International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.03, March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Community Detection in a Network with Categorical Relationships through Graph Models

Jaishri Gothania[1], Dr. Shashi Kant Rathore[2]

[1]*Ph.D Scholar, Department of Computer Science and Engineering,*
Career Point University, Kota, Rajasthan, (India)
[2]*Assistant Professor, Department of Computer Science and Engineering,*
Career Point University, Kota, Rajasthan, (India)

ABSTRACT

The study of networks is an active area of research due to its capability of modeling many real-world complex systems. The research for community detection in social networks aims at analyzing networks to extract useful information from it. Data of the nodes in network can be numeric, categorical or hybrid. The relationships among data points are generally limited to either binary or fuzzy. Conventional analysis through clustering decides this relationship based on distance or any other similarity measure between two data points and detect them as a same community and cluster them together if found similar. With time, a new kind of relationship called categorical relationship was observed between data points. This paper focuses on exploring works that handle community detection in a network having categorical relationships and related problems of community discovery or data mining.

Keywords: Community detection, Fuzzy relationships, Agonistic, etc.

I.INTRODUCTION

Community detection is an unsupervised machine learning approach aiming at categorizing similar data points among a set of data and grouping them together in a bundle, specifically called a cluster. It is of wide importance in areas of pattern analysis, statistical data analysis, image analysis, information retrieval, bioinformatics etc. Effective Community detection takes into account two aspects; the nature of the data points to be clustered and the relationships between these data points. The data can be numeric, categorical or mixed. The relationships between the data points are observed to be binary, fuzzy or the newly observed categorical. All the previous research works were limited to picking a numeric value called distance or any other similarity measure between data points to cluster them accordingly. This similarity can be perfectly deduced if we find what relationship two data points are holding for each other. Whereas binary relationship categorized the data points as similar or dissimilar with respect to any similarity measure used and clustering them accordingly, fuzzy relationship pointed out a percentage of similarity or dissimilarity between data points with the less similar ones more probable to lie in the same cluster. Both the binary and the fuzzy relationships involved computation on the actual representations of the objects.

Our goal is to study the factors which sway relationships and to study the categorical and correlations between data points or data objects. A fundamental problem related to these social networks is the discovery of "clusters" or "communities". A community is a subset of data objects in a graph or a cluster of densely connected data nodes of a network.

II.COMMUNITY DETECTION IN SOCIAL NETWORKS

Now these days social networking(SN) gain popularity due to its ease of use ,we all know that social network facilitates users to interact, communicate and share on World Wide Web. Recently social networks become vogue due to its popularity, commerciality and trendiness.

A brief overview of some of the previous work in the area of community detection to give the reader a sense of current methods.

A. Disjoint Community Detection

As pointed out by Kelly et al[1] the majority of current methods work treat the problem of locating communities as a hierarchical partitioning problem. According to this approach, the community structure of a network is assumed to be hierarchical; individuals form disjoint groups which become subgroups of larger groups until one group, comprising the whole society, is formed. Such methods for a tree of subgroup relations called a dendrogram. A dendrogram allows the community structure of a network to be at various resolutions.

B. Overlapping Communities

Kelly et al[1] also observed that while hierarchical grouping is valid for some types of networks, e.g., organizational networks or taxonomies, intuition and experience suggest that social networks contain pairs of communities that overlap while not containing each other as a sub-community. Consider an individual in a social network representing "friendship".He or she may have friendship relations across many different social circles, such as those formed in the workplace, by a family unit, by a religious group, or by social clubs. In this case, assuming social structure of the network to be hierarchical might lead to missing important information about members' attachment to the numerous social circles with which they concurrently interact.

III.COMMUNITY DISCOVERY METHODS IN COMPLEX NETWORKS

In recent years detection of communities in complex network has attracted a lot of attention. To discover such communities researchers are putting their effort by applying different methodologies.

