Using Triads to Identify Local Community Structure in Social Networks

Justin Fagnan Osmar Zaïane Denilson Barbosa
University of Alberta, Canada
Email:{fagnan, zaiane, denilson}@ualberta.ca

Abstract

We present our novel community mining algorithm that uses only local information to accurately identify communities, outliers, and hubs in social networks. The main component of our algorithm is the T metric, which evaluates the relative quality of a community by considering the number of internal and external triads (3-node cliques) it contains. Furthermore we propose an intuitive statistical method based on our T metric, which correctly identifies outlier and hub nodes within each discovered community. Finally, we evaluate our approach on a series of ground-truth networks and show that our method outperforms the state-of-the-art in community mining algorithms.

I. Introduction

Data can often be expressed in the form of an Information Network that stores the entities, represented as nodes, and the relationships between them, represented as edges. One can mine these networks by employing a variety of techniques that identify communities, which are tightly-knit groups of nodes that interact more within the group than outside of the group.

The challenge of how to detect such communities has been a central problem studied in the field of social network analysis in the past two decades. To this end, researchers have proposed a variety of techniques that discover communities by considering the entire network structure, that is, they require global knowledge of the network [1], [2], [3]. Unfortunately, they realized that these global techniques do not scale well when considering extremely large information networks, such as Facebook or the World Wide Web, which are becoming increasingly popular and contain hundreds of millions or billions of nodes [1].

To remedy this problem, researchers have recently proposed local methods that detect communities by only considering local information and therefore are not sensitive to the size of the network [4], [5], [6]. These local methods generally require some metric that determines the relative quality of a community, and indeed, many such metrics have been proposed. However, the existing metrics often suffer from poor outlier detection [4], [7] and the discovery of incorrect communities in simple ground truth networks [5], [8].

In this paper we aim to solve both of these problems by presenting our T metric, which defines the relative quality of a community by considering the number of internal and external triads it contains. We apply our T metric within a modified version of Clauset's local framework [4] to greedily discover communities while achieving more accurate outlier and hub detection when compared to previous approaches. We also show that our framework, combined with the T metric, leads to increased accuracy on a variety of ground truth networks when compared to the existing techniques.

II. Related Work

A variety of community mining techniques have been proposed that employ either the divisive or agglomerative framework to detect communities using global information [9]. The most well-known of these approaches is Newman's Q-Modularity [2] metric which considers the number of edges within a community minus the expected number of such edges in a random network. However, Fortunato and Barthelemy [10] have shown that Modularity-based metrics suffer from a resolution limit, in that they cannot detect communities smaller than some threshold. Furthermore, it is unclear how to detect outlier nodes with Modularity-based methods.

In addition, researchers have realized that it is computationally intractable to consider global information for many of the large scale networks that they wish to analyze [1]. To address this concern, Clauset [4] introduced his local community mining framework that explores the network through local expansion and thus is not sensitive to the network size. His method requires a metric to determine the quality of each discovered community and a variety of such metrics have been proposed, including Clauset's own R metric, the M metric from Luo et al., and the L metric from Chen et al. [4], [7], [5]. All of these metrics evaluate a community by considering how edges are distributed within the community relative to outside of the community. These metrics, however, fail to accurately identify outliers and achieve low scores on many ground truth networks, as shown in our evaluation section.

Palla et al. [6] have also proposed their Clique Percolation Method which identifies communities by rolling a k-clique around the network until it is unable to reach any unexplored nodes. The nodes covered while rolling are considered the discovered community and then the process continues on a different section of the network. Although their algorithm benefits from being local and intuitive, it is also very sensitive to the parameter choice for k -i.e. the size of the clique, and thus the algorithm can be difficult to apply in practice.

We find that our approach lies somewhere between Clauset's Local Framework and the Clique Percolation Method, in that our T metric favours communities that contain triads (cliques

of size 3), but it discovers these communities through local expansion making it scalable for very large networks.

III. Our Approach

Our approach is a two stage algorithm that first detects communities by applying our T metric within the local community framework and then employs an additional stage to identify outliers/hubs in the discovered community.

The local community framework we apply in this paper has been adapted from Clauset's local framework [4] and can be summarized as follows. First, we initialize the community with a single node and place all of its neighbours into the shell set. Then, for each iteration, we greedily select the node from the shell set that, when included in the community, maximizes the T metric. We add this selected node to the community and all of its neighbours to the shell set. This process continues until there are no nodes in the shell set which would further maximize the T metric. At this point, a community is discovered and the algorithm restarts on another node in the network. In order to prevent overlapping communities we ensure that all nodes which are assigned to one community cannot belong to any other community. A depiction of the shell set is shown in Figure 1. Our major deviation from Clauset's original framework is that we do not keep track of the boundary set. Also, as we will explain later, we have added an additional stage to the framework that detects both outliers and hubs.

