A Comparison of Community Detection In Social Network Using Modularity

Thanda Tin Yu
University of Computer Studies, Mandalay
thandaryu@gmail.com

Abstract

Social network analysis has undergone a renaissance with the ubiquity and quantity of content from social media, web pages, and sensors. This content is a rich data source for constructing and analyzing social networks. This paper is addressing the problems in constructing the community structure of a networks. Graph analytics have proven to be valuable tools in solving this challenges. Network become an intensive subject of research for example in computer science, networking, network sciences etc., a growing need for valid and useful dataset is presented. Useful ways of addressing this problem are sampling based on the nodes (user)ids in the social network until sufficient amount of data has been obtained. This paper is presented the community detection algorithm such as modularity methods. Then compare the given dataset Vs different data set.

1. Introduction

Community detection provides an important way to further understand and apply social networks. That is, identifying communities or groups in which nodes are densely connected inside while loosely connected outside through some methods[5].

At present, there are many different detection approach and two kinds of among them are applied more. The first is traditional topology-based community detection approach, which maps the real world network into a graph structure with nodes representing users and edges representing the interaction relation between users. Community detection approach based on graph partition or clustering tries to detect sub-graph with high density such as clustering based on vertex similarity, latent space model approximation, spectral clustering and modularity maximization. And discovering communities consisting of similar users is an important problem and can find practical applications in sociology, biology, computer science and other areas. There had been some related work about how to similarity between users based on which people are grouped into communities. One common approach is to treat communities as group of nodes in social network that the connection among themselves are more densely than the rest of network, which makes the community detection a graph clustering problem.

1.1. Related works

William M.Compbell et.al (2013), they presented rich data source for constructing and analyzing the social network using modularity optimization to find the analysis and prediction of individual and group behavior result on real-world data[14] . Benjamin A. Miller et.al (2014), they presented community detection based on topic distance between users depends on the bookmarking relationships between users and tags [3]. Benjamin A. Miller et.al (2013), they presented Varity of techniques exist to analyze graph data sets, that detection and estimation in the classical setting of vector spaces with Gaussian noise using real world dataset. Santo Fortunato (2010) he presented the significance of clustering and how methods should be tested and compared aginst each other to description of application to real networks . Michel Plantiem Michel Crampes (2013) he observed that the development has allowed demonstrating how to share knowledge and information among social network users. Sminu Izudheen et al (2011), they presented that The use of spectral optimization of triangular modularity as and effective method to identify on real biological data . Ming Cheung (2001), he presented that apply the two biological networks such as a collaboration network and a food web that to detect about the edge betweenness. Ming Cheung(2015) presented User shared images are proved to be an easier and effective to discover user connections that investigate user shares images from two social networks, Sky rock and 163 Welbo, in which follower/followee relationship show relatively higher similarity with more practical prediction method.

2. Community detection

Many social networks exhibit community structure are groups of nodes that have high connectivity within a group and low connectivity across groups. Communities roughly correspond to organizations and groups in real social networks. For the purposes of this paper that the communities are disjoint, that is, membership in one community precludes membership in another. In this paper apply modularity in the problem in real dataset. The aim is to partition a set of people into the distinct university group.

2.1. Social Network with Graph

To be partitioning into the distinct group, first to construct the social network, considering from the user(id)s to Mutual friend which crawls nodes of a network in communities are visited one after another. Then social network is extracted from the original data source, it must be stored in structured form that automatic analysis, retrieval, and manipulation are equivalent. The main difference among them is in multiple fields, Two such representations are knowledge representation and graphs.

2.1.1. Knowledge representation and graphs

For every input datum (e.g. text, speech, image), analysts produces a set of objects, attributes and predicates conforming to an ontology that describes structured information in the document. An ontology based on standards for information extraction, primarily the Automated Content Extraction (ACE) protocol is common. An example extraction from a document might be Member (Bob, Karate Club) where Bob is an object of type per (person) and Karate Club is an object of type organization. An important point is that representation is usually limited to binary predicates, i.e. relationships of the form Relation(entity, entity).Another property is that object can have attributes. For instance, it is possible to extract ATT-age(Tina,20).The knowledge representation approach is equivalent to a relational database model. Each predicate correspond to a table as shown in fig 1.

