Multi-source Fusion Approaches for Efficient Community Detection in Social Networks

Mehrafarin Adami, Mohammad Nadimi
Faculty of Computer Engineering
Islamic Azad University, Najafabad branch
Isfahan, Iran
mehrafarina@sco.iaun.ac.ir, nadimi@iaun.ac.ir

Reza Ghaemi
Computer Engineering Department
Islamic Azad University, Quchan branch
Quchan, Iran
rezaghaemi@iauq.ac.ir

Abstract— Social network community detection is an important issue for means of effective advertisements, accurate recommender systems and tracking changes. Since social networks consist of several data sources, they can be informative that yield self-descriptive communities. Accuracy, efficiency and scalability of community detection can be enhanced by selection of these data sources and determination of the appropriate combining approach. Traditional approaches express concern over community detection algorithms. To prevent semantic confliction and managing multiple data sources, a process is necessary to bring all necessary steps together. In this paper this process is clarified. Moreover, several effective approaches for combining the data sources are compared.

Keywords: Social networks, Community detection process, Multi-source fusion approaches

I. INTRODUCTION

Digitalizing official and administrative affairs and shortened distance between people has attracted people to virtual social networks. Users can do their daily chores and communicate there. In this process different relations among users are formed. It is so convenient these days that if users find a new virtual place with more services, an ideal environment, they will join it exponentially and start discussions, share photos and other media to show recent or important events in real lives, i.e. there are several types of behaviors and different characteristics in these virtual communities just like what we see in real societies, therefore it is important to analyze the communities in a social network.

Through the community detection process (CDP), it is possible to categorize common relations between users and analyze each related part of a network, community, separately in more detail. Changes can also be shown by tracking communities.

In spite of the great number of valuable methods in the literature, there are some limitations; e.g. it is not clear which algorithm to use for CDP that fits network features. A large number of methods are introduced but they have been tested on datasets with few numbers of users, and just one relation type between them. Therefore it cannot be adapted to community features such as community structures, existence of hierarchy or overlapping. Almost all of these algorithms need input parameters such as number of communities, minimum and maximum number of users or number of common neighbors in a community and the shape of communities which have direct influence on the result. The other limitation relates to the nature of methods. For Example, modularity-based algorithms cannot handle free-scale datasets due to expensive computations. Other critical limitations such as the need for the determination of cluster centroids, effect of outliers and so forth. In addition to these technical limitations there is another type of challenges related to application-based research, data collection and pre-processing of networks. Community ddetecting methods are applied on artificial networks. Analyzers consider assumptions and default values to start CDP. Therefore the real data change to a filtered, branched limited graph which is usually able to show just one kind of relation. These restrictions indicate that it is impossible to analyze all relations among all users of virtual communities. Consequently, it is irrational to expect a reality in which all these relations can be categorized and studied.

Although there are several contributions about community detection in social networks, the process of community detection is not cleared yet, therefore in this paper, first the process is investigated and its steps are explained, where both application and technique developments are sequentially considered. This process is not dependent on using a special or predefined dataset(s) or method(s).Second several fusion approaches, which can manage different data sources and the diversity of relations between users, are discussed. Each approach can be used based on the number of sources, available data types and the importance of maintaining each source semantic. The rest of the paper is organized as follows: related work on technique- and application- developments and hybrid methods are described in section 2, CDP is investigated in detail in section 3, multi-source fusion approaches are described in section 4, and finally conclusion is in section 5.

II. RELATED WORK

Community detection approaches, summarized in two aspects that go back to the technical and application sides of the issue as discussed in the following two subsections. Recently multiple sources are used to improve results of CDP that will be discussed in 2.3.

A. Technique Development

Since community detection methods were first a graph partitioning problem [1] proposed in the early 1970's, one the most effective algorithms using central measurements was proposed by Girvan and Newman [2]. Later several community detection methods have been designed that fall into 5 categories: methods based on similarity parameters, clustering, spectral methods, probabilistic models and graph partitioning. Table 1 summarized these methods and critical limitations.

Technical algorithms have been tested by benchmarked or synthesized datasets with known input parameters such as: number of communities, the existence or absence of hierarchical relationships and existence or absence of soft membership among users, so that if some of these parameters were not known, lack of them would not pose a problem for research, because algorithm designers focus on methods not on the hidden properties of datasets or thoroughness of their side information. Additionally, these methods are useful for handling only one category of data sources.

