International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

# Efficient Implementation of Community Detection in Large Networks Using Framework Model

**Saradha[1], Dr. P. Arul[2]**
[1]Research Scholar, Bharathiar University, Coimbatore, Tamilnadu, India
[2]Assistant Professor, Department of Info. Technology, Govt Arts College, Trichy, Tamilnadu, India

## ABSTRACT

Given a large network, local community detection aims at finding the community that contains a set of query nodes and also maximizes (minimizes) goodness metric. Furthermore, due to the inconvenience or impossibility of obtaining the complete network information in many situations, the detection becomes more challenging. This problem has recently drawn intense research interest. Various goodness metrics have been proposed. And most of them base on the statistical features of community structures, such as the internal density or external sparseness. However, the metrics often result in unsatisfactory results by either including irrelevant subgraphs of high density, or pulling in outliers which accidentally match the metric for the time being. Furthermore, when in a highly overlapping environment such as social networks, the unconventional community structures make these metrics usually end up with a quite trivial detection result. We engage in an in-depth benchmarking study of community detection in social networks. We formulate a generalized community detection procedure and propose a procedure-oriented framework for benchmarking. This framework enables us to evaluate and compare various approaches to community detection systematically and thoroughly under identical experimental conditions. Upon that we can analyze and diagnose the inherent defect of existing approaches deeply, and further make effective improvements correspondingly. We have re-implemented ten state-of-the-art representative algorithms upon this framework and make comprehensive evaluations of multiple aspects, including the efficiency evaluation, performance evaluations, sensitivity evaluations, etc. We discuss their merits and faults in depth, and draw a set of take-away interesting conclusions.

**Keywords:** Community Detection, Framework, Social, Large Networks.

## I. INTRODUCTION

Community or modular structure is considered to be a significant property of real-world social networks as it often accounts for the functionality of the system. Despite the ambiguity in the definition of *community*, numerous techniques have been developed for both efficient and effective community detection. Random walks, spectral clustering, modularity maximization, differential equations, and statistical mechanics have all been used previously. Much of the focus within community detection has been on identifying *disjoint* communities. This type of detection assumes that the network can be partitioned into dense regions in which nodes have more connections to each other than to the rest of the network.

Despite the differences of Social Media networks with respect to the entities and the type of relations they model, they present a significant source of intelligence since they encode the online activities and inputs of masses of Social Media participants. Not only is it possible by analyzing such networks to gain insights into the social phenomena and

processes that take place in our world, but one can also extract actionable knowledge that can be beneficial in several information management and retrieval tasks, such as online content navigation and recommendation.

However, the analysis of such networks poses serious challenges to data mining methods, since these networks are almost invariably characterized by huge scales and a highly dynamic nature. Despite the differences of Social Media networks with respect to the entities and the type of relations they model, they present a significant source of intelligence since they encode the online activities and inputs of masses of Social Media participants. Not only is it possible by analyzing such networks to gain insights into the social phenomena and processes that take place in our world, but one can also extract actionable knowledge that can be beneficial in several information management and retrieval tasks, such as online content navigation and recommendation. However, the analysis of such networks poses serious challenges to data mining methods, since these networks are almost invariably characterized by huge scales and a highly dynamic nature.

The major challenges usually encountered in the problem of community detection in social media data are highlighted below:

**Scalability**

The amount of online social media content over the internet is raising everyday at a tremendous rate. Currently, the size of social networks is in scale of billions of nodes and connections. As the network is expanding, both the space requirement to store the network and time complexity to process the network would increase exponentially. This imposes a great challenge to the conventional community detection algorithms. Traditional community detection methods often deals with thousands of nodes or more.

**Heterogeneity**

Raw social media networks comprise multiple types of edges and vertices. Usually, they are represented as hyper graphs or k-partite graphs. Majority of community detection algorithms are not applicable to hyper graphs or k-partite graphs. For that reason, it is common practice to extract simplified network forms that depict partial aspects of the complex interactions of the original network.

**Evolution**

Due to highly dynamic nature of social media data, the evolving nature of network should be taken into account for network analysis applications. So far, the discussion on community detection has progressed under the silent assumption that the network under consideration is static. Time awareness should be incorporated in the community detection approaches.

**Evaluation**

The lack of reliable ground-truth makes the evaluation extremely difficult. Currently the performance of community detection methods is evaluated by manual inspection. Such anecdotal evaluation procedures require extensive manual effort, are non-comprehensive and limited to small networks.

