Proceedings of the Federated Conference on Computer Science and Information Systems pp. 473–477
DOI: 10.15439/2020F88
ISSN 2300-5963 ACSIS, Vol. 21

On the Community Discovery Methods for Complex Networks: A Case Study

Kirubel W. Afrassa, Genco Cosgun
Computer Engineering Dept.
Yildiz Technical University
Istanbul, Turkey
{kirubel.afrassa, genco.cosgun}@std.yildiz.edu.tr

Ulku F. Gursoy, Enes M. Yildiz
R&D Center
Intellica Business Intelligence Consultancy
Istanbul, Turkey
{ulku.gursoy, enes.yildiz}@intellica.net

Mehmet S. Aktas
Computer Engineering Dept.
Yildiz Technical University
Istanbul, Turkey
aktas@yildiz.edu.tr

***Abstract*—The inherent knowledge discovery problem regarding networks that represent complex real world phenomenon is a popular research topic. Specifically, in social network analysis (SNA), several community discovery techniques with various approaches have been put forward to distinguish closely related entities. Identifying the relevant techniques to utilize based on the context of the application is a key difficulty researchers face. In this study we propose a methodology for classifying these techniques, visualize a prototype, and analyze the performance and quality of selected approaches over a real world call detail record (CDR) data set.**

***Index Terms*—community discovery, community detection algorithms, visualization, CDR**

I. Introduction

ONE of the many applications of networks is community discovery. Intuitively, community ensues entities that are closer to each other within any arbitrary group, than outside it. Closeness maybe defined by common properties, similar roles or various measurements made on entity interaction. In a network, entities can be characterized by nodes and interactions among them can be embodied using edges. Despite the huge literature available on communities represented in a network, scholars do not have an agreement on what a network with communities corresponds to. However, the widely accepted definition is the *planted l-partition model* [1]. In this model p_{in} and p_{out} signify probability of each node being connected to nodes in its group and different groups respectively. If $p_{in} > p_{out}$ the network has communities present otherwise, the graph is random.

Community discovery within a network of such description can be viewed as maximizing the number of edges between any k groups within that community and minimizing the number of edges outside each of those groups. In terms of nodes, it can also be expressed as a generalization of a data mining problem that is analogous to unsupervised node clustering. But this definition doesn't account for nodes relational behavior. Different community discovery applications and algorithms use diverse and specialized interpretations for community detection. Consequently, there are many types of community discovery algorithms. They mainly constitute varying definitions based on node relationship and the definition of a community. Therefore, we propose a methodology for classifying different community discovery techniques in social network analysis (SNA) in order to narrow down the multitude of available methods. After that, we present a case study on a large scale call detail record (CDR) data set using the selected approaches. The selected approaches are implemented and evaluated based on a strategic ground truth definition for unlabeled data sets. Moreover, the performance and scalability of the selected algorithms are tested on both CDR and YouTube data sets. As part of the experimental study we also develop a visualization software that illustrates networks and discovered communities.

The rest of this paper is organized as follows. Section II reviews literature, Section III introduces the methodology used in choosing community discovery methods, on Section IV a prototype of discovered communities visualizer is presented, Section V defines different metrics in order to evaluate the performance of community discovery algorithms over larger scale social network and CDR data sets and Section VI summarizes the results obtained and possible future works.

This study is funded by Scientific and Technological Research Council of Turkey (TÜBİTAK) project no. 3181048

II. Literature Review

Community detection techniques have been widely used for variety of purposes. Among these SNA is a common application. Social-based metaheuristic optimization algorithms have been used in order to identify overlapping communities [2] [3]. Moreover, recognized communities can pertain to improve performances of other operations. For instance, community discovery is used to boost low accuracy ranking algorithms in identifying top information spreaders [4].