TABLE1.MAIN SOURCES USED IN DIFFERENT COMMUNITY DETECTION METHODS (K: NUMBER OF COMMUNITIES)

METHOD-NAME	Some algorithms	Main Source	Important limitations
Methods by use of similarity parameters	[3, 4]	Links (explicit relations)	Always high value doesn't mean better communities.[5]
Clustering	[6, 7]	Links or profile features	Centroid specification and outliers decrease the quality- K as input
Spectral methods (using eigenvectors)	[8, 9]	links	High computation
Probabilistic models	[10, 11]	Both Link, content	Depends on drichlet allocation thresholds /High computation
Graph partitioning	[12]	links	Hard membership- k as input/Bisection limitations

B. Application Development

This type of work is related to instances where the social network and its users relations is important, like comparison of well-known social networks [13], the interdependence of Facebook users [14], network of scientific authors [15] and weblog content as a social networks [16].

A social network analyzer, based on some background knowledge about these kinds of networks, starts to develop a graph which represents special predefined feature not available in all cases. The emphasis is on making a comprehensible and more reliable dataset rather than benchmarked datasets and finally, they choose a technical method to detect hidden communities. These methods are described in Table 1 and are applicable to linked users, where users are represented by nodes, and relations among them are represented by links (links can be made on the basis of other sources like content similarities, but here it means explicit relations like friend lists, etc.), on the other hand probabilistic models like Bayesian methods and clustering approaches are applicable to other sources like content- or user-based features like age and sex. Being familiar to nature of technical methods is essential for choosing the right algorithm, Santo Fortunato believed, "people rely blindly on some algorithms instead of others for reasons that have nothing to do with the actual performance of the algorithms, like popularity".[17]

C. Hybrid Methods

Both technique- and application- developments have their own limitations. Therefore there has been a growing interest in community detection approaches for the combination of methods or for using multiple sources together, which yield improved results, overcome individual algorithm drawbacks and help to decrease computational overhead [11, 18-21] . Some of hybrid methods are described following.

To overcome individual algorithm drawbacks, HCDF is presented by Henderson et al., which uses Latent Dirichlet Allocation on Graphs as the core Bayesian method for community detection. A key aspect of HCDF is its effectiveness on incorporating hints from a number of other community detection algorithms [11]. The other research uses attribute and relationship sources for community detection. Attribute data, such as demographic information and relationship data can yield more accurate results than classical algorithms that either use only attributes or only relationships. by using these two data sources, JointClust algorithm discovers meaningful and accurate clustering [21]. Furthermore, ccommunities can comprise users with common topics. Since usually document collection is a voluminous task, Li *et.al* had found a solution to make a scalable community detection method by utilizing text contents as well as relations. [18]

III. COMMUNITY DETECTION PROCESS

Traditional methods are applied to nodes and edges without any side information, without clarifying node`s real life characteristics or different types of their associations. This gray presentation of networks seems to be far from real communities. In this section whatever necessary for community detection is investigated as a process Fig. 1 shows the main steps of community detection process.