**Privacy**

Privacy is a big concern in social media. Facebook, Google often appear in debates about privacy Simple anonymization does not necessarily protect privacy. As private information is involved, a secure and trustable system is critical. Hence, lot of valuable information is not made available due to security concerns.

## II. LITERATURE RECIEW

Discovering the groups in a network where individuals “group memberships” are not explicitly given is the concept behind the community detection. Community detection attempts to solve the problem which is the identification of groups of vertices

(nodes) that are more densely connected to each other than to the other remaining network. Detecting communities is of great importance in sociology, biology and computer science where systems are often represented as graphs. Real networks are not random graphs, as they display big in homogeneities revealing a high order and organization. The distribution of edges is not only globally, but also locally inhomogeneous, with high concentration of edges within special groups of vertices, and low concentrations between these groups. This feature of real network is called "community structure or clustering".

In a community, nodes are connected with each other based on their human relationships like friendship, colleague etc. In computer science, communities can be regarded as sub-graphs of networks. The whole complex network can be generated as a graph, which is consisted of many sub-graphs. Community detection attempts to solve the problem which is the identification of groups of vertices, that are 'more densely connected' to each other than to the rest of the network. Detecting and analyzing the community structure of network ranges to important findings in wide range of domains like biology to social sciences to web. Community detection has increasing interest in applying on social media not only as a means of understanding the underlying phenomena taking place in such systems, but also to exploit its results in intelligent services and applications e.g. automatic event detection in social media content.

"**Detecting Community Structure in Networks**" by **M.E.J Newman**, In this paper, the focus is given on reviewing the different algorithmic method for finding community of densely connected vertices in the network data. The discussion of some of the traditional approaches, such as spectral graph partitioning and hierarchical clustering is also been done, but further it was found number of shortcomings as far as the concern for the analysis of the large real-world network. There was also the description of the, methods based on Iterative removal of "Between community edges", which also includes the "between's-based method" of Girvan & Newman and Monte Carlo resample variation ,proposed by Tyler and also the algorithm based on "counts of short loops", which was proposed by Radicchi.

"**Identifying overlapping communities in networks using evolutionary method**" by **W.Zhan, J.Guan & H.Chen, proposes,** In this paper, the presentation of an encoding scheme for an overlapping partition of a network is done. The two informativeness measure for a node is proposed and presents an evolutionary scheme between two segments over the population. This evolutionary method was for detecting overlapping community structure in the network. For the representation of the overlapping part of the network , there has been developed an encoding scheme composed of two segments, the first one represents a disjoint partition and the other one represents an extension of the partition which allows the multiple membership.

"**Personalized recommendations based on time-weighted overlapping community detection**" by **H. Feng& J.Tian, proposes,** In this paper, a recommendation based approach called TOTAR (Temporal overlapping community detection using time weighted Association Rules) is proposed which is based on time weighted overlapping community detection and association rule mining. The different approaches have been incorporated to synchronize the time effects in the proposed algorithm to improve its performance and predict the user's dynamic interests over the time. Different data sets has also been used from MOVIELENS and NETFLIX and performance is compared with other algorithms especially the accuracy and diversity [5].

## III. PROPOSED WORK

In this paper, we have designed a benchmark for community detection. Our benchmark consists of four core modules:

- **Setup**, including a set of algorithms, real-world and synthetic datasets, parameter configurations, and a unified graph model converted from the datasets;
- **Detection Framework**, a generalized detection procedure with high abstraction of the common workflow of community detection
- **Diagnoses**, which provide targeted diagnoses on these algorithms based on our framework, leading to directions of improvement over the existing work.
- **Evaluation**, a comprehensive evaluation system for community detection from different aspects

The benchmark contains a universal framework which abstracts the key factors, phases and steps from many approaches to com-munity detection tasks, and makes it easy to implement classical or latest algorithms for comparison.