In SNA and other applications, visualization tools have been made with their own individual variations. While hierarchical community structure and fence-sitting nodes visualization is performed on [5], there are also tools that use some nuance of Newman's modularity optimization algorithm for clustering, prior to visualization [6], [7]. CDR data has been impactful in the analysis of varying models and its application is growing exceedingly. Among these usages, urban sensing and planning [8], [9], traffic engineering [10], [11], predicting energy consumption [12], improved churn prediction using both CDR data and community detection [13] can be cited. Graph data

analysis have been studied in different research fields such as provenance field [14]–[22]. Provenance graph is mainly used to understand the data lineage. Different from previous work in provenance research filed, in this study, we focus on analysing graph data to identify the sub networks.

III. Methodology

The variety of community discovery algorithms are more diversified by their abilities to support different types of networks. The main categories to consider would be their capability to support data sets that are directed or weighted and whether or not communities overlap.

Overlapping community discovery refers to a node's ability to be a member of multiple communities at the same time.

On the other hand, in non-overlapping community detection, no two or more communities share a common node. The computational complexity of non-overlapping community discovery algorithms is generally good because of the deducted operation of identifying nodes that are a member of multiple communities.

In real world SNA community discovery is unsupervised i.e. there is no ground truth data to justify the acquired results. And most community discovery algorithms just as other unsupervised learning processes, involve hyper parameters and initialization procedures that may lead to degenerate results or local minima / maxima. The additional diverse approaches explained thus far point out that careful selection of community discovery algorithms is necessary for increasing the quality of obtained results. Therefore, from the above broadly classified approaches, we believe researchers should aim to identify algorithms that are suitable for their data set and the respective outcome required.

Considering this application driven proposition, in this case study, we identify two models that are prominent in terms of their approach and relevancy to SNA.

These approaches are III-A Diffusion Model and III-B Motif-based Model. As it will be self-evident, we aim to take advantage of the different perspective these two approaches provide on community definition and detection.

A. Diffusion Model

One of the main approaches of community detection algorithms is Diffusion. Algorithms that take on diffusion model are not only used in community discovery, but also in viral marketing and churn analysis.

Group of nodes that are clustered through propagation of the same or similar properties summarizes community detection using this model [23], [24]. For example, in social network of a university, similarity among students may be defined by their common hobbies or lessons. These common information make nodes densely connected thus, creating a community. Similarity and shared information is also the basis of influence a node has in its community.

This approach makes the analysis of group dynamics apparent since the behavior of nodes is very closely related to the influence that drives it. For instance, if a highly influencer node leaves a network, eventually, it affects the existence of other nodes of the same community. And what is more, there is a high possibility that the community it belongs to scatters since the influencer node was crucial. Likewise, a node is attracted to a community that is more similar to itself.

B. Motif-based Model

In the above community detection approach, low-level connectivity can be seen as a theme. On other hand, motif-based approaches aim to address insights that can be gained by considering high-level connectivity. These methods achieve this through detecting dense subgraphs that appear in the network a lot more than those in a randomized network. The substructures are defined by a distinct pattern of interaction between nodes. The implication being that these sets of nodes within the hypergraph reflect a specific function or relationship.

Network motifs were first proposed by [25] and can be formally denoted as $M = \{V_{\mathbb{M}}, E_{\mathbb{M}}\}$ where $V_{\mathbb{M}}$ is a set of m nodes and $E_{\mathbb{M}}$ is a set of edges between $m-1$ (line motif) and $\frac{m(m-1)}{2}$ (clique motif) in the motif $\mathbb{M}$ [26]. But generally a network motif consists between 3-8 number of nodes [27]. This is because higher-order motifs are structurally complicated.

There are many variations implemented on this core approach that seek to enhance different defects or achieve a certain goal [26], [28], [29].

This perspective of over-watching the organization of a network for community detection must be incorporated with other techniques that advance towards addressing the negligence of lower-order connectivity and enhance the ability to find multi-layered motifs.

IV. Prototype

To demonstrate community discovery in a network we use Dash [33] Python framework for web applications that extends Cytoscape.js [34] and renders it. Fig. 1 shows a visualization of Label Propagation algorithm over a Twitch data set described in Section V-A. The shape of the networks represented in this figure can be adjusted using an interactive graphical user interface (GUI). Moreover, from the GUI, the user can choose among the two algorithms mentioned so far and other example data sets. After selection of these preferences, the network is rendered in the user's browser along with basic statistical, centrality and network defining properties. Due to space limitation these details are omitted in the figure. The network depicted on the left shows the original network and on the right we see the community substructure of nodes that belong to a selected community.