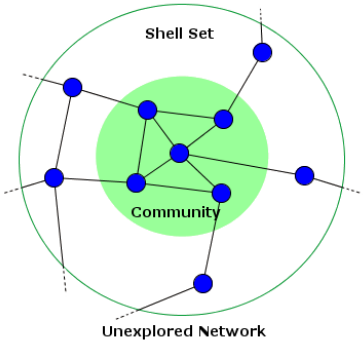


Fig. 1. A depiction of the local algorithm.

It is important to note that the selection of the starting node for the local framework can have a dramatic effect on the accuracy of the algorithm. In particular, our evaluation in Section 4 shows that randomly selecting the starting node can result in very poor community structure and accuracy. We hypothesize that a good starting node will have a high degree because it allows for a large neighbourhood to be considered in the first iteration of the framework. However without global knowledge it is impossible to select the optimal starting node, as we do not know the degree of every node. To remedy this we propose a local approach, that first chooses a node at random and then explores its immediate neighbourhood and selects the node with the highest degree. This node becomes our starting node. Note that we have used the degree metric, instead of PageRank or Betweenness, because it can be computed locally without consulting the entire network. We briefly considered other metrics, such as Estimated Closeness [11], but they did not outperform the degree metric in our test cases.

A. Local Community Metric T

Given the local community framework we can see that the role of the T metric is to determine whether or not a node should be included in the community. Intuitively, our metric favours nodes that form many triads with nodes within the community and few triads with nodes outside of the community. We define these quantities as T_{in} and T_{ex}, respectively. We define a triad as a collection of three nodes that are fully connected, aka, a 3-node clique. Our intuition is that all members of a triad are tightly bonded together and thus are more likely to belong to the same community. More formally, we present our T metric as:

$$T = T_{in} * T_{diff}$$

Where

$$T_{diff} = \begin{cases} T_{in} - T_{ex} & \text{if } T_{in} \geq T_{ex} \\ 0 & \text{otherwise} \end{cases}$$

$$T_{in} = \frac{1}{6} * \sum_{i \in C, j \in C, k \in C} A_{i,j} * A_{j,k} * A_{i,k}$$

$$T_{ex} = \frac{1}{2} * \sum_{i \in C, j \in S, k \in S} A_{i,j} * A_{i,k} * A_{j,k}$$

Where C is the set of nodes in the community, S is the set of nodes in the shell set, and A is the adjacency matrix such that $A_{i,j}$ is 1 if nodes i and j share an edge. We divide the T_{in} score by 6 to prevent double counting all permutations of the same triad, for example, 'ABC', 'ACB','BCA', 'BAC', 'CAB', and 'CBA' all refer to the same triad between nodes A, B, and C. For T_{ex} we only divide by 2 because the limitation that $i \in C$ reduces the number of permutations.

We have bounded T_{diff}, and thus T, to be non-negative because all of the nodes in the initial stages of the community will belong to more external triads than internal ones. If left unbounded, this would result in a negative T_{diff} score that would penalize well connected nodes; yet these are the very nodes that we believe should be included first. Thus, we set the T_{diff} score to zero in these cases and let the tie-breaking step determine the best node.

This tie-breaking step is a critical part of the metric because there are many cases where multiple nodes result in the same T score, yet are qualitatively different. For example, consider a node X that when included in the community has a T_{in} score of 49, T_{ex} score of 48, and thus a T score of 49. Also consider a node Y that has a T_{in} score of 7, T_{ex} score of 0, and thus also a T score of 49. Clearly, node Y is a better choice to include in the community because it directly contributes to the internal score without a negative influence on the external

score. We capture this intuition by always selecting the node with the lowest T_{ex} score in the event of a tie.

It is important to note that we are not considering triads that have two nodes in the community and one node in the shell set. This is because such a triad could be classified as either external or internal depending on whether the target node is chosen to be included in the community, or placed back into the shell set. Thus it does not make intuitive sense to assign this triad to either set.

Furthermore we are aware that our metric is dissimilar from many of the existing approaches in that it does not try to maximize a ratio of internal to external scores. This is because we feel that the difficulty associated with dividing by zero results in a biased metric that favours nodes with no external relations. For example, consider a metric that counts the number of edges. Also consider two nodes in the shell set: one with 2 internal edges, 0 external edges, and one with 10 internal edges, 1 external edge. If the ratio of internal to external edges of the community is 100:10, then the first node will be included, but the second will not. We find this approach to be counter-intuitive, especially given that as the ratio score of the community increases, so does the idiosyncrasy of such examples.