A. Density Based Community Detection

As pointed out by Coscia et al[2] the community is defined as a group in which there are many edges between vertices, but between groups there are fewer edges. The aim of a community detection algorithm is to divide the vertices of a network into some number k of groups, while maximizing the number of edges inside these groups and minimizing the number of edges that run between vertices in different groups. In density based community detection they consider the connection between two vertices a particular kind of action. Hence, if they group entities by maximizing their common actions, we also group them by maximizing the edges inside the

community. Community discovery is exactly the same if the edge creation is the only action recorded in the network representation.

B. Vertex similarity-based Community Detection

As pointed out by Fortunato [3], it is natural to assume that communities are groups of vertices that are similar to each other. One can compute the similarity between each pair of vertices with respect to some reference property, local or global, irrespectively of whether or not they are connected by an edge. Each vertex ends up in the cluster whose vertices are the most similar to it. By considering an evolving setting in our problem representation, together with the presence or absence of a particular property (i.e. a label of the vertex), we can model the similarity measures as the similarity of the set of actions.

C. Action-based Community Detection

Entities can be grouped by the set of actions they perform inside the network. For example, in [4] a multi-mode network is considered in which users are connected to queries and ads. Two users are seen as being part of the same community if they are connected to the same queries (i.e. they perform the same actions) even if they are not directly linked to each other. The discovery of communities based on this method can be performed considering or not the presence of a direct link between entities.

D. Influence Propagation based Community Detection

In [5] the concept of a "tribe" has been introduced, a tribe is defined as a set of entities that are influenced by the same leaders. A node is a leader if it has performed an action and, within a chosen time bound after this action, a sufficient number of other users have performed the same action. The role of social ties in this influence spread is considered. Thus, according to our definition, the set of users that frequently perform the same actions due to the influence of their leaders are considered as being a community.

IV.APPLICATIONS OF COMMUNITY DETECTION

Community detection involves the collection of information from a (usually fairly large) number "unit". These units may be people, or organizations, or towns, or families, or departments, etc; the information collected may be of any kind - eg financial information or opinions in the case of surveys of people, or information about numbers of employees and organizational structures in the case of a community detection in an organizations. In sociology ,biology and computer science disciplines where systems and networks are often represented as graphs, there discovering communities have great importance.

Structure: Community discovery is to define the community exactly as a very precise and almost immutable structure of edges. Often these structures are defined as a combination of smaller networks[2].

Closeness: A community can also be defined as a group of entities that can reach each of its own community companions with very few hops on the edges of the graph, while the entities outside the community are significantly farther apart[2].

Bridge Detection: The community discovery approaches based on the concept that communities are dense parts of the graph among which there are very few edges that can break the network down into pieces if they are removed. These edges are "bridges" and the components of the network resulting from their removal are the desired communities[2].

Link Clustering: Instead of clustering the nodes of a network, it is the relation that belongs to a community, not the node. There-fore they cluster the edges of the network and thus the nodes belong to the set of communities of their edges[2].

Diffusion: Communities are groups of nodes that can be influenced by the diffusion of a certain property or information inside the network. In addition, the community definition can be narrowed down to the groups that are only influenced by the very same set of diffusion sources[2].

Internal Density: In this we can discover community by directly detecting the denser areas of the network[2].

No Definition: There are a number of community discovery frameworks which do not have a basic definition of the characteristic of the community they want to explore. Instead they define various operations and algorithms to combine the results of various community discovery approaches and then use the target method community definition for their results[2].

Feature Distance: A community is composed of entities which present everywhere and share a very precise set of features, with similar values (i.e. defining a distance measure on their features, the entities are all close to each other). A common feature can be an edge or any attribute linked to the entity (in our problem definition: the action). Usually, these approaches propose this community definition in order to apply classical data mining clustering techniques, such as the Minimum Description Length principle [6, 7].