The cone shaped circular figure depicts hierarchy based on connectivity i.e., the most inner nodes found at the center have the most number of edges. Similarly connectivity decreases going out further in to the outer arcs. Fig. 2 zooms in on the most inner circle of the network and discovered community substructures. In fig. 2 (left) Node 166 (colored red to show membership) on the network can be identified as one of the

Fig. 1: Twitch data set visualization (Left). Discovered Community (Right).

most well connected nodes therefore, it is depicted at the very center of the network. And as anticipated, as shown in fig. 2 (right), the same node is at the center of it's community.

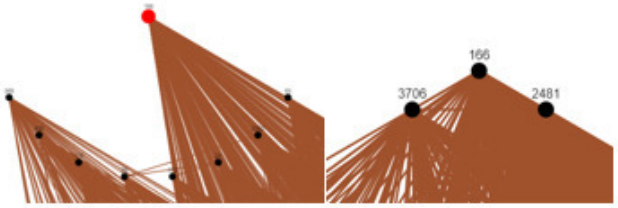


Fig. 2: Node 166 in the network (Left). Node 166 in the its Community (Right).

V. Evaluation

A. Data Set

In this paper we have used two test data sets. The first is a small Twitch [36] data set of 7126 nodes and 35324 edges for visualizing a prototype of a network and discovered communities. The second is a YouTube data set provided by [35]. It defines over 1.1 million nodes as YouTube users and almost 3 million edges represent user friendships. We use this data set for evaluating our ground truth method that is later used in the case study. Both data sets can also be found Stanford Network Analysis Project (SNAP) [32]. The CDR data set is a weighted edge list that consists of over 1.8 million number of nodes and almost 1.6 million number of edges.

B. Algorithm Selection

In order to implement the selected approaches in Section III over the unlabeled CDR data, we have picked two algorithms. Each of these algorithms are selected because they provide a unique advancements on top of the core approaches described and to take advantage of both lower-order and higher-order community detection. These algorithms are:

1) Label Propagation [30]: Each node in the network has a label attached that denotes its community membership. A node in the network joins a community based on the maximum number of neighbors that have a particular label. Therefore, the label propagates through the network quickly and at the end, each node will have been assigned a label. Consequently, nodes with the same label are clustered together as a member of one community. The time complexity of this algorithm is $O\ (m+n)$ where n and m are number of nodes and edges.

2) Edge enhancement approach for Motif-aware community detection (EdMot) [31]: This method fundamentally performs just as explained in Section III-B with measure improvements that other motif-based approaches neglect. That is, the fragmentation problem which is resulted from isolated nodes. This algorithm addresses this issue by deriving a clique from each partitioned module and rewiring the original network. This solution is shown to have increased the quality of higher-order community detection. The time complexity of this algorithm is $O\ (m^{1.5}+n\log n)$ where $O\ (m^{1.5})$ is the time required to find triangle motifs.

C. Data Input and Output

Data input and output format in the case study is made to be uniform for simplicity and consistency. A network that represents a data set is described as an edge list just as shown in Table I . After fitting the graph with the selected model, the output is a list of nodes that are indexed by a community id as shown in Table II.

TABLE I: Edge List Representation of a Social Network

From Node	To Node	Weight (if applicable)
1	6	290
2	7	79
3	545	388
4	210	12
...	...	...

TABLE II: Example Community Membership Output

Community id	Member nodes
1	1, 2, 4, 5...
2	3, 7, 8, 9...
3	6, 10, 11...
...	...

