Global Journal of Pure and Applied Mathematics.
ISSN 0973-1768 Volume 14, Number 5 (2018), pp. 635-648

http://www.ripublication.com

Community Detection Framework in Large Networks Using Procedure Oriednted Framework

Mrs. Saradha
Research Scholar,
Bharathiar University,
Coimbatore, India.

Dr. P. Arul
Assistant Professor,
Dept of Info. Technology,
Govt Arts College, Trichy, India.

Abstract

With the development of Internet and computer science, more and more people join social networks. People communicate with each other and express their opinions on the social media, which forms a complex network relationship. Individuals in the social networks form a "relation structure" through various connections which produces a large amount of information dissemination. This "relation structure" is the community that we are going to research. Community detection is very important to reveal the structure of social networks, dig to people's views, analyse the information dissemination and grasp as well as control the public sentiment. During the last decade, many approaches have been proposed to solve this challenging problem in diverse ways, i.e. different measures or data structures. Unfortunately, experimental reports on existing techniques fell short in validity and integrity since many comparisons were not based on a unified code base or merely discussed in theory. 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 analyse 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. In addition, we present how we can make diagnoses for these algorithms resulting in significant improvements.

Keywords: Community, Detection, Social, large Network, Procedure, Framework.

1. INTRODUCTION

Many types of real-world datasets can be modeled with *networks*. A network provides a powerful mathematical tool to represent the relations in the data. Networks generated from real-world data are often divided into four categories, social, information, technological, and biological networks [1]. A *social network* is a network connecting the people who contact or interact with each other. Social networks are not limited to "online social networks" such as Facebook, Twitter, or LinkedIn. Other examples of social networks are the network of people collaboration, co-authorships, and co-appearance, as well as networks of communication between people such as telephone calls and emails. An *information network* is a network of entities containing information such as World Wide Web, network of citations, and word co-occurrence networks. A *technical network* refers to a manmade network such as the Internet, the electric power grid, networks of roads, railways, and airline routes. A *biological network* represents a biological system such as a network of metabolic pathways, protein-protein interactions, the food web, and the network of blood vessels.

Community detection aims to reveal groups of items in our network that are more closely knit towards each other, than towards the rest of the network. Although communities may arise in all types of networks, our intuitive understanding of them is particularly clear in a social network. In a social network, the connected items are social factors, such as persons or organizations, and the ties refer to social interaction or communication between pairs of actors. A group of friends together with their phone records, or internet chat logs, can most certainly be considered a social network. At the same time, it is clear that the group must also be part of a much larger network, and within this larger network they likely constitute exactly what we are looking to find, that is a community. Communities may also be defined in a hierarchical matter. The network of all high-school students in some city, for example, may be broken down into communities comprised of students that all go to the same school. These communities may again be refined into groups of friends, or even students that share the same classes. Many methods for finding communities, in fact outputs such a hierarchic structure, and it is often so that each level in the hierarchy has a different interpretation.

An excessively studied structural property of real-world networks is their community structure. The community structure captures the tendency of nodes in the network to group together with other similar nodes into communities. This property has been observed in many real-world networks. Despite excessive studies of the community structure of networks, there is no consensus on a single quantitative definition for the concept of *community* and different studies have used different definitions. A community, also known as a *cluster*, is usually thought of as a group of nodes that have many connections to each other and few connections to the rest of the network. Identifying communities in a network can provide valuable information about the structural properties of the network, the interactions among nodes in the communities, and the role of the nodes in each community.

2. LITERATURE REVIEW

Detecting communities in networks is comparable to partitioning sets of data into similar clusters. The main difference is that data clustering allows for grouping any pair of data points together, whereas community detection only allows for directly grouping linked nodes together.

Community detection is a stimulating field of research. There are various community detection algorithms available but most of the algorithms are able to detect disjoint communities only. As overlapping community detection is comparatively new approach less algorithms are present for this approach. Some of these works are described below.

Wangsheng Zhang, Gang Pan_, ZhaohuiWu, Shijian Li, et al., proposes, Complex networks describe a wide range of systems in nature and society. To understand the complex networks, it is crucial to investigate their internal structure. In this paper, we propose an online community detection method for large complex networks, which make it possible to process networks edge-by-edge in a serial fashion. We investigate the generative mechanism of complex networks and propose a split mechanism based on the degree of the nodes to create new community. Our method has linear time complexity.