V.CORRELATION CLUSTERING

Bansal et al proposed Correlation Clustering in [8] which was successful enough to eradicate all the issues encountered in the traditional clustering algorithms. Instead of some distance/similarity measure, it uses a similarity relation to consider the objects similar if they hold this relation and dissimilar if not. The clustering methodology required edge-labeled graphs with edges signed as positive or negative. Clustering depends on edge labels and can have any number of clusters. Clustering is based on the notion of maximizing agreements and minimizing disagreements. Here, agreement means the sum of number of positively signed edges inside clusters and number of negatively signed edges between clusters. Disagreement, therefore, means the sum of number of negatively signed edges inside clusters and number of positively signed edges between clusters. Mathematically expressed, for a graph G = (V ; E), where V is the set of objects to be clustered and E edges denoting relationships between V, a function $s: E \rightarrow \{+, -\}$ is defined to assign a sign for each edge, with sign + denoting the similarity and - denoting dissimilarity. Therefore, for correlation clustering, a signed graph as (G; s) is used. Any similarity distance or real distance is used for the signing of edges.

A. A note on Conventional Clustering

Therefore, for correlation clustering, a signed graph as (G; s) is used. Any similarity distance or real distance is used for the signing of edges. The data in the form of audio files, video files, texts, documents, records etc is continuously increasing giving rise to the need of determining patterns of related data from bulk of data for future uses like storage, searching, sorting, updating etc. Clustering, an exploratory task of data mining, aims to analyze and group related data in clusters. Earlier, data used to be considered only numeric. With time, data was further classified to be of categorical nature or a mix of numeric and categorical. Categorical data involves grouping of data in terms of the attributes the data holds, for example, age or blood group of a person, state of a country, type of rock etc. Mixed data contains both numeric and categorical attributes of data. There have been proposed umpteen numbers of clustering algorithms for numeric, categorical and mixed data. All the clustering approaches can be further classified into:

***Hierarchical Clustering Approach*:** Include the clustering algorithms that seek to build a hierarchy of clusters. The clustering technique can be further divided into

- *Agglomerative or "Bottom-Up" approach*: The clustering algorithm involving an agglomerative approach for clustering starts with every data point in its own cluster with merging between clusters on their way up to the hierarchy.
- *Divisive or "Top-Down" approach:* The algorithm with a divisive approach begin with data points in their own clusters and merging with the others on the way down the hierarchy.

Partitioning approach: The data points are decomposed or partitioned into disjoint set of clusters with subsequent iterations of the algorithm. The algorithms run in an iterative fashion until convergence or till all the data points are not clustered.

Density-Based Approaches: It is a subpart of the partitioning approaches with areas of high density denoted as clusters and the remaining data points as outliers or border points.

Grid-based approaches: This is again a partitioning approach which partitions the attribute space covered by data objects into segments/cells/regions. Thus it utilizes the topology of the data space. This space-partitioning is then used for data partitioning according to membership in regions.

Machine Learning approaches: These approaches use a sample of pre-classified data to train themselves. A relation between the attributes and categories is established which is used to categorize actual data.

High-Dimensional Data clustering approaches: With high dimensions, clustering becomes a tedious task, mostly because of the lack of separation of data points at such high dimensions, also referred to as the "Dimensionality curse".

B. Edge Labeled Graphs

An edge-labeled graph G = (V; E; L; Lo; f), where the set of vertices V corresponds to the objects to be clustered, the set of edges E comprises all unordered pairs within V having some relation (i.e., whose relation is represented by a label other than the Lo label), and the function f assigns to each edge in E a label from L. Here, the label L

shows relationship between two vertices, represented through edges. More than two vertices having same relationships can also be joined through similarly labeled edges joining these vertices.

C. Agnoistic learning

The correlation clustering problem by Bansal et al in [8] represents function f in a given limited hypothesis language. This can be termed as agonistic learning [9,10]. When the clustering is not perfect, that is all the positive edges cannot be put in a single cluster leaving behind the negative edges, a trivial solution is agreeing with half of the edge labels for clustering. For example, in case of more positive edges and less negative edges, all the vertices could be put in a single big cluster and if not, then each vertex would lie in a different cluster. The observation results agreeing with atleast half of the edge labels can correspond to an error atmost ½ using either all positive or all negative hypotheses.

