Social Networking, 2016, 5, 11-15
Published Online January 2016 in SciRes. http://www.scirp.org/journal/sn
http://dx.doi.org/10.4236/sn.2016.51002

A Combination Approach to Community Detection in Social Networks by Utilizing Structural and Attribute Data

Nasif Muslim

Department of Computer Science and Engineering, United International University (UIU), Dhaka, Bangladesh
Email: nasif@cse.uiu.ac.bd

Received 23 October 2015; accepted 28 December 2015; published 31 December 2015

Abstract

Community detection is one of the important tasks of social network analysis. It has significant practical importance for achieving cost-effective solutions for problems in the area of search engine optimization, spam detection, viral marketing, counter-terrorism, epidemic modeling, etc. In recent years, there has been an exponential growth of online social platforms such as Twitter, Facebook, Google+, Pinterest and Tumblr, as people can easily connect to each other in the Internet era overcoming geographical barriers. This has brought about new forms of social interaction, dialogue, exchange and collaboration across diverse social networks of unprecedented scales. At the same time, it presents new challenges and demands more effective, as well as scalable, graph-mining techniques because the extraction of novel and useful knowledge from massive amount of graph data holds the key to the analysis of social networks in a much larger scale. In this research paper, the problem to find communities within social networks is considered. Existing community detection techniques utilize the topological structure of the social network, but a proper combination of the available attribute data, which represents the properties of the participants or actors, and the structure data of the social network graph is promising for the detection of more accurate and meaningful communities.

Keywords

Social Networks, Clustering, Community

1. Introduction

A social network graph (sometimes called an "activity graph") [1] contains nodes representing actors (generally

How to cite this paper: Muslim, N. (2016) A Combination Approach to Community Detection in Social Networks by Utilizing Structural and Attribute Data. *Social Networking*, **5**, 11-15. http://dx.doi.org/10.4236/sn.2016.51002

people or organizations), and edges representing relationships or communications between the actors. In graph representation, a community structure consists of several nodes which shows dense internal connection compared to the rest of the network. It is one of the well-known graph patterns observed in large social networks, and it reflects the fact the human society is organized into groups, families, filial groups, tribes, villages, friendship circles, professional circles, etc. The graph structure of social networks is influenced by the underlying human social behavior, and there is a strong connection between association and similarity in human society. Those with similar interests or of the same kind tend to form groups, which are shown by the age-old adage: "birds of a feather flock together." This principle, the homophily principle [2], implies the nodes that belong to a cluster or community in a social network are expected to have homogeneous characteristics. The identification of communities hidden within the structure of large complex network is a challenging problem which has attracted a considerable amount of interest. It has been widely studied in the literature, and there have been significant advancements with contributions from different fields, such as statistical physics. The fact that there is no agreed-upon definition for a community adds to the challenge. Different community detection methods are developed from various applications of specific needs which establish, explicitly or implicitly, its own definition of community. This means the definition of a community depends on the application domain and the properties of the graph under consideration.

Community detection is analogous to the clustering problem, and it can be viewed as a data mining analysis on graphs: an unsupervised classification of its nodes. However, classical data mining clustering algorithms [3] like K-means, Expectation-Maximization do not deal with graph data where the entities are nodes connected to each other through edges. On the other hand, the well-known graph clustering techniques, like clustering based on normalized cut [4], modularity [5] or structural density [6], utilize the relationships of the network to partition the graph into several densely connected components, but do not take the properties of the nodes into account.

Most community detection techniques focus mainly on the topological structure of the graph. However, online social networks are a rich source of information that can help profiling the users because they register their patterns of social interaction, dialogue, exchange and collaboration. Hence, in social network graph, nodes have attributes, and can be represented as attributed graphs, [7] [8] where an attribute vector associated with a node represent properties of the node like demographic features, personal preferences, and features characterizing the user's utilization of specific social network functions or applications. Since the nodes forming a community are expected to have homogeneous characteristics because of the homophily principle, it can be useful to take into account the node properties in the clustering process for increasing the accuracy of the partitions. Under this assumption, we can argue that, considering both structural data and attribute data simultaneously, instead of separately, can inform us the properties of topological structure and the impact and the role of the attributes on the topological structure which will be helpful for more meaningful and accurate community detection.