D. Ground Truth Definition

In our work, since we aim to discover communities from unlabeled CDR data set, we use intersectional communities as ground truth. Specifically, let C_L and C_M be a set of all communities discovered by *Label Propagation* and *EdMot* respectively, then $\forall C_l \in \mathbb{C}_\mathbb{L}$ and $\forall C_m \in \mathbb{C}_\mathbb{M}$, we compute the agreement ratio defined as:

$$Agreement(C_l, C_m) = \frac{\big|\,(C_l \cap C_m)\,\big|}{\big|\,(C_l \cup C_m)\,\big| - \big|\,(C_l \cap C_m)\,\big|} \times 100 \tag{1}$$

Subsequently, those common communities with an agreement threshold $\tau > 40\%$ are included in the set of intersectional ground truth communities C_G. Once ground truth is determined, output of each algorithm, C_L and C_M, is evaluated against C_G using a matrix as shown in Table III.

During evaluation (1) is applied by replacing C_L with C_G when evaluating C_M and vise versa.

Furthermore, (1) also solves the label assignment problem in discovered communities. That is, when evaluating community detection algorithms in such a way, there is high likelihood that *community id* assigned to the output of an algorithm having a high agreement ratio to a community in the ground truth labeled differently. For instance, sets of nodes labeled as C_2 from Table III, are highly similar to C_0 in the ground truth data. This is due to various approaches each algorithm uses and there is no computationally inexpensive way to control the label assignment problem.

TABLE III: Ground Truth Agreement Evaluation Matrix

		Algorithm Output		
		C_0 (3,4,7,8)	C_1 (2,3,4,5,7)	C_2 (1,2,4,5)
Ground Truth	C_2	14.28%	50%	100%
	C_1	50%	42.84%	12.5%
	C_0	16.66%	33.3%	40%

In this arbitrary example, in Table III, we assume nodes (1,2,4,5) $\in C_0$, (2,3,7,8,9) $\in C_1$ and (4,5,6) $\in C_2$ as the ground truth C_G.

E. Evaluation Metrics

We have summarized our result matrix for different algorithm comparisons using three metrics. These are of importance in order to measure algorithms success rate.

1) The number of correct communities: For a community detection algorithm output, an agreement ratio of p along a ground truth agreement comparison matrix column k (i.e., single community), we set a threshold value t, such that if $p_k > t$ then the community detected is accepted as successful. The total number of correct communities η from an algorithm is considered as the first metric.

2) The rate of correct communities per number of ground-truth communities: For a number of ground truth communities N^G and correctly identified communities η, the rate of correct communities discovered is $\rho = \frac{\eta}{N^G}$. This allows for precision evaluation.

3) The mean agreement rate of correct communities: For communities discovered above the threshold value t $(C_1, C_2, ..., C_n)$ with respective agreement rate of $(A_1, A_2, ..., A_n)$, we evaluate the mean agreement rate $\mu = \frac{(A_1, A_2, ..., A_n)}{n}$. This metrics permits us to measure the mean correctness of discovered communities. Even though identified communities are above the set threshold value, evaluating the average agreement rate against the intersectional ground truth communities summarizes an algorithms success well.

Moving forward for better presentation of results, we will use the symbols described in Table IV to represent the above metrics.

TABLE IV: Symbol Representation of Metrics

Symbol	Corresponding Metric
N^G	Number of intersectional ground truth communities
ν	Number of discovered communities
η	Number of correct communities
t	Correctness threshold
ρ	Rate of correct communities per number of ground-truth communities
μ	Mean agreement rate of correct communities

F. Performance Evaluation

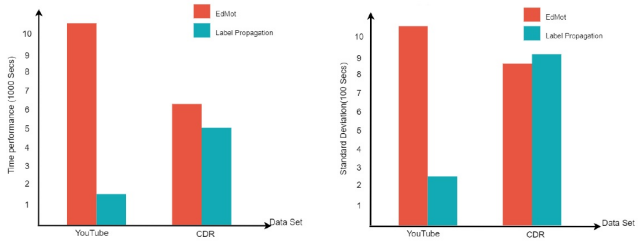
In order to run the following experiments, Amazon EC2 Linux version 5.3.0-1019 instance with 31GB of RAM was used.