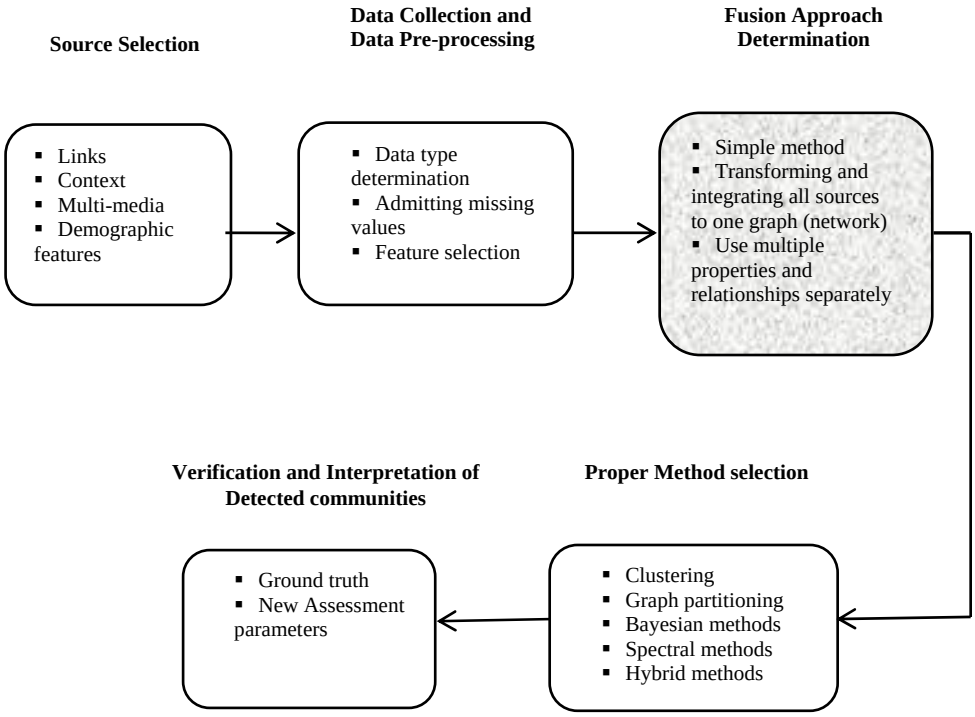

Figure 1. Community Detection Process (CDP)

A. Source Selection

Data source selection is the first important step, because the other steps are affected by this step. Sources are contents (e.g. opinions, comments and weblogs), links (co-author, friend list or reference to other web pages As mentioned before in this paper links are explicit connections between users), attributes and features (such as profile information) and other media sources such as tagged photos or shared videos. There is not a general method to handle all of them together yet. Depending on the importance of accuracy and reality level of the social network, it can include each or all of these data sources.

B. Data Collection and Data Pre-processing

Using multiple sources is like the application of a vertical partitioned database. So it is necessary to determine the final variable types of data and describe the data prior to store them.

Raw data can decrease the results quality, because there are incomplete, inconsistent and noisy data in the dataset, a pre-processing map is necessary to make data clean and consistent. This map is drawn on the basis of the accepted quality and the availability of background knowledge, so it is a rather arbitrary step. The important techniques are discussed in the following.

- **Fill the null values**: by using link prediction methods, missing relations can be detected [22]. If nodes have properties and there are some null fields, filling them manually by background knowledge or by means of that property as a default value. Another strategy is removing nulls by the most probable value.[23]
- **Select distinctive features**: If the total number of variables is huge, feature selection can be helpful to reduce data and decreasing data dimensions, e.g. since there are a lot of words in each document, the selection of keywords or highly frequent words is vital. Feature selection is applicable to user properties too, if the features with no new information can be eliminated. Feature selection decreases runtime of the algorithm and facilitates results interpretation.[24]

C. Fusion Approach Determination

Different methods are recently proposed to use multiple sources to overcome some limitations like scalability and interpretability. Some instances are combination of content and link sources, using media information and combination of user's attributes and links [19, 21, 25-27]. Although using multiple sources can produce better results in terms of quality and inclusiveness, but it has its own complexity: How much is the impact of each source? How to aggregate or mix these sources? How to detect effective variables of each source? In [28] the well-known link and content methods are reviewed and the effect of each source is analyzed. Finally, a mixture model is presented, though it is unable to improve single source results. After declaring what is significant to the aim of CDP, it is important to determine the right multi-source fusion approach, in order to facilitate the selection of the right algorithm. These approaches will be referred to in section 4.

D. Proper Method Selection

Based on the fusion approach and data types of the network, an algorithm or a framework needs to be designed or selected. There are several useful methods for different kinds of graphs like (un)weighted graphs or (un)directed graphs, Also determination of existence of soft membership or hierarchical relations that may be hidden in communities is important. The method can be as simple as explained methods in Table 1 or it can be as complex as hybrid types. The nature of algorithms and their input parameters, limitations and other factors like their cost or complexity are reviewed in several surveys such as [17, 26, 29, 30].

E. Verification and Interpretation of Detected Communities

The accuracy and reality of results (communities) need to be evaluated. If there is ground truth to compare the results with, and if there is no parameter which is compatible to all sources and semantics to choose from, use a new assessment parameter. By using multisource fusion approach, traditional parameters may produce low value which does not mean the low accuracy of the results, because traditional parameters are based on one single source and one single concept. For example, the quality parameter in Divisive algorithms is modularity, which includes links. But now the informative network is divided into communities based on multiple sources which have their own effects on CDP.

IV. MULTI-SOURCE FUSION APPROACHES

A. Simple Method

This method applies to networks when edges represents one kind of relationships, therefore a consistent network made by a single type of nodes or edges. For example, in case of the scientific authors network, all of the co-authors, library assistants and supervisors are of the same type and different relationships between them are considered to be a simple edge. Methods described in table 1 can be used as the right algorithm for the CDP.

