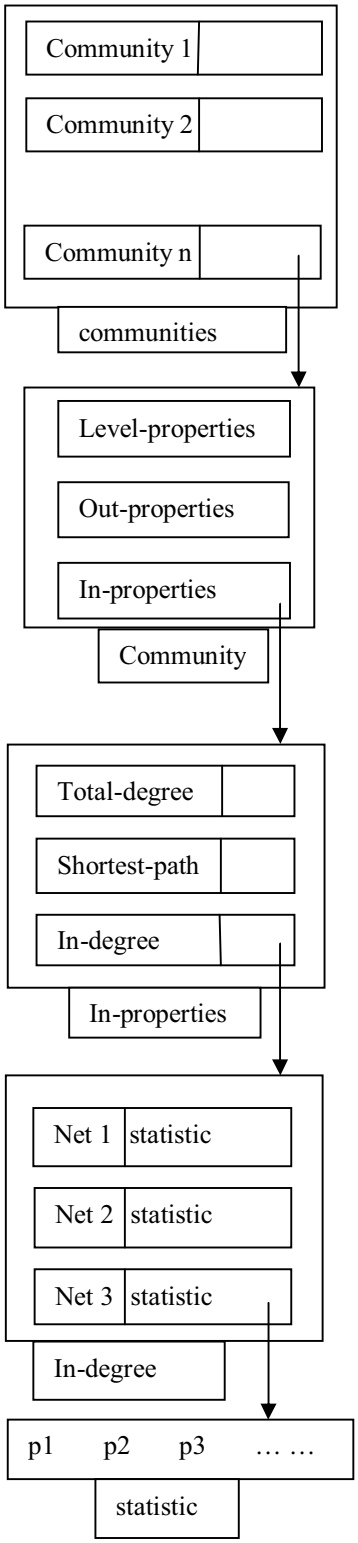

Figure 4

The data structure is shown as Figure 4.
We use a data structure like BigTable to save the statistic information of the communities in a network.

In the "communities" structure, each row is response to one community and the index is the community identification which is a number. So the community identification is like row key in the BigTable. Here we use HashMap to implement it. Every record in HashMap is response to one row here. Key is the community identification and value is response to the statistic information of community structure.

In the "Community" structure, there are three properties in total, in-properties, out-properties and level-properties. These three kinds of properties are response to columns in BigTable.

As to "in-properties" structure, it contains in-degree, shortest path and so on. The property name belonged to in-properties is response to column key in the BigTable. A HashMap is used to save all statistic information of the in-properties. Key is the property name and value is the sequence which is the statistic for that specific property.

To every "In-degree" structure, the net identification is like the timestamp in BigTable. A HashMap is used to save the statistic information for one property at different time. Key is the net identification; value is an array which is used to show the statistic of the specific property at the given time.

This data structure can describe the community structure previously and search easily. Other property data can be added to the structure in time. It is easy to analyze community structure based on this structure.

4 Experiments

The experiment is based on the telecommunication network. Clique percolation method is used as the community detection method and the parameter k equals to 4. The network has 436856 nodes, 763885 edges and 1086 communities. The characteristic of some communities is shown as Figure 5.

Community Id	Node number	Community total-degree	Communiy in-degree
1	55	1622	426
2	118	4208	1368
3	364	11666	5776
4	36	1507	272
5	2241	92336	35612

Figure 5

The result is shown as Figure 6. The first community is black. The second community is blue. The third community is green. The fourth community is red. The fifth community is pink. The

solid line presents the distribution of total degree. The dotted line presents the distribution of the in-degree.

From this figure, we can compare the same property (such as in-degree) of the five different communities. The blue, green, red and pink lines are similar because they assemble most of the nodes about 75% in the first part. But the black line is different from others because it has only 45% nodes in the first part. So, we can say that the first community is different from other four communities. It is about half the percentage of other four communities in the first part, but it is almost at least twice percentage of other four communities in other parts. Althouth the last four communities have different size, they can have similar structure for in-degree property. This similarity is difficult to be shown by using the previous methods because they are too sensitive to the community size.

For the same community, some properties can not correspond with each other. The inconsistent can be shown by the blue line and the pink line. For the blue line, althouth 80% nodes are in the first part for the total-degree property, only 35% nodes are in the first part for the in-degree property. For the pink line, althouth 80% nodes are in the first part for the total-degree property, no nodes are in the first part for the in-degree property. This means most of the nodes have small total degree, but few nodes have very small in degree.

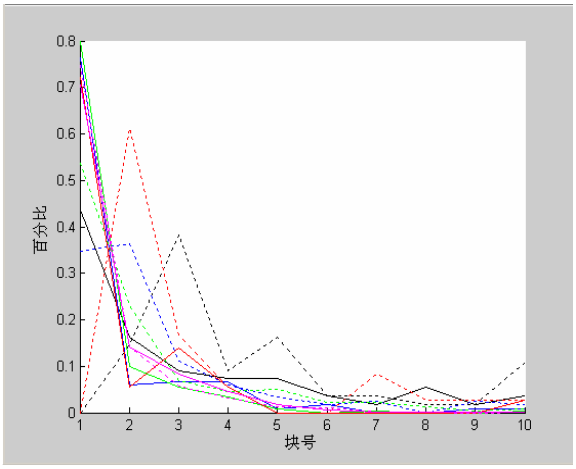

Figure 6 Black, blue, green, red, pink

5 Conclusions

This method is proposed in order to describe community structure but not depend on the community size. In this method, the statistic rule can be changed dynamically to get the different description of the same community structure. It can show the instrinct of the community structure. In this way, the community with different size can have similar structure, which is impossiable for other methods. It is meaningful for community clustring and community evolution. A data structure is also proposed to save the community structure information for the purpose of searching it quickly.

Acknowledgements

This work is supported by the National Natural Science Foundation of China, No. 61074128 and the Fundamental Research Funds for the Central Universities.

References

[1] Chang F, Dean J, Ghemawat S, Wilson C. Bigtable: A Distributed Storage System for Structured Data. OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006.

[2] Leskovec, J, Lang, K, Dasgupta, A. (2008). Statistical Properties of Community Structure in Large Social and Information Networks. Proceeding of the 17th international conference on World Wide Web WWW 08, 7, 695. ACM Press.

[3] Lancichinetti A, Kivela M. Characterizing the community structure of complex networks. PLoS ONE 5(8): e11976. doi:10.1371/journal.pone.0011976.

[4] Fortunato S. Community detection in graphs. Complex Networks Lagrange Laboratory, ISI Foundation, Viale S. Severo 65, 10133, Torino, I-ITALY.

[5] Girvan M, Newman M E J. Community structure in social and biological network. Proc, Natl. Acad. Sci. ,2001,99:7821~7826

[6] Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 205, 435(7043): 814~818

[7] Zachary W. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 1977,33: 452~473

[8] Pothen A, Simon H, Liou K P. Partitioning sparse matrices with eigenvectors of graphs. SIAM J. Matrix Anal. Appl, 1990, 11:430

[9] Newman M E J. Fast algorithm for detecting community structure in networks. Phys. Rev. E, 2004, 69:066133