D. Chromatic Correlation Clustering

Inspired from the Correlation Clustering by Bansal et al [8], Bonchi et al[11,12] extended the work to assigning colors to edges instead of signs. These colors acted as labels to the edges. Similarly colored edges showed similar relations between the adjoining vertices and hence were expected to fall in the same cluster. An objective function was introduced for ensuring that the edges within a cluster are as much as possible, of same color. The contributions by Bonchi et al are briefly discussed below

Chromatic Balls Algorithm: A randomized algorithm for solving the chromatic clustering problem and providing approximation guarantee till the maximum degree of the graph.

Lazy Chromatic Balls Algorithm: One of the two alternative algorithms for overcoming the issues of Chromatic Balls; optimizes the proposed objective function iteratively.

Multi Chromatic Balls: For relations between objects denoted by a single label, Chromatic Correlation Clustering problem is a novel concept. For relations denoted by a set of labels, a generalized version of the Chromatic Correlation Clustering problem, Multi Chromatic Clustering problem has been defined and as a solution, Multi Chromatic Balls algorithm is proposed.

Informed Chromatic Balls: The issues in traditional clustering algorithm of not being able to cluster data objects having categorical relations called for clustering using edge labeled graphs, capable of representing both independent and co-independent relations. Correlation Clustering, followed by Chromatic Correlation Clustering proved as successful solutions to the problem of handling categorical relations. An Informed Chromatic Balls algorithm gives its contribution in the direction of revisiting the work of Bonchi et al [9,10]. An Informed Chromatic Balls algorithm is presented to increase the probability of better solution of the algorithm keeping its advantage of speed retained.

VI.CONCLUSION

Community detection is a relevant problem in current scenario of social networking among people. The social media provides many types of relationships to exist among people or other entities of the network. These are best represented through labeled graphs. This converts community discovery into a component discovery

problem for graphs based on labels. Hierarchical and other clustering methods are not suitable for this. A recent technique called chromatic correlation clustering and its variants are more effective to partition the network graphs into disjoint components corresponding to communities. It can be converted to overlapping communities also. This makes chromatic correlation clustering a basic problem to define community discovery challenge.

REFERENCES

[1] Stephen Kelley, Mark Gold berg ,Malik Magdon-Ismail, Konstantin Mertsalov, and Al Wallace. Defining and Discovering Communities in Social Networks. Handbook of Optimization in Complex Network, pages 139-168,2012.

[2] Michele Coscia, Fosca Giannotti, Dino Pedreschi.A classification of Community Discovery Methods in Complex Networks. Statistical Analysis and Data Mining journal, Special issue:Network.Volume 4, Issue 5, pages 512-546, June 18,2012.

[3] S. Fortunato, "Community detection in graphs," Physics Re-ports, vol. 486, no. 3-5, pp. 75 – 174, 2010.

[4] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, "Community evo-lution in dynamic multi-mode networks," in KDD '08: Pro-ceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, (New York, NY, USA), pp. 677–685, ACM, 2008.

[5] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Discovering leaders from community actions," in CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management, (New York, NY, USA), pp. 499–508, ACM, 2008.

[6] J. Rissanen, "Modelling by the shortest data description," Automatica, vol. 14, pp. 465–471, 1978.

[7] P. D. Grnwald, The Minimum Description Length Principle, vol. 1 of MIT Press Books. The MIT Press, 2007.

[8] N. Bansal, A. Blum and S. Chawla, "Correlation Clustering", Machine Learning, Vol. 56, pp. 89-113,2004.

[9] M. Kearns, "Efficient noise-tolerant learning from statistical queries", Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, pp. 392–401, 1993.

[10] M. J. Kearns, R. E. Schapire, and L. M. Sellie, "Toward efficient agnostic learning", Machine Learning, Vol. 17, No. 2-3, pp.115–142, 1994.

[11] F. Bonchi, A. Gionis, F. Gullo and A. Ukkonen , "Chromatic Correlation Clustering", Proceedings Of The 18th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining (KDD '12), pp. 1321-1329, 2012.

[12] F. Bonchi, A. Gionis, F. Gullo, Charalampos, E. Tsourakakis and A. Ukkonen, "Chromatic Correlation Clustering", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 9, Issue .4, No. 34, 2015.