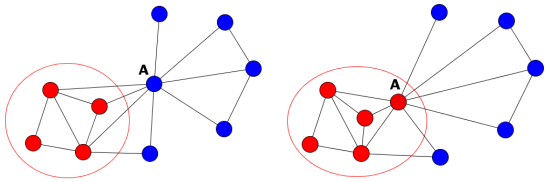
B. Incremental Formula

Although the formulae given above are relatively simple, it would be computationally demanding to count the number of triads every time a node is considered, thus we also present an incremental formula for computing the T_{in} and T_{ex} scores based on the previous scores. More formally:

$$T_{in}\prime = T_{in} + \frac{1}{2} \sum_{\substack{i \in N(X) \\ j \in N(X) \\ i \neq j}} A_{j,i} * C_i * C_j$$

$$\begin{aligned} T_{ex}\prime = T_{ex} \\ + \frac{1}{2} \sum_{\substack{i \in N(X) \\ j \in N(X) \\ i \neq j}} A_{j,i} * (1 - C_i) * (1 - C_j) \\ - \sum_{\substack{i \in N(X) \\ j \in N(X) \\ i \neq j}} A_{j,i} * (1 - C_i) * C_j \end{aligned}$$

Where T_{in} and T_{ex} are the scores before including node X in the community, $N(X)$ is the neighbourhood of node X, $A_{i,j}$ is the adjacency matrix, and C_n is 1 if n is in the community, 0 otherwise. Here, the last term in $T_{ex}\prime$ represents the number of triads that contain one node in the community and one node outside of the community. These triads are discounted because they were considered external triads prior to node X being included in the community, but now are considered uncounted triads. Note that we divide the second term in $T_{ex}\prime$ by 2 to avoid double counting both permutations of the same triad. A visual example of the incremental formula can be seen in Figure 2.



(a) Before including node A. T_{in} = 2, T_{ex} = 1.

(b) After including node A. $T_{in}\prime$ = 2 + 2 = 4, $T_{ex}\prime$ = 1 + 2 - 1 = 2.

Fig. 2. An example of the incremental T calculation. Nodes within the circle are part of the community.

C. Outlier and Hub Detection

Although our T metric is used to identify communities, it does not directly solve the problems of pruning outliers from these communities or detecting hubs. The notion of an outlier can be summarized as a node that is weakly connected to the community but does not belong to any other community. Whereas a hub refers to a node that is strongly connected to many communities, without truly belonging to any individual community. To identify such nodes we have added an additional stage to the local framework that further processes each community after it has been discovered. In particular, we iterate through the entire community and record the number of internal triads that each node belongs to. We then compute the average, T_{inAvg}, and standard deviation, T_{inStd}, of this score.

While iterating through the community we label a node as a hub if it participates in more external than internal triads; which follows from the observation that this node may belong to many other communities. However, it is not sufficient to detect a hub by only considering a single community. Thus, we allow nodes with the 'hub' label to join more than one community. This way if two or more communities label the same node as a hub, then it must be a true hub. On the other hand, if only a single community labels it as a hub then, by definition, it cannot be a hub and we remove its label.

To detect outliers we rely on the statistical distribution of the internal triads in the community. More specifically, a node is an outlier if it satisfies the following criteria:

$$\begin{aligned} T_{in}(X) &< \lfloor T_{inAvg} - T_{inStd} \rfloor \\ T_{ex}(X) &= 0 \end{aligned}$$

Where $T_{in}(X)$ is the number of internal triads that node X participates in and likewise for $T_{ex}(X)$. We believe that this definition best captures the intuitive understanding of an outlier, in that any node participating in significantly fewer triads than the average must be a weak member of that community. We have opted to use only one standard deviation

based on our empirical analysis. We should point out that there are a variety of well-known statistical approaches to determine outliers, such as those proposed by Chauvenet, Grubbs, or Peirce [12], [13], [14]. Unfortunately we could not apply these methods as their assumption that the data is distributed normally does not hold in our scale-free social networks.

IV. Evaluation

To rigorously evaluate our proposed framework we have compared it to a variety of popular community mining algorithms on a series of well-known ground truth networks. We have employed the Adjusted Rand Index (ARI) to compute a quantitative score that indicates how closely the results returned by each algorithm match that of the ground truth. More specifically, this index compares two sets of results and returns a score that ranges from 0, which indicates a completely random match, and 1, which indicates a perfect match.

We performed an evaluation against all known ground-truth networks, which are summarized in Table I.