2. Background

A community is intuitively understood as a set of entities where each entity is closer, in the network sense, to the other entities within the community than to the entities outside it [9]. One of the widely-used definition of communities is based on the number of edges within a group (density) compared to the number of edges between different groups. A community is thought of as a group of nodes that has more and/or better-connected edges between its members (intra-cluster edges) than between its members and the remainder of the network (inter-clusters edges) [10].

Community detection methods can be broadly classified [11] into four categories: 1) node-centric community detection, 2) group-centric community detection, 3) network-centric community detection, and 4) hierarchy-centric community detection. Community detection based on node-centric criteria requires each node in a group to satisfy certain properties like mutuality, reachability, or degrees. On the other hand, community detection based on group-centric criteria means each group has to satisfy certain requirements like density, and the connections inside a group are considered as a whole. Network-centric community detection has to consider the connections of the whole network. It aims to partition the nodes into a number of disjoint sets. A group in this case is not defined independently. Typically, some quantitative criterion of the network partition is optimized. Hierarchy-centric community detection constructs a hierarchical structure of communities which facilitates the examination of communities at different granularity. There are mainly three types of hierarchical clustering: agglomerative, divisive, and structure search. Newman-Girvan algorithm [5], is a divisive hierarchical clustering algorithm which used Modularity to measure the quality of the community.

Community detection in social network graph can be viewed as an unsupervised classification of its nodes, *i.e.*, graphs clustering. While most graph clustering methods focus only on the topological structure of a graph, there are a few methods that aims to partition the graph depending solely on the similarity of the attributes, *i.e.*, the node properties. Some recent methods attempt to consider both the graph topological structure and the node properties. We describe below some of the representative methods which demonstrate that utilizing both the structural data and the attribute data simultaneously is promising.

Steinhaeuser *et al.* [12] utilize the node attributes for weighting the edges of the graph. They define a novel metric, node attribute similarity (NAS), to assign meaningful weights to the edges. The weight of an edge depends on the attribute-similarity of two nodes connected by it. Communities are identified using a thresholding method. Any pair of nodes whose normalized edge weight exceeds the threshold is placed in the same community. The proposed community detection method is relatively simpler and highly scalable which is able to produce extremely high modularity scores (values over 0.9), as shown by experimental evaluation on a large graph with 1.3 million nodes and 1.2 million edges that represents a real-world social network constructed from cellular phone records. It is reported that the attribute values alone contain some extremely valuable information about the community structure.

In [13], the authors study the evolution of Google+ social network by augmenting the social network structure with four users attributes (School, Major, Employer and City). Measurements of the network structure of the resultant Google+ Social-Attribute Network (SAN) is done by means of several canonical network metrics such as the reciprocity, density, clustering coefficient, and degree distribution [11] [14] [15]. Then, by generalizing the definition of the network metrics, a set of attribute metrics such as attribute density, attribute distance and diameter, attribute clustering coefficient is obtained. By measuring the network structure as well as the attribute structure by means of the defined metrics, it is possible to gain a more fine-grained understanding of a social network. In particular, how the attributes influence the social structure of the Google+ SAN is studied in [13]. It is observed that the nodes sharing common attributes are likely to have higher reciprocity, and that some attributes have much stronger influence than others.

Elhadi *et al.* [16] propose a novel clustering method, termed Selection method, which does not use a modified objective function to combine the structure data, and the attribute data, rather makes the choice to utilize either the structure data, or the attribute data depending on the type of graph and the level of information in the attributed graph.

Gunnemann *et al.* [17] propose a combined clustering model, and the algorithm GAMER (Graph & Attribute Miner) for detection of densely connected subgraph (clusters) that exhibit attribute similarity in subsets of the dimensions (subspace clusters). It is empirically shown for synthetic and real world datasets that clustering quality is improved by simultaneous use of both graph data and attributes data.

3. Community Detection Schemes

We have designed four schemes, which use structural and attribute similarity of the nodes for detection of community structure. Each of the schemes will deliver separate outputs. To find the set of community which represent the real community structure of the social network, outputs of different schemes are combined by using Consensus clustering [18]. Brief descriptions of the schemes are given below.