An alternate representation of social network data is to view the knowledge representation structure as a graph. Entities are converted to nodes in the graph which can have different types e.g, people, organizations, and events.

Knows

Tina	John
Tina	Fred
John	Tom
……	……

ATT-age

Tina	20
Fred	32
John	47
…..	… …

Figure 1. Knowledge representation in a relational database. Standard knowledge representation schemes usually involve binary predicates defining relationships between entities. This knowledge representation approach is equivalent to a relational database model as shown above for the predicades, Knows and ATT-Age

3. Clustering methods

Multiple methods for community detection have been proposed in the literature. Many of these methods are analogous to clustering methods. This paper is expressed three methods representatives of standard approaches: modularity optimization, spectral clustering and Infomap.

3.1. Modularity optimization

Modularity optimization is a popular method for community detection. Modularity is an estimate of the “goodness” of a partition based on a comparison between the given graph and a random graph with the same expected degree distribution as the original graph. The method proposed by Clauset, Newman, and Moore is a modularity-based algorithm that address this problem. This paper is presented by spectral clustering algorithm as follow.

3.2. Spectral clustering

Clustering is a popular data mining technique that is used to place data elements into related groups of “similar behaviour”. The traditional clustering algorithm is the so-called k-means algorithm. However, k-means has some well-known problems, i.e. it does not work well on clusters with not well-defined centers, it is difficult to choose the number k of clusters to construct upfront and different initial centers can lead to different final clusters.

In recent years, spectral clustering has become popular and widely used since its results often outperform the outcomes of the k-means algorithm. Spectral clustering is a more advanced algorithm compared to k-means as it uses several mathematical concepts (i.e. degree matrices, weight matrices, similarity matrices, similarity graphs, graph Laplacians, eigenvalues and eigenvectors) in order to divide similar data points in the same group and dissimilar data points in different groups.

Spectral clustering methods are common graph based approach to clustering of data. Spectral clustering methods are attractive, easy to implement reasonably fast especially for sparse data set up to thousands. Spectral methods for community detection rely upon normalized cuts for clustering [4]. A cut partitions a graph into separate parts by removing edges: see Figure 2 as an example. Spectral clustering partitions a graph into two sub graphs by using the best cut such that within community connections are high and across community connections are low. It can be shown that a relaxation of this discrete optimization problem is equivalent to examining the eigenvectors of Laplacian of the graph[14].

,

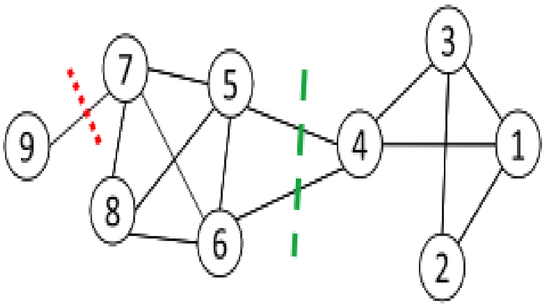

Figure 2 . Spectral clustering of a graph relies on recursive binary partitions or "cut" of the graph, (is illustrated by red dotted line), the "best cut "(shown by green dotted line).