Vinh-Loc Dao, C´ecile Bothorel, and Philippe Lenca, et al, proposes, evaluating a network partition just only via conventional quality metrics – such as modularity, conductance or normalized mutual of information – is usually insufficient. Indeed, global quality scores of a network partition or its clusters do not provide many ideas about their structural characteristics. Furthermore, quality metrics often fail to reach an agreement especially in networks whose modular structures are not very obvious. Evaluating the goodness of network partitions in function of desired structural properties is still a challenge. This descriptive approach also helps to clarify the composition of communities in real world networks. The methodology hence brings us a step closer to the understanding of modular structures in complex networks.

Emilio Ferrara et al., proposes, Understanding social dynamics that govern human phenomena, such as communications and social relationships is a major problem in current computational social sciences. In particular, given the unprecedented success of online social networks

(OSNs), in this paper we are concerned with the analysis of aggregation patterns and social dynamics occurring among users of the largest OSN as the date: Facebook. To this purpose, we acquired a sample of this network containing millions of users and their social relationships; then, we unveiled the communities representing the aggregation units among which users gather and interact; finally, we analyzed the statistical features of such a network of communities, discovering and characterizing some specific organization patterns followed by individuals interacting in online social networks, that emerge considering different sampling techniques and clustering

methodologies.

Shweta Bansal, Sanjukta Bhowmick, Prashant Paymal, et al., proposes, Dynamic complex networks are used to model the evolving relationships between entities in widely varying fields of research such as epidemiology, ecology, sociology, and economics. In the study of complex networks, a network is said to have community structure if it divides naturally into groups of vertices with dense connections within groups and sparser connections between groups. Detecting the evolution of communities within dynamically changing networks is crucial to understanding the underlying dynamic processes driving complex systems. However, much of the current work in community detection is based on static networks. Most community detection methods treat each new configuration for every time step as a separate network and the communities have to be recomputed as a whole. Our method takes advantage of community information from previous time steps and thereby improves efficiency while maintaining the quality of community detection.

Michael T. Schaub, Jean-Charles Delvenne, et al., proposes, Community detection, the decomposition of a graph into essential building blocks, has been a core research topic in network science over the past years. Since a precise notion of what constitutes a community has remained evasive, community detection algorithms have often been compared on benchmark graphs with a particular form of assortative community structure and classified based on the mathematical techniques

they employ. However, this comparison can be misleading because apparent similarities in their mathematical machinery can disguise different goals and reasons for why we want to employ community detection in the first place. Here we provide a focused review of these different motivations that underpin community detection. This problem-driven classification is useful in applied network science, where it is important to select an appropriate algorithm for the given purpose.

3. COMMUNITY DETECTION - DEFINITIONS

There are many definitions of community detection. No definition is universally acclaimed. Often, the definition is based on the type of system under consideration or application dependent. After a careful review of these definitions, it can be concluded that communities are the subgraphs in which nodes are densely connected to each other when compared to the rest of the network. There are two types of community definitions; local and global. Local definitions focus on the subgraph under study but neglect the rest of the graph. On the other hand, in the case of global definitions, communities are defined with respect to the graph as a whole.

Local Definition

In social networks, a community often means a group whose participants are all

friends with each other. In graph theory, such a group is termed as a clique in whom every two distinct nodes are adjacent. This is a rather strict definition of community. According to this definition, if a single pair of nodes is not connected to other nodes in the network, it can be termed a community. Also a subgraph in which all but one pair of nodes is connected would not qualify as a community. Another important limitation of this definition comes about if we want to understand the hierarchical roles of nodes within the community. Using the definition of community as a clique, it is simply not possible.

Global Definition

In contrast to local definitions, in a global definition of community, we specify not only the relationship between nodes within the community, but also their relationship to nodes outside the community. A community is defined as a subgraph in which nodes are densely connected to each other and sparsely to the rest of the network. Within such a context, a global property of the graph is used in an algorithm which delivers communities. A key idea in the literature is that if a network has community structure, it is different from a random graph. Several definitions [18] draw on this notion. The random graph defined by [15], will not have community structure. As any two nodes of the graph have the same probability of being connected to each other, as a result, there will not be any special group or community. In the literature, the null model is defined as a random graph which matches the given graph in some of its structural properties. This null model is used as a comparison tool in order to find out whether the original graph exhibits community structure or not.