3.1. First Scheme

In this scheme, the structural similarity between the nodes is utilized to detect community. To measure the structural similarity between the nodes, two structural metrics: node degree and clustering coefficient are used. For each node, degree and clustering coefficient value are calculated individually. These metric values are used as co-ordinate values for a particular node in 2-dimensional space (xi, xj) = (node degree, clustering coefficient). Now, K-means clustering is applied to find the set of communities. However, it is not always clear what is the exact value of K. To find out the value of K, community strength is measured using Modularity after each iteration. The iteration process continues until Modularity value does not improve. The final output of this scheme will be a set of communities.

3.2. Second Scheme

In this scheme, the attribute similarity between the nodes is exploited to detect community. To measure the attribute similarity between the nodes, the social network graph will be augmented with user attributes [13]. For each node i with attribute node a, an undirected link between i and a in the social network graph is created. For each attribute node, two attribute metrics: attribute degree and attribute clustering coefficient values are calculated. These metric values are used as co-ordinate values for a particular node in 2-dimensional space (xi, xj) = (attribute degree, attribute clustering coefficient). Now, K-means clustering is applied to find the set of communities. To find the precise value of K, community strength is measured for each community using Modularity. The iteration process continues until Modularity value does not improve. A set of communities are generated as the output of this scheme.

3.3. Third Scheme

In this scheme, the structural similarity and the attribute similarity between the nodes are considered simultaneously to detect community. For a particular node, co-ordinate value in 1-dimensional space is linear combination of structural metric and attribute metric: w1 × structural metric + w2 × attribute metric. Weight w1 corresponds to the structural metric and Weight w2 corresponds to the attribute metric.

Large social network graphs show some distinct patterns and properties which are not visible in smaller networks. One of the most well-known characteristics of large social network graph is scale-free distribution. It means that the degree distribution of a social network follows a power-law. Large social network also shows strong community effect. It can be hypothesized that if distribution is a good fit to the power-law distribution then more communities can be found. If node degree distribution is a good fit to the power-law distribution than attribute degree distribution then the number of communities detected using structural metrics will be higher. The weight is calculated based on the fitness to the power-law distribution.

To measure the fitness to the power-law, degree distribution is converted into log-log scale which will make the curve of the power law distribution into a straight line. Linear Regression is used to calculate the equation of the best-fitting straight line. Total residual value is measured which will be the value of the weight. Residual = Observed value – Predicted value.

Now, agglomerative hierarchical clustering is used to find the set of communities. In case of agglomerative hierarchical clustering, a community pair is merged and if doing so results in the increase of overall Modularity, then the process of merging continues until no merge can be found to improve the Modularity score. The final output of this scheme will be a set of communities.

3.4. Fourth Scheme

It is another scheme, where to detect the community structure; attribute similarity between the nodes is exploited. However, some attributes or dimensions are not relevant for all communities. To solve this problem, for each attribute node, value of the attribute coefficient in the augmented social network graph [13], is calculated. By applying a threshold value, most relevant attributes of the social network is selected. Now, subspace clustering is considered. Two fold clustering as used in GAMER algorithm [17] will be employed to find the set of communities. Alternatively, a variant of the GAMER algorithm is going to be considered where Modularity score is used instead of density score.

4. Evaluation

The community detection schemes will be quantitatively evaluated on datasets representing large real-world social networks. For this reason, we are going to need a collection of large real-world social network data in graph representation, where group memberships (ground-truth information) of nodes are explicitly stated, and the properties or attributes of the nodes are known.

A large collection of benchmark datasets, with ground-truth information, and also with both structural and attribute data do not yet exist, to the best of our knowledge. To solve this problem, a collection of benchmark datasets with both structural and attribute data will be gathered by crawling [14] online social networks using the public web interface provided by the sites. Also, we would utilize existing datasets of real-world social networks gathered by researchers, like the dataset used in [7].

5. Conclusion

The absence of a universal definition of a community has implications for community structure detection and evaluation. It is hard to evaluate the identified communities and obtain the comparison of different community detection methods. In this paper, four schemes are designed which utilize structural and attribute data for community detection. These schemes will improve the quality of community detection.