Random walk on graph

– Start in state s with probability P(S1=s)

– Move to next state with probability P(Si =s | Si-1=s')

3.3 Infomap

A graph can be converted to a Markov model in which a random walker on the nodes has a high probability of transitioning to within-community nodes and a low probability of transitioning outside of the community. The problem of finding the best cluster of a graph can be as the compressing the node sequence from the random walk process in an information theoretic sense. The objective of an Infomap is to arrive at a two-level description (lossless code) that exploits both the network's structure and the fact that a random walker is likely to spend long periods of time within certain clusters of nodes. Then the search is for a module partition M of N nodes into m clusters that minimizes the following expected description length of a single step in a random walk on the graph:

$$L(M) = q_{\cap} H(Q) + \sum_{i=1}^{m} p'_{\cup} H(P')$$

This equation comprises two terms: first is the entropy of the movement between clusters, and second is the entropy of movements within cluster, each of which is weighted respectively by the frequency with which it occurs in the particular partitioning.

$q_{\cap}$ =the probability of the random walk switches cluster on any given step

$H(Q)$=the entropy of the top-level clusters

$H(P')$= the entropy of within-cluster movements

$p'_{\cup}$ =the fraction of within-cluster movements that occurs in cluster i. Random walk on graph start in state s with probability P (S1=s) and then move to next state with probability P(Si=s | Si-1=s')

4. Experiment

The first step in the experiments is to exploit the Zachary's karate club, data to obtain a social network for analysis. Queries are designed in SQL to extract people and their mentions in document. Then, a network of documents and individual is constructed on the basis of document co-occurrence. It gets the resulting graph. Community detection methods are applied to the spectral clustering. For the first step, the graph is split using a tree structure. The colors indicate the final communities shown in figure3.

This paper proposes the user(id)s in social network especially facebook from mutual friends to cluster the group of university using the spectral clustering method.

The precision and recall of the algorithms can be quantitatively measured. For any two individuals, it is built if they are same karate group (or not) by using the truth tables from Zachary's karate club which is compared to the predicated membership obtained from the community detection algorithm. The true positive (TP) occurs when both the groups and the communities are the same for the two individuals. A false positive (FP) occurs when the group are not the same, but individuals are placed in the same community. At last, a false negative (FN) occurs when the groups are the same.

The two measures of performance are

$$precision = \frac{TP}{TP+Fp} \text{ and } recall = \frac{TP}{TP+FN}$$

The spectral clustering algorithm has a threshold that allows a wide variation in the trade-off precision versus recall. In general, the trade-off is due to the community(cluster)size. The algorithms can either produce small clusters that are highly accurate but have better recall. Overall, users of these algorithms will have to determine what operating point is best fitted to their application[10].

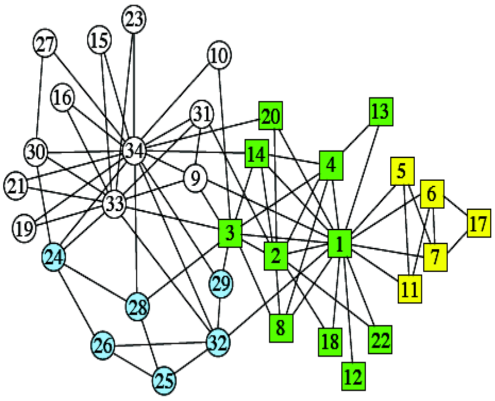

Figure 3 . Zachary’s karate club, the color correspond to the best partition found by optimizing the modularity[10]

4.1 Verification and Evaluation

To prove the correctness of user profile network in term of community detection that compare the results of community assignments to existing approaches. However , the variety of community detection algorithms is too large to compare against by using spectral methods in other data set as shown in table1.

As all mentioned approaches are not directly providing a map of community ids to node ids the choose the partition resulting merge of nodes leading to a maximum of modularity. This partition is then compared to the output of algorithms found in the same result as indicated by color as shown in example of Zachary’s karate club.[6]

A comparison of the mentioned methods on some selected datasets in given in the following table 1. The dataset are compare against contain “Zachary ‘s karate club”, Given and Newman’s”, American college football games” and the network of all Digg user described in Tang et.al [5].