3. PROBLEM STATEMENT

Community structures are quite common in real networks. Social Networks often include community groups based on common location, interests, occupation etc. Metabolic Networks have communities based on functional groupings. Citation Networks form communities by research topic. Being able to identify these sub-structures within a network can provide insight into how network function and topology affect each other.

In the existing approach, they design an end-to-end framework for identifying communities from raw, noisy social media data. The framework is composed of two important phases. First, they introduce a new method of converting the raw, noisy social media data into a weighted entity-entity co-occurrence based consistency network. This includes a simple iterative noise removal procedure for cleaning the entity consistency network by removing noisy entity pairs. Secondly, they propose an approach for identifying coherent communities from the weighted entity network, by introducing novel notions of communityness and community, based on eigenvector centrality.

In our proposed approach, we introduce the new method called procedure oriented framework. We use this framework to solve three different problems from two distinct domains.

- The first problem involves detecting communities from raw social media data and showing the application of the communities discovered in a recommendation engine setting.
- The second problem is, given a set of communities of discovered by traditional community detection methods, we need to identify loose communities among them and partition them into compact ones.
- The third problem is about showing the application of such framework in an Image Annotation scenario in presence of noisy labels. The problem of image annotation is defined to be, given an unknown image, we need to predict labels which best describes the semantics of the image.

4. EXISTING SYSTEM APPROACH

The last decade has witnessed the birth of a new field of interest and research in the study Of complex networks, i.e. networks whose structure is irregular, complex and dynamically evolving in time, with the main focus moving from the analysis of small networks to that of systems with thousands or millions of nodes, and with a renewed attention to the properties of networks of dynamical units. Networks are all around us, and we are ourselves, as individuals, the units of a network of social relationships of different kinds and, as biological systems, the delicate result of a network of biochemical reactions.

We have discussed community dynamics and reviewed complex network structural parameters. We highlighted the importance of network centrality or degree centrality and network robustness for community detection. Centrality is correlated with degree. We discussed network or degree centrality (weighted Laplacian centrality) based on modified Palladian, weighted micro-community centrality. We also discussed and introduced algorithm for k-clique sub-community and optimal partition of k-clique sub-community for weighted modularity optimization and overlapping community detection based on degree and weighted micro-community centrality.

These new matrices and algorithms are helpful in identifying hidden level vulnerabilities. We analyzed real-world large-scale complex networks and carried out comparison of different community detection algorithms. Our results indicated certain relationship between degree centrality and modularity optimization. Network centrality and robustness will help for supervised community detection in overlapping communities. These algorithms will be useful for finding communities of densely connected vertices in network data. Scalable nature of this algorithm is valuable for analyzing more complex large-scale networks. It is also an interesting problem about the selection of the parameter k in our method. We will further investigate how to

determine an appropriate k for a given network later. In our future work we will put forward functional dynamics of complex network by incorporating network centrality and weighted clustering coefficient for identifying micro level communities and their associated relationship.

Disadvantages:

- Maximum Clique Problem: The maximum clique is the maximal clique with the maximum cardinality or weight.
- Maximal Clique problem: The maximal clique is the clique that is not a subset of any other clique (independent set).
- When the number of cliques starts to grow exponentially we cannot calculate the solution in polynomial time. The problem then becomes the NP-Hard problem.
- For a given graph, find all the cliques with the maximum cardinality.
- Listing all the maximal cliques.
- The decision problem of analysing does the graph Z contains a clique larger than a given size.

5. PROPOSED SYSTEM STRUCTURE

Detecting communities is a challenging and interesting area of research in the domain of complex networks. The goal of community detection algorithm is to find group of nodes of interest in a given network. At present, social network has emerged as one of the popular means of communication. Social networks describe a structure of relationships between individuals. A community in social network can be considered as a set of users, where each user interacts more frequently with users within the group than with users outside the group. Community detection has been widely used in social network analysis to study the behaviour and interaction patterns of people in social network.