B. Transforming and integrating all sources into one graph (network)

This approach applies to sources that are made homogenous. Different node properties and different relations can be managed by this approach, but these sources will not be recognizable any more after being transformed to one data type or after their integration as a general property. One strategy for this approach is transforming all of the source variables into one type and simply applying a clustering method to detect communities. The next strategy is assigning weights to edges on the basis of the number of their common activities or features. The more similar the activities the higher weight will be assigned to the nodes or edges. For the scientific authors network, if two authors have similar field of research, they are in a stronger relationship, or if they use similar keywords, their specialties are of the same filed thus it will put more weight on their links. Moser et al., adopt an informative graph for the representation of both entity-entity similarity and entity-relationship matrices. Attributes and relationships are transformed into an undirected graph, in which each node represents a data object and each edge

a relationship between the data objects associated with the corresponding nodes. The attribute values of a data object are attached to the corresponding node as a weight vector. [21]

Transforming and integrating all sources is not reliable when more than two sources need to be gathered. Adjusting weights to edges needs special consideration: what is a proper criterion for assigning weights to each source or how to keep sources effective and meaningful, and finally how to quantify them into numbers with no semantic confliction?

Introducing the multi-source related similarity parameters can be useful to prevent semantic confliction. In the scientific authors network "co-citation" and "coupling" are two parameters that evaluate the similarity between two authors in terms of subject matter, and number of co-citations as a basis for giving weights to links based on these similarity parameters. These parameters however have drawbacks, which are beyond the scope of this paper and need further research.[25, 31]

C. Using multiple properties and relationships (multiple sources) separately

This approach has no limitation in the number of sources, data types and different kinds of relations. Each node and each edge have properties to show all of the network aspects. Division of an informative network makes self-descriptive communities. If a network is presented as a graph, a multisource network is compatible to colored graph theory.

1) Using overlaps between communities to find soft-membership of each node.

Each source needs its community detection method e.g. clustering for user features and graph partitioning for link data. In other words, there are some communities for each source. To illustrate the matter, let's look at an example: family members have a lot of common interactions with each other, so they are linked strongly. On the other hand, each family member may have co-workers with the same demographic features to them. So it is possible for them to have their own community in a general network too. Finding the overlaps or similarities between peer communities of each source needs to be considered in order to find the finally overlapped communities. There are some methods like those clustering similarity parameters such as Rand-index. Furthermore, to find peer communities, parameters need to be adapted to this purpose.

2) Multi-source Integration

This approach is the same as hybrid methods discussed in related work. Combining or mixing different data sources and different community detection methods has positive effects on scalability, accuracy, etc. Probability models are used extensively as a main method of CDP.

Hybrid methods can handle multiple sources or several algorithms together, but solving some problems made it a complex process, problems like how to handle high volume text datasets or how to gather multidimensional data which raises the question of which type of data fits the aim of community detection in a better way.

3) Multi-source Aggregation

Extracted data and information from sources seem to be helpful for approximation of algorithm inputs and obtaining high quality results. This approach uses information and data from the main and peripheral sources for CDP, as a means for the complementary information to improve the main algorithm which handles the main source. In [32] influence concept and user links have been intertwined. By choosing influential nodes of a social network, the algorithm speculate community cores which help to improve community detection results.

Fig. 2 shows fusion approaches in a reversed pyramid. Simple method (SM), Transforming and integrating all sources into one graph (Transforming), Multi-source Aggregation (MS Aggregation), Multi-source Integration (MS Integration), Using overlaps between communities (Peer communities overlaps). Development complexity, the number of sources and self-descriptiveness of each approach will be increased in top levels.

Figure 2. Comparison between multi-source fusion approaches

V. CONCLUSIONS

Although many methods have been proposed to detect communities in the social network, the community detection process (CDP) was not cleared yet. Therefore, in this paper, this process is introduced, including five main steps: source selection, data collection and data preprocessing, fusion approach determination, proper methods selection and finally, verification and interpretation of detected communities. The main contribution was to determine a fusion approach by which the proper community detection method can be selected. Then, several fusion approaches are compared. Transformation approach is applicable to handle few sources which assign weights to edges or nodes based on their similarities. However, this is a situation where semantic confliction might be happened. Other approaches are proper to handle several sources such as content, link, media and other features, separately. To make consistent results overlaps can be used. Furthermore, hybrid methods handle sources in sequential steps or using other sources as complementary information which can be used to estimate algorithm input parameters.