Table 1. Comparison of well known community detection on different dataset

Data set	Method	Number of communities	Modularity
Karate club	Original partition	2	0.36
	Louvian method	4	0.42
	Fast and greedy method	3	0.38
	Random walk method	5	0.35
Foot ball	Original partition	12	0.554
	Louvian method	10	0.604
	Fast and greedy method	5	0.544
	Random walk method	9	0.603
Digg	Louvian method	26646	0.178
	Fast and greedy method	37591	0.303
	Random walk method	78308	0.142

As given in table 1, methods are comparable to well known procedures when compared in terms of modularity. Expect the last data set, a large scale directed network of all users of Digg.com where the number of detected communities is higher than given by the Louvian or fast and greedy method[15].

5. Conclusion

This paper is presented the (user) ids in the social network until sufficient amount of data has been obtained. Then studies the community detection in social network especially facebook account (user)ids from mutual friends to cluster the group of university using relational database such as entities. First to extract from the database. In graph user(id)s is nodes and their relation of the friend is edges in social network. This paper is presented the community detection algorithm such as spectral modularity methods. Then compare the given dataset Vs different data set such as Zachary’s karate club using many other methods to get the modularity and cluster individual group. Another data set such as football club and Digg is presented . So this paper is expressed the comparison of clustering in given data set and other data set using community detection method (spectral clustering method) to produce small cluster that highly accurate or larger cluster are less accurate.

References

[1] Benjamin A. Miller, Nadya T, Bliss, Patrick J. Wolfe, and Michelle S.Beard," detection theory for graph", Lincoln Laboratory Journal, 2013.

[2] Ding Y. Community detection: Topological vs. topical, Journal of informatics.2011:5(4):498-514.

[3] Hongtao Lui Chen, Mao Lin, YU Wu, "Community detection based on topic distance in social tagging network", Indonesian Journal of Electrical Engineering, vol 12, No5, May,2014, pp 4038-4049.

[4] J.Shi and J.Malik, "Normalized Cuts and Image Segmentation,"IEEE transaction Analysis and Machine Intelligence, vol.70, no.6,2004,pp 066111-1-6.

[5] M.Garvan and M.E.J. Newman, "community structure in social and biological networks," Proceeding of the National Academy of Science, vol 99,pp.7821-7826, June 2002.

[6] Michel Plantiem Michel Crampes, "Survey on Social Community detection ",hal.archives-ouvertes.fr/ha;-00804234, Mar, 2013.

[7] Michelle Girvan and E.J Newman," A method for community detection in protein networks using spectral optimization", International Journal of Database Management Systems(IJDMS), Dec 2001.

[8] Michelle Girvan and M.E.J Newman, "Community structure in social and biological networks", Santa Fe Institute,1399 , Dec, 2001.

[9] Ming Cheungm" Connection discovery using big data of user shared images in social media", IEEE Transaction on multimedia, Feb, 2015.

[10] Santo Fortunato, "Community detection in graph", Complex Networks and System Lagrange Laboratory, ISI foundation, Viale S. Severo 65, 10133, Torino, 1-1tly, January 2010.

[11] Sminu Izudheen and Sheena Mathew, "A method for community detection in protein networks using spectral optimization", International Journal of Database Management Systems(IJDMS), Nov, 2011.

[12] S. Tang, N .Blenn, C. Doerr, and P. Van Mieghem, "Digging in the dig Social News Website", IEEE Transactions on Multimedia,vol.13, pp.1163-1175, oct-2011.

[13] W. W. Zachary, "An Information Flow Model for Conflict Model for and Fission in Small Group", Journal of Anthropological Research, vol.33, no.4,1977.

[14] William M. Compbell, Charlie K. Dagli, and Clifford J.Weinstein, "social network analysis with content and graphs", Lincoln Laboratory Journal, 2013.

[15] Norbert Blenn, Christian Doerr, Bas Van Kester, Piet Van Mieghem" Crawling and Detecting Community Structure in Online Social Networks using Local Information", 11th international IFIC TC 6 Networking conferences, Prague, Czech Republic, May, 2012.