Of particular interest to us are the NCAA Football network, which contains many small communities and the Political Blogs network, which contains over 1000 nodes in two very large communities. We expect that many algorithms will have difficulty capturing both the small and large scale communities. Furthermore, we note that the Strike and Karate networks each contain one node that shares a single edge with both ground truth communities. Thus, using only the information in the network it is impossible to assign these nodes to the correct community every time. To prevent 'lucky' selections from biasing the results we have allowed these two nodes to belong to either of the communities without any penalty to the Adjusted Rand Index (ARI) score. The algorithms we compare our T metric against include:

MaxMin Modularity This is an agglomerative algorithm proposed by Chen et al. [3] as an improvement over Newman's Q-Modularity-based approach in that it also considers the number of unrelated node pairs within the community.

Clique Percolation Modularity (CPM) Please see the section on Related Work. For our evaluation we have selected the best result between the parameter value of K = 3, 4, and 5.

Local L This is a local algorithm proposed by Chen et al. [5] that employs Clauset's local framework and the L metric to discover communities by maximizing the ratio of internal average degree over external average degree.

Local R This is the original local algorithm proposed by Clauset [4] that tries to maximize the number of edges leading from the boundary set of a community to its core and minimize the number of external edges. There is no outlier detection.

Local M This is also a local algorithm proposed by Luo et al. [7] that employs Clauset's framework and the M metric to discover communities by maximizing the ratio of internal edges over external edges. There is no outlier detection.

When evaluating our T metric we also want to determine what the optimal strategy is for selecting a starting node in the local framework. Thus we present two evaluations. In the first, our framework selects the starting node by exploring the local neighbourhood of a randomly selected node and choosing the one with highest degree. In the second, our framework simply selects a starting node at random. To mitigate the effects of this randomness we have run each local algorithm ten times and reported the average score of these runs. We hope to show our proposed approach for selecting the starting node is significantly better than the current random approach.

The results of our evaluation are summarized in Table II, which contains the Adjusted Rand Index (ARI) scores, and Table I, which contains the number of detected communities.

As we can see in Table II, our T metric matches or outperforms the existing algorithms on nearly every ground truth network, with the exception of the Mexican Politics network. In fact we notice that all of the algorithms perform very poorly on this network.

Furthermore, in Table I we notice that our algorithm is the only method which identifies the correct number of communities for a majority of the ground truth networks. We feel that this is an important evaluation tool in that ARI score can often be misleading when we don't consider the number of communities. This is exemplified by the fact that the M metric achieved a reasonably high ARI score in the Political Blogs network, even though it detected a multitude of extraneous communities. Additionally, we can see that our outlier detection method performed exceptionally well given that it accurately identified all of the outliers and all but one of the hubs in the NCAA Football Network. More importantly, contrary to Local L, our method did not identify any outliers or hubs in the other outlier-free networks.

Finally, our method of selecting the starting node appears to be somewhat better than the random approach when applied to our T metric. We can see noticeable improvements when using our approach on denser networks such as Political Blogs and Political Books. In addition to this evaluation we have also applied the L metric within our proposed framework to determine if it is our framework that provides the increased accuracy, or if it is the T metric itself. We hypothesized that perhaps our outlier and hub detection stage was responsible for our excellent results. This was not the case, and our results, which we do not present here for the sake of brevity, indicate that the L metric performs very poorly in our framework. Thus, we are more confident in claiming that the performance of our algorithm is largely attributable to our T metric.

V. Conclusion and Future Work

In this paper we have presented a modified local framework that employs our novel T metric, which considers the number of internal and external triads in a community. In addition, we detailed our modifications to Clauset's local framework in order to achieve improved outlier detection and better starting node selection when compared to previous approaches. We performed a rigorous evaluation against a variety of existing community mining algorithms and showed that our method outperforms all of these algorithms on a variety of ground truth networks. Furthermore we showed that our starting

TABLE I
AN OVERVIEW OF THE GROUND TRUTH NETWORKS USED IN OUR EVALUATION AND THE NUMBER OF COMMUNITIES DETECTED BY EACH ALGORITHM.

Ground Truth Network	Number of Nodes	Number of Edges	Number of Communities	MaxMin	CPM	Local L	Local R	Local M	Local T
Zachary's Karate Club [15]	34	78	2	2	3	6	3	3	2
Strike [16]	24	38	3	3	6	5	5	3	3
Political Blogs [17]	1224	19087	2	-	-	94	20	66	3
Political Books [18]	105	441	3	2	4	10	7	4	3
Mexican Politics [19]	35	117	2	3	1	3	3	2	1
NCAA Football [20]	180	788	11 + outliers + hubs	5	12	12	13	11	11+ outliers + hubs

TABLE II
EVALUATION RESULTS. FOR THE X/Y CELLS, X INDICATES THE AVERAGE SCORE WHEN SELECTING THE STARTING NODES WITH THE MAXIMUM LOCAL DEGREE, AND Y INDICATES THE AVERAGE SCORE WHEN RANDOMLY SELECTING THE STARTING NODES. A DASH INDICATES THAT THE ALGORITHM DID NOT COMPLETE WHEN PROCESSING THE NETWORK.