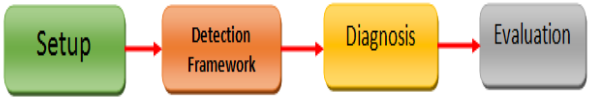


**Figure 1.** Proposed Structure

Moreover, it consists of a comprehensive suite of widely-recognized metrics for evaluation of various concerned aspects, including the efficiency evaluation on the time cost, performance evaluations on accuracy and effectiveness, sensitivity evaluations on network density and mixture degree, and additional evaluations on community distribution and the ability to avoid excessive outliers. By modularizing and separating key factors and steps, our framework allows us to study the strength and weakness of each algorithm thoroughly, and make diagnoses and targeted prescriptions for improvement. In this benchmark we provide a common code base with algorithms implemented in the same environment, and thus make the comparison more fair and credible.

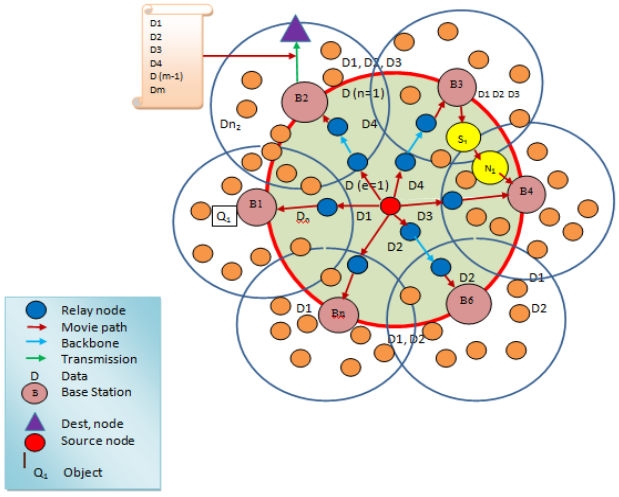


**Figure 2.** Large Network

We conduct in-depth evaluations for the community detection algorithms within our framework using the proposed benchmark, which covers the efficiency, accuracy, effectiveness, density sensitivity, mixture sensitivity, outliers, community distribution and diagnosis effects. We introduce the datasets and parameter configurations in the benchmark at first, and then report our thorough evaluation methodology and results. We summarize our findings and rate the algorithms intuitively at last. All experiments are conducted on a computer running Windows Server 2008 with an Intel Xeon 2.0 GHz CPU and 256 GB RAM.

We make the following main contributions:

- We propose a novel procedure-oriented framework by formulating a generic workflow of community detection via abstracting and modularizing the key factors and steps.
- We review the family of community detection approaches, and re-implement ten state-of-the-art representative algorithms in a common code base (using standard C++) by mapping them to the framework based on their specifics.
- We make in-depth evaluations on these approaches based onour benchmark using both real-world and synthetic datasets.

- We draw a set of interesting take-away conclusions, and provide intuitive and brief ratings on concerned algorithms.
- We also present how to make diagnoses for existing approaches, leading to significant performance improvements

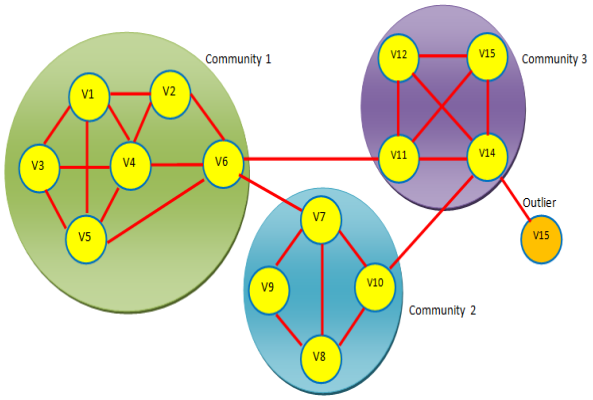


**Figure 3.** Community Detection

## IV. RESULTS & DISCUSSIONS

In semiconductor manufacturing, previous studies have shown that wafers processed in normal equipment conditions are located near each other in the feature space [16]. Based on observations, we consider four illustrative examples that address two input features.

### i) Accuracy

The following result shows that the proposed approach provides close to the saliency approach and better than other approaches.