Our benchmark consists of four core modules: (1) Setup, including a set of algorithms, real world and synthetic datasets, parameter configurations, and a unified graph model converted from the datasets;

(2) Detection Framework, a generalized detection procedure with high abstraction of the common workflow of community detection; (3) Diagnoses, which provide targeted diagnoses on these algorithms based on our framework, leading to directions of improvement over the existing work; (4) Evaluation, a comprehensive evaluation system for community detection from different aspects

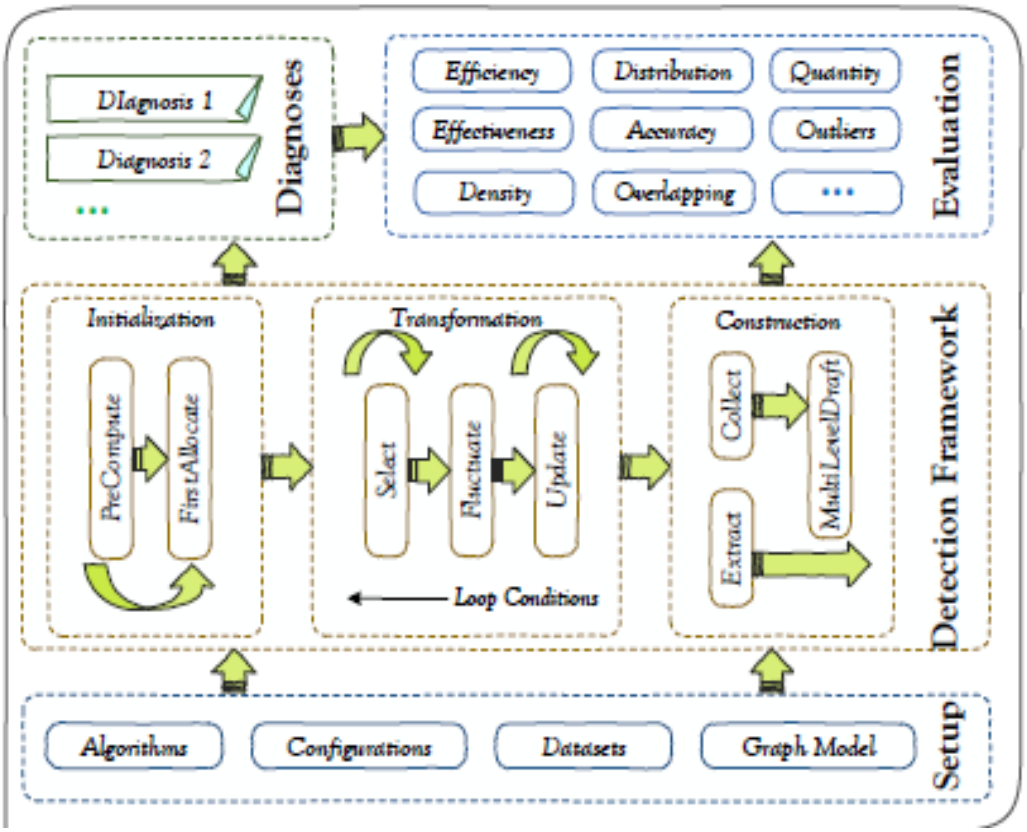

Fig 1: - Benchmark for Community Detection

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

We have conducted a comprehensive benchmarking study which focuses on the in-depth analysis, evaluation and comparison of the extensive work. To the best of our knowledge, this is the first work on the benchmarking study with a generalized framework on non-overlapping community detection techniques. 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 on our 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.

Existing algorithms usually solve the community detection problem with various methods based on different assumptions. This makes it difficult to comparatively analyze these algorithms thoroughly. For the sake of a better understanding of the underlying principles of community detection algorithms, we abstract two fundamental concepts, including the propinquity measure and the revelatory structure, which play critical roles in the community detection task and can be used to distinguish different approaches.

Community detection has been studied unremittingly all these years, and a particularly large number of effective approaches have been proposed. In this study we focus on the fundamental problem of non-overlapping community detection, which aims at finding the definite group (community) that each node belongs to in the graph.

Modularity is a function that evaluates the quality of a given partition of nodes in a graph as good communities. It is based on evaluating how much the graph, and the given partition differ from null model.

Consider a graph G = (V;E) and let |V| = n and |E| = m. Suppose we are given a community structure e for the graph where e = C1;C2; ::Ck is a partition of V into communities. Define d(v) to be the degree of node v. Let A be the adjacency matrix for G where Avw = 1 means there is an edge from node v to node w and Avw = 0 means there is no edge from node v to node w.