TABLE V: Accuracy Performance

Metrics	YouTube		CDR*	
	EdMot	Label P.	EdMot	Label P.
N^G	4114		16908	
ν	9451	62790	253589	50589
$\eta,\ t > 50\%$	2754	4100	11882	16908
ρ	66.94%	99.65%	70.27%	100%
μ	76.47%	99.58%	66.74%	99.98%

* communities with number of nodes < 4 were removed.

Fig. 3 shows the time performance and standard deviation over 30 runs.



(a) Time Performance. (b) Standard Deviation.

Fig. 3: Time Performance Analysis for YouTube and CDR Data sets..

VI. Conclusion and Future Work

Community discovery in unlabeled real world data sets is subjected to a lot of uncertainty due to its inherit unsupervised nature. Although many promising advances are made continuously, there is room for improvement. In the scope of this study, in order to increase the quality of acquired results, we have shown that selective and context-driven methodology is necessary. And the results demonstrate intersectional ground truth can be used to strengthen the available community discovery algorithms by enforcing correctness through double approval scheme. For instance, from the 50,589 number of communities originally found by Label Propagation in the CDR data set, 16,908 of them were approved by EdMot with average intersectional ground truth agreement rate of 99,98%. This can be interpreted as it has successfully discovered 16,908 number of communities.

In the future, this work can be utilized in order to improve churn prediction models in the telecommunication sector by identifying influential entities.

ACKNOWLEDGMENT

We thank Intellica Business Intelligence Consultancy for providing us the CDR data set and for their continuous support in this case study. This study is supported by TUBITAK TEYDEB under the project ID 3181048.