**Accuracy = (TP+TN)/(TP+FN+FP+TN)**

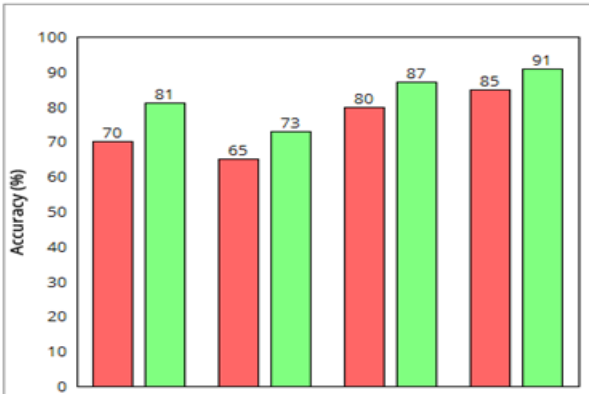


**Figure 6.1.** Accuracy for Community Detection

### ii) Precision

The proposed incremental clustering approach works as an extension to the saliency approach. It enhances the precision for faults having large smoothed regions and provides better recall than other approaches.

**Precision = TP/ (TP+FP)**

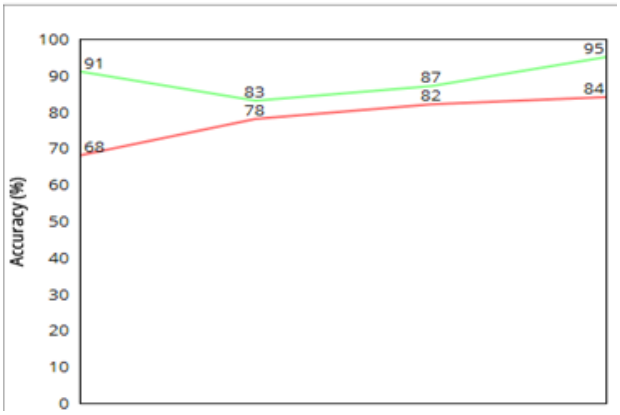


**Figure 6.2.** The Precision for Fault Detection

## V. CONCLUSIONS

There are some challenges we have faced in our work, which indicates some limitations on our studies. Firstly, all of our algorithms are exact, in the sense that we do not approximate a graph analytic. This may not be realistic for real-world use cases, but we believe that our contributions can be used as a building block for further analyses. Secondly, evaluating the results of a new graph analytics based on the ground-truth information is quite challenging. For the dense subgraph discovery and community detection problems, existing ground-truth information is quite dependent to the domain. For instance, the densest region in a protein-protein interaction network may not correspond to something meaningful and larger subgraphs with lower densities are more of interest.

## VI. REFERENCES

[1]. Q. K. Telesford, K. E. Joyce, S. Hayasaka, J. H. Burdette, and P. J. Laurienti, "The ubiquity of

small-world networks," Brain Connectivity, vol. 1, no. 5, pp. 367–375, 2011.

[2]. S. Fortunato, "Community detection in graphs," Physics Reports, vol. 486, no. 3, pp. 75–174, 2010.

[3]. A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of Facebook networks," Physica A: Statistical Mechanics and its Applications, vol. 391, no. 16, pp. 4165–4180, 2012.

[4]. R. Guimera, S. Mossa, A. Turtschi, and L. N. Amaral, "The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles," Proceedings of the National Academy of Sciences, vol. 102, no. 22, pp. 7794–7799, 2005.

[5]. J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," ACM Computing Surveys, vol. 45, no. 4, p. 43, 2013.

[6]. M. Salath´e and J. H. Jones, "Dynamics and control of diseases in networks with community structure," PLoS Computational Biology, vol. 6, no. 4, p. e1000736, 2010.

[7]. Pretty good privacy network dataset { KONECT, September 2016.

[8]. U. rovira i virgili network dataset { KONECT, January 2016.

[9]. R. Aktunc, I. H. Toroslu, M. Ozer, and H. Davulcu. A dynamic modularity based community detection algorithm for large-scale networks: Dslm. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, ASONAM '15, pages 1177{1183, 2015.

[10]. C. Fan, K. Xiao, B. Xiu, and G. Lv. A fuzzy clustering algorithm to detect criminals without prior information. In 2014 IEEE/ACM International Conference on Ad-vances in Social Networks Analysis and Mining (ASONAM 2014), pages 238{243, Aug 2014.

[11]. S. Fortunato. Community detection in graphs. Physics Reports, 486(35):75 { 174, 2010.

[12]. J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection, June 2014.

[13]. X. Liu and T. Murata. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. Physica A: Statistical Mechanics and its Applications, 389(7):1493 { 1500, 2010.

[14]. P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. Generalized louvain method for community detection in large networks. In 2011 11th International Conference on Intelligent Systems Design and Applications, pages 88{93, 2011.