Mining Anomalous Sub-graphs in Graph Data Using Non-negative Matrix Factorization

N.N.R. Ranga Suri[1], Musti Narasimha Murty[2], and Gopalasamy Athithan[1,3]

[1] Centre for AI and Robotics (CAIR), Bangalore, India
{rangasuri,athithan.g}@gmail.com
[2] Dept of CSA, Indian Institute of Science (IISc), Bangalore, India
mnm@csa.iisc.ernet.in
[3] Presently working at Scientific Analysis Group (SAG), Delhi, India

Abstract. Mining graph data has been an important data mining task due to its significance in network analysis and many other contemporary applications. Detecting anomalies in graph data is challenging due to the unsupervised nature of the problem and the size of the data itself to be dealt with. Recent research efforts in this direction have explored graph data for identifying anomalous nodes and anomalous edges of a given graph. However, in many real life applications where the data is inherently networked in nature, the requirement is to detect anomalous sub-graphs with distinguishing characteristics such as *near cliques*, etc. In this context, we propose a novel method for addressing the anomalous sub-graph mining problem through community detection by employing the non-negative matrix factorization technique. Anomalous sub-graphs are identified by applying some existing techniques on the detected communities for measuring their deviation from the normal characteristics. We demonstrate the effectiveness of the proposed method through experimental evaluation on various benchmark graph data sets.

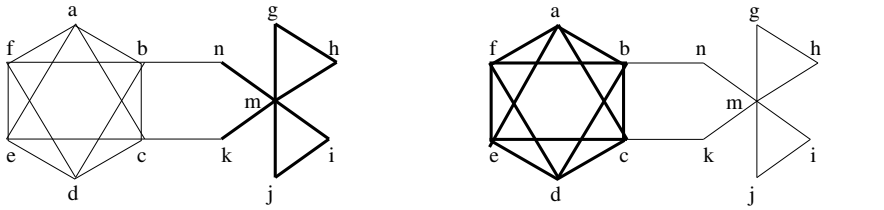
Keywords: Data mining, Mining graph data, Anomalous sub-graphs, Community detection.

1 Introduction

An emerging research problem related to graph mining is to discover anomalies, also known as outliers, in graph data [2,10,3]. The objective is to identify the sub-graphs displaying anomalous characteristics in the input graph. This problem assumes significance as finding out a close group of individuals in a social network (forming a clique like pattern) or identifying minimal interactions among a set of nodes in a communication network (forming a tree like pattern) helps in further analysis of this sub-set data in an objective manner. Thus, anomaly detection in graph data turns out to be an important mining task.

A recent work on anomaly detection in graph data [1], makes use of the notation of *egonet* of a node defined as the induced sub-graph of its 1-step neighbors as shown in Fig. 1(a). According to this work, it is found that the number of

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 88–93, 2013.



(a) Egonet of node m (bold lines) (b) An arbitrary sub-graph (bold lines)

Fig. 1. Examples of sub-graphs of different types

nodes and the number of edges of a normal egonet satisfy a power law relationship. As a result, the anomalous egonets (*near cliques* and *near stars* as defined in [1]) are identified by employing the established power law relationship.

It is clear that this result is applicable only to identify anomalous sub-graphs that are egonets structurally. However, there can exist various anomalous sub-graphs which are arbitrary sub-graphs of a given graph, not necessarily being sub-graphs of the egonet type. For example, an arbitrary sub-graph as shown in Fig. 1(b) qualifies to be an anomalous sub-graph (near clique) that is not an egonet structurally.

The above discussion clearly motivates the strong need to develop a generic approach to identify anomalous sub-graphs. In this paper, we propose a novel method for identifying anomalous arbitrary sub-graphs of a given graph. Similar to [1], the novel method is intended to detect *near clique* and *near tree* types of anomalies. In contrast to a near star anomaly, a near tree anomaly represents a rooted hierarchical structure, which is not an egonet. Firstly, the set of all possible connected sub-graphs is determined through community detection [4] by employing the Non-negative Matrix