REFERENCES

[1] A. Condon and R. M. Karp, "Algorithms for graph partitioning on the planted partition model", Random Struct Algor 18, pp.116–140, 2001.
[2] F. Altunbey and B. Alatas, "Overlapping Community Detection in Social Networks Using Parliamentary Optimization Algorithm", International Journal of Computer Networks and Applications (IJCNA) 2, no. 1, pp. 12–19, 2015.
[3] N. Du, B. Wu, X. Pei, B. Wang, and L. Xu, "Community detection in large-scale social networks", The 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD '07). Association for Computing Machinery, New York, NY, USA, pp. 16-–25, 2007.
[4] M. S. Khan, A. W. A. Wahab, T. Herawan, G. Mujtaba, S. Danjuma and M. A. Al-Garadi, "Virtual Community Detection Through the Association between Prime Nodes in Online Social Networks and Its Application to Ranking Algorithms", in IEEE Access, vol. 4, pp. 9614-9624, 2016.
[5] L. Runpeng, H. Jun and W. Xiaofan, "VCD: A network visualization tool based on community detection", 2012 12th International Conference on Control, Automation and Systems, JeJu Island, pp. 1221–1226, 2012.
[6] M. Crampes, M. Plantie, "A unified community detection, visualization and analysis method", Advances in complex systems, vol. 17, no. 01, pp. 1450001, 2014.
[7] J. David Cruz, C. Bothorel, and F. Poulet, "Community detection and visualization in social networks: Integrating structural and semantic information", ACM Trans. Intell. Syst. Technol. 5, 1, Article 11 pp. 26, 2013.
[8] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "A tale of one city: Using cellular network data for urban planning", IEEE Pervasive Computing, 10(4), pp. 18-–26, 2011.
[9] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: A survey of research", ACM Comput. Surv., 47(2):25:1–25:20, 2014.
[10] L. Alexander, S. Jiang, M. Murga, and M. C. Gonz´alez, "Origin–destination trips by purpose and time of day inferred from mobile phone data", Transportation Research Part C: Emerging Technologies, 58 pp. 240-–250, 2015.
[11] O. J¨arv, R. Ahas, E. Saluveer, B. Derudder, and F. Witlox, "Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records", PloS one, 7(11) pp. 1-–12, 2012.
[12] A. Bogomolov, B. Lepri, R. Larcher, F. Antonelli, F. Pianesi, and A. Pentland, "Energy consumption prediction using people dynamics derived from cellular network data", EPJ Data Science, 5(1):13, 2016.
[13] K. Kim, C. Jun, J. Lee, "Improved churn prediction in telecommunication industry by analyzing a large network", Expert Systems with Applications, vol. 41, Issue 15, pp. 6575–6584, 2014.
[14] Baeth, M.J. et al. (2019). Detecting misinformation in social networks using provenance data, CONCURR COMP-PRACT E, 31(3).
[15] Baeth M. J. et al. (2018) An approach to custom privacy policy violation detection problems using big social provenance data, CONCURR COMP-PRACT E, 30(21).
[16] Baeth, M.J. et al. (2017). Detecting misinformation in social networks using provenance data, SKG-17.
[17] Baeth, M.J. et al. (2015). On the Detection of Information Pollution and Violation of Copyrights in the Social Web, SOCA-15.
[18] Dundar, B. et al. (2016) A Big Data Processing Framework for Self Healing Internet of Things Applications, SKG-16.
[19] Aktas, M.S. et al. (2019), Provenance aware run-time verification of things for selfhealing Internet of Things applications, CONCURR COMP-PRACT E, DOI: 10.1002/cpe.4263.
[20] Aktaş M.S., (2018) Hybrid cloud computing monitoring software architecture, CONCURR COMP-PRACT E, 30(21).
[21] Riveni, M. et al. (2019). Application of provenance in social computing: A case study, CONCURR COMP-PRACT E, 31(3).
[22] Tas, Y. et al. (2016) An Approach to Standalone Provenance Systems for Big Provenance Data, SKG-16.
[23] Newman, Mark EJ. "The structure and function of complex networks." SIAM review 45.2, pp. 167–256, 2003.
[24] C. Michele, F. Giannotti, and D. Pedreschi. "A classification for community discovery methods in complex networks." Statistical Analysis and Data Mining: The ASA Data Science Journal 4.5, pp. 512-546, 2011.
[25] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks", Science, vol. 298, no. 5594, pp. 824—827, October 2002.
[26] L. Huang, C. Wang, H. Chao, "Higher-Order Multi-Layer Community Detection", The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), pp. 9945–9946, 2019.
[27] S. Yu, J. Xu, C. Zhang, F. Xia, Z. Almakhadmeh and A. Tolba, "Motifs in Big Networks: Methods and Applications", in IEEE Access, vol. 7, pp. 183322–183338, 2019, doi: 10.1109/ACCESS.2019.2960044.
[28] A. Benson, D. Gleich, and J. Leskovec, "Higher-order organization of complex networks", Science 353, 6295, pp. 163-–166 2016.
[29] L. Huang, C. Wang, and H. Chao, "A Harmonic Motif Modularity Approach for Multi-layer Network Community Detection" IEEE International Conference on Data Mining, ICDM, Singapore, November 17-20, pp. 1043-—1048, 2018.
[30] U. N. Raghavan, R. Albert, and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks,", Physical Review E, vol. 76, p. 036106, 2007.
[31] P. Li, L. Huang, C. Wang, and J. Lai, "EdMot: An Edge Enhancement Approach for Motif-aware Community Detection.", The 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '19), Association for Computing Machinery, New York, NY, USA, pp. 479-–487, 2019, DOI:https://doi.org/10.1145/3292500.3330882
[32] Stanford Network Analysis Project snap,https://snap.stanford.edu/,Accessed: 2020-05-25
[33] Dash framework, https://dash.plotly.com/, Accessed: 2020-05-25
[34] Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD, "Cytoscape.js: a graph theory library for visualisation and analysis", Bioinformatics, 32, (2), pp.309–311, 2015
[35] Jaewon Yang, Jure Leskovec, "Defining and evaluating network communities based on ground-truth", Knowl. Inf. Syst. 42, 1 (January 2015), 181–213. DOI:https://doi.org/10.1007/s10115-013-0693-z
[36] B. Rozemberczki, C. Allen and R.Sarkar, "Multi-scale Attributed Node Embedding", 1909.13021, cs.LG, 2019.