	MaxMin	CPM	Local L	Local R	Local M	Local T
Zachary's Karate Club	**1**	0.15	0.32	0.52	0.47	**1** / 0.9
Strike	**1**	0.36	0.37	0.71	0.76	**1** / 1
Political Blogs	-	-	0.06	0.62	0.66	**0.88** / 0.65
Political Books	0.64	0.63	0.22	0.55	0.57	**0.66** / 0.57
Mexican Politics	**0.36**	0.14	0.09	0.19	0.3	0
NCAA Football	0.15	0.983	0.96	0.28	0.28	**0.996** / 0.94

node selection method is superior to the current practice of randomly selecting a node, and that our outlier detection can correctly detect outliers, or the absence of, in all of the ground truth networks we evaluated.

Future work includes investigating how our metric can be applied to weighted networks by, for example, computing the weight of a triad as the average of its edge weights. One could also consider signed networks by utilizing the signed triads approached proposed by Leskovec et al. [21]. Alternatively, our metric could be applied to directed networks, although it is currently unclear what a directed triad would mean in this context. We also believe that our work can easily be extended to support the detection of overlapping communities by adding nodes back into the network after they have been placed into a community. Unfortunately there is no overlapping ground truth and thus such an approach can only be evaluated in theory.

Finally, as discussed in the paper, we believe that deciding how to select a starting node in the local framework is an entirely open problem that warrants further study. In particular, we are interested in discovering what structural properties of node result in it being a good seed node for a community and how these properties are related to existing metrics, such as PageRank, Degree, or Betweenness. Perhaps possible insights can be gained by studying how database researchers have solved the similar problem of selecting cluster centroids; but it is not immediately obvious that these techniques could easily be applied within the context of community mining.

REFERENCES

[1] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, p. 026113, Feb 2004.

[2] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Phys. Rev. E*, vol. 69, no. 6, p. 066133, Jun 2004.

[3] J. Chen, O. Zaïane, and R. Goebel, "Detecting communities in social networks using max-min modularity," in *SIAM International Conference on Data Mining (SDM'09)*, 2009.

[4] A. Clauset, "Finding local community structure in networks," *Phys. Rev. E*, vol. 72, no. 2, p. 026132, Aug 2005.

[5] J. Chen, O. R. Zaïane, and R. Goebel, "Local community identification in social networks," in *ASONAM*, 2009, pp. 237–242.

[6] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, p. 814, Jun 2005.

[7] F. Luo, J. Z. Wang, and E. Promislow, "Exploring local community structures in large networks," in *in WI 06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, 2006, pp. 233–239.

[8] M. Rosvall and C. T. Bergstrom, in *Proc. Natl. Acad. Sci*, vol. 105, 2008, pp. 1118–1123.

[9] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75–174, 2010.

[10] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, January 2007.

[11] D. Eppstein and J. Wang, "Fast approximation of centrality," *Proceedings of the twelftth annual ACM-SIAM symposium on Discrete algorithms*, 2001.

[12] W. Chauvenet, "A manual of spherical and practical astronomy v. ii." in *Technometrics 11*, 1863.

[13] F. E. Grubbs, "Procedures for detecting outlying observations in samples," in *Technometrics 11*, 1969, pp. 1–21.

[14] B. Peirce, "Criterion for the rejection of doubtful observations," in *Astronomical Journal II 45*, 1852.

[15] W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

[16] J. Michael, "Labor dispute reconciliation in a forest products manufacturing facility," in *Forest Products Journal*, vol. 47, 1997, pp. 41–45.

[17] L. Adamic and N. Glance, "The political blogosphere and the 2004 us election," in *in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

[18] V. Krebs, "Unpublished," in *http://www.orgnet.com/*, accessed in 2010.

[19] J. Gil-Mendieta and S. Schmidt, "The political network in mexico," in *in: Social Networks 18*, vol. 4, 1996, pp. 355–381.

[20] X. Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "Scan: a structural clustering algorithm for networks," in *in: KDD*, 2007, pp. 824–833.

[21] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the 28th international conference on Human factors in computing systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 1361–1370. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753532