To obtain a null model, the following procedure is described in [3]. For every edge in the graph, we cut it in a half so we have two stubs. The total number of stubs is 2m. We now reattach the stubs at random to obtain a null graph G'. If (v;w) 2 ε, the probability that v and w are connected in G' is

$$\frac{d(v)d(w)}{2m}$$

Therefore, the difference between the actual number of edges and the expected number of edges between v and w is

$$A_{vw} - \frac{d(v)d(w)}{2m}$$

The modularity of the partition e is now defined to be the summation of the difference over all edges within communities [2]. In particular:

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{d(v)d(w)}{2m} \right] \delta(c_v, c_w)$$

Composed of two fundamental concepts and a generalized procedure, our framework is beneficial for understanding and analyzing the community detection approaches. The two concepts are the basis for solving the problem of community detection, and different implementations may lead to quite different performances, even for the same approach. The procedure is the modularization of the critical steps of this problem, and uncovers the generic detection workflow, making it easy to study various approaches deeply within the identical framework.

6. BENCHMARK EVALUATION

In this section, 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.

6.1. Performance Evaluation Parameters

The following performance parameters are commonly used in Dimensionality Reduction technique evaluation. The existing approach is compared with proposed scheme using these evaluation parameters. The performance of the TC process can be measured by one or more of the following methods:

6.1.1. Recall and Precision

They are two well known measures of effectiveness in text mining. While Recall is a measure of correctly predicted documents by the system among the positive documents, Precision is a measure of correctly predicted documents by the system among all the predicted documents. The system is evaluated in terms of precision,

recall and Fmeasure. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of existing relevant documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

$$precision = \frac{number\ of\ correct\ results}{number\ of\ all\ returned\ results}$$

$$recall = \frac{number\ of\ correct\ results}{total\ number\ of\ actual\ results}$$

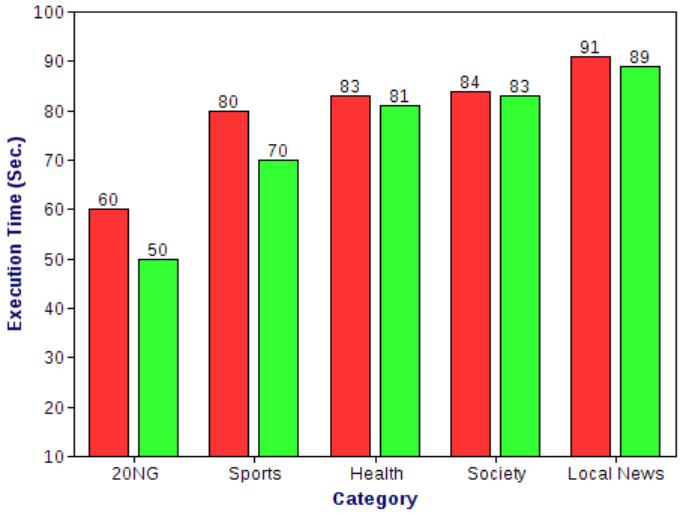

Fig 2: - Recall Parameter Evaluation

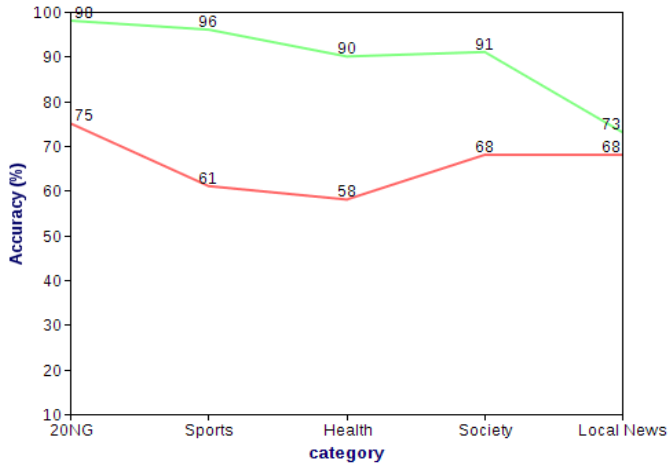

Fig 3: - Precision Parameter Evaluation

6.1.2. F-Measure

F-measure combines precision and recall and is the harmonic mean of precision and recall.

$$F-measure = 2 \times \frac{precision \times recall}{precision + recall}$$

Several experiments were conducted with different query documents and the precision, recall and F-measure of the output was calculated. This higher improvement in precision value can compromise for the very small percentage of drop in the recall value. Moreover, the F-measure which combines precision and recall is much improved for similarity than existing system.