REFERENCES

[1] A. Pothen, "Graph partitioning algorithms with applications to scientific computing," in Parallel Numerical Algorithms, ed: Springer, 1997, pp. 323-368.

[2] M. Girvan and M. E. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences, vol. 99, pp. 7821-7826, 2002.

[3] J. Shi and J. Malik, "Normalized cuts and image segmentation," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 22, pp. 888-905, 2000.

[4] M. E. Newman, "Detecting community structure in networks," The European Physical Journal B-Condensed Matter and Complex Systems, vol. 38, pp. 321-330, 2004.

[5] A. Lancichinetti and S. Fortunato, "Limits of modularity maximization in community detection," Physical Review E, vol. 84, p. 066122, 2011.

[6] H. N. Djidjev, "A scalable multilevel algorithm for graph clustering and community structure detection," in Algorithms and Models for the Web-Graph, ed: Springer, 2008, pp. 117-128.

[7] Z. Ye, S. Hu, and J. Yu, "Adaptive clustering algorithm for community detection in complex networks," Physical Review E, vol. 78, p. 046115, 2008.

[8] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in neural information processing systems, vol. 2, pp. 849-856, 2002.

[9] M. Mitrović and B. Tadić, "Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities," Physical Review E, vol. 80, p. 026123, 2009.

[10] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," in Proceedings of the 15th international conference on World Wide Web, 2006, pp. 173-182.

[11] K. Henderson, T. Eliassi-Rad, S. Papadimitriou, and C. Faloutsos, "HCDF: A Hybrid Community Discovery Framework," in SDM, 2010, pp. 754-765.

[12] B. Kernighan, Lin, S, " An efficient heuristic procedure for partitioning graphs," Bell system technical journal, vol. 49, pp. 291-307, 1970.

[13] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in Proceedings of the 16th international conference on World Wide Web, 2007, pp. 835-844.

[14] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in Proceedings of the 4th ACM European conference on Computer systems, 2009, pp. 205-218.

[15] D. Greene and P. Cunningham, "Multi-view clustering for mining heterogeneous social network data," presented at the 31st European Conference on Information Retrieval 2009.

[16] T. Nguyen, D. Phung, B. Adams, T. Tran, and S. Venkatesh, "Hyper-community detection in the blogosphere," in Proceedings of second ACM SIGMM workshop on Social media, 2010, pp. 21-26.

[17] S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, pp. 75-174, 2010.

[18] H. Li, Z. Nie, W.-C. Lee, L. Giles, and J.-R. Wen, "Scalable community discovery on textual data with relations," in Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 1203-1212.

[19] N. F. Chikhi, B. Rothenburger, and N. Aussenac-Gilles, "Combining link and content information for scientific topics discovery," in Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on, 2008, pp. 211-214.

[20] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 2009, pp. 927-936.

[21] F. Moser, R. Ge, and M. Ester, "Joint cluster analysis of attribute and relationship data withouta-priori specification of the number of clusters," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 510-519.

[22] M. Al Hasan and M. J. Zaki, "A survey of link prediction in social networks," in Social network data analytics, ed: Springer, 2011, pp. 243-275.

[23] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: Morgan kaufmann, 2006.

[24] R. Xu and D. Wunsch, "Survey of clustering algorithms," Neural Networks, IEEE Transactions on, vol. 16, pp. 645-678, 2005.

[25] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 487-494.

[26] C. C. Aggarwal, An introduction to social network data analytics: Springer, 2011.

[27] L. Cao, G. Qi, S.-F. Tsai, M.-H. Tsai, A. Del Pozo, T. S. Huang, et al., "Multimedia Information Networks in Social Media," in Social Network Data Analytics, ed: Springer, 2011, pp. 413-445.

[28] K. Yang, "Combining Text-and Link-based Retrieval Methods for Web IR," in TREC, 2001.

[29] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," Physical Review E, vol. 78, p. 046110, 2008.

[30] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," Data Mining and Knowledge Discovery, vol. 24, pp. 515-554, 2012.

[31] Y. Wang and M. Kitsuregawa, "On combining link and contents information for web page clustering," in Database and expert systems applications, 2002, pp. 902-913.

[32] R. R. Khorasgani, J. Chen, and O. R. Zaïane, "Top leaders community detection approach in information networks," in Proceedings of the 4th Workshop on Social Network Mining and Analysis, 2010.