References

[1] Cook, D.J. and Holder, L.B. (2007) Mining Graph Data. John Wiley & Sons, Inc., Hoboken.

[2] McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001) Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, **27**, 415-444. http://dx.doi.org/10.1146/annurev.soc.27.1.415

[3] Kaufman, L. and Rousseeuw, P.J. (2005) Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley, Hoboken. http://dx.doi.org/10.1002/9780470316801.ch1

[4] Shi, J. and Malik, J. (2000) Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**, 888-905. http://dx.doi.org/10.1109/34.868688

[5] Newman, M. and Girvan, M. (2004) Finding and Evaluating Community Structure in Networks. *Physical Review E*, **69**, 1-16. http://dx.doi.org/10.1103/PhysRevE.69.026113

[6] Xu, X., Yuruk, N., Feng, Z. and Schweiger, T.A.J. (2007) Scan: A Structural Clustering Algorithm for Networks. *International Conference on Knowledge Discovery and Data Mining* (*KDD*'07), San Jose, 824-833. http://dx.doi.org/10.1145/1281192.1281280

[7] Combe, D., Largeron, C., Egyed-Zsigmond, E. and Gery, M. (2012) Combining Relations and Text in Scientific Network Clustering. 2012 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Istanbul, 26-29 August 2012, 1248-1253. http://dx.doi.org/10.1109/ASONAM.2012.215

[8] Zhou, Y., Cheng, H. and Yu, J. (2009) Graph Clustering Based on Structural/Attribute Similarities. *Proceedings of the VLDB Endowment*, 2, 718-729. http://dx.doi.org/10.14778/1687627.1687709

[9] Coscia, M., Giannotti, F. and Pedreschi, D. (2011) A Classification for Community Discovery Methods in Complex Networks. *Statistical Analysis and Data Mining Journal*. http://dx.doi.org/10.1002/sam.10133

[10] Leskovec, J., Lang, K.J., Dasgupta, A. and Mahoney, M.W. (2008) Statistical Properties of Community Structure in Large Social and Information Networks. *Proceedings of* 17*th International Conference on World Wide Web*, *WWW*'08, 695-704. http://dx.doi.org/10.1145/1367497.1367591

[11] Aggarwal, C.C. and Wang, H.X. (2010) Managing and Mining Graph Data. Springer, New York.

[12] Steinhaeuser, K. and Chawla, N. (2008) Community Detection in a Large Real-World Social Network. *Social Computing*, *Behavioral Modeling*, *and Prediction*, 168-175. http://dx.doi.org/10.1007/978-0-387-77672-9_19

[13] Gong, N.Z.Q., Xu, W.C., Huang, L., Mittal, P., Stefanov, E., Sekar, V. and Song, D. (2012) Evolution of Social-Attribute Networks: Measurements, Modeling, and Implications Using Google+. *Internet Measurement Conference*, 131-144. http://dx.doi.org/10.1145/2398776.2398792

[14] Mislove, A., Marcon, M., Gummadi, K.P. and Bhattacharjee, B. (2007) Measurement and Analysis of Online Social Networks. *Proceedings of the* 7*th ACM SIGCOMM Conference on Internet Measurement*, 29-42. http://dx.doi.org/10.1145/1298306.1298311

[15] Kossinets, G. and Watts, D. (2006) Empirical Analysis of an Evolving Social Network. *Science*, **311**, 88-90. http://dx.doi.org/10.1126/science.1116869

[16] Elhadi, H. and Agam, G. (2013) Structure and Attributes Community Detection: Comparative Analysis of Composite, Ensemble and Selection Methods. *The* 7*th SNA-KDD Workshop*'13 (*SNA-KDD*'13).

[17] Gunnemann, S., Farber, I., Boden, B. and Seidl, T. (2010) Subspace Clustering Meets Dense Subgraph Mining: A Synthesis of Two Paradigms. *Proceedings of the IEEE International Conference on Data Mining*, Sydney, 13-17 December 2010, 845-850. http://dx.doi.org/10.1109/ICDM.2010.95

[18] Lancichinetti, A. and Fortunato, S. (2012) Consensus Clustering in Complex Networks. *Scientific Reports*, **2**, 336. http://dx.doi.org/10.1038/srep00336