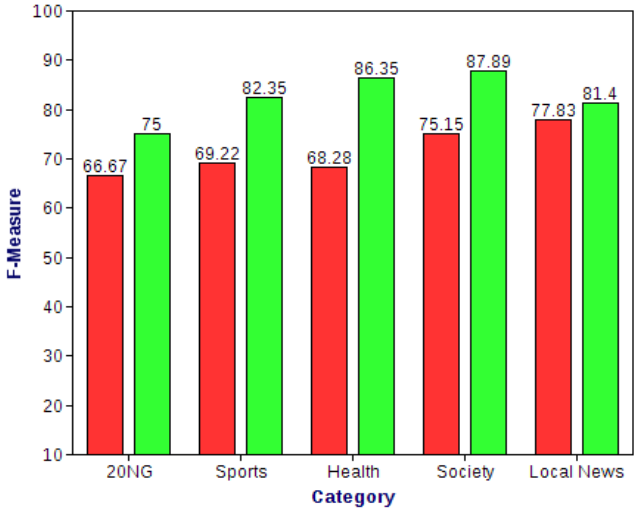

Fig 4: - F-Measure Parameter Evaluation

7. CONCLUSION

In this paper, we conduct a comprehensive benchmarking study on approaches to community detection in social networks. Within the proposed benchmark, we formulate a generalized procedure oriented framework, with high abstraction and nice modularization of the fundamental factors and critical steps of this problem. We have re-implemented ten state-of-the-art representative algorithms by mapping them to the proposed framework, and make in-depth evaluations on them based on our benchmark using both real-world and synthetic datasets. We discuss their merits and faults thoroughly, draw a set of interesting take-away conclusions, and provide intuitive ratings. In addition, we present how to make diagnoses for these algorithms based on our framework, and report significant improvements in the experimental study.

REFERENCES

[1] A. Prat-P´erez, D. Dominguez-Sal, and J.-L. Larriba-Pey. High quality, scalable and parallel community detection for large real graphs. WWW, pages 225–236, 2014.

[2] A. Prat-P´erez, D. Dominguez-Sal, J. M. Brunat, and J.-L. Larriba-Pey. Shaping communities out of triangles. CIKM, pages 1677–1681, 2012.

[3] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. ICDM, pages 745–754, 2012.

[4] F. Zhao and A. K. Tung. Large scale cohesive subgraphs discovery for social network visual analysis. VLDB, pages 85–96, 2012.

[5] 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.

[6] T. Alzahrani and K. J. Horadam. Community Detection in Bipartite Networks: Algorithms and Case studies, pages 25{50. 2016.

[7] 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.

[8] G. Laurent, J. Saram • aki, and M. Karsai. From calls to communities: a model for time-varying social networks. The European Physical Journal B, 88(11):1{10, 2015.

[9] Yang J, Leskovec J (2012) Defining and evaluating network communities based on ground-truth. In: Proceedings of the Twelfth International Conference on Data Mining. pp. 745–754.

[10] Lu¨ L, Zhou T (2011) Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications 390: 1150–1170.

[11] Zhang W, Pan G, Wu Z, Li S (2013) Online community detection for large complex networks. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. pp. 1903–1909.

[12] H. Aksu, M. Canim, Y. Chang, I. Korpeoglu, and O. Ulusoy. Distributed k-core view materialization and maintenance for large dynamic graphs. Knowledge and Data Engineering, IEEE Transactions on, PP(99):1{1, 2014.

[13] M. Baglioni, F. Geraci, M. Pellegrini, and E. Lastres. Fast exact computation of betweenness centrality in social networks. In ASONAM, 2012.

[14] W. Cui, Y. Xiao, H. Wang, Y. Lu, and W. Wang. Online search of

overlapping communities. In SIGMOD, 2013.

[15] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In WSDM, pages 635–644, 2011.

[16] V. Belak, S. Lam, and C. Hayes. Towards maximising cross-community information diffusion. In 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 171–178, Aug 2012.

[17] M. Chen, T. Nguyen, and B. Szymanski. A new metric for quality of network community structure. ASE Human Journal, 1(4):226–240, 2013.

[18] P. Chen and S. Redner. Community structure of the physical review citation network. J. Informetrics, 4(3):278–290, 2010.

[19] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. PLoS ONE, 6(9):e24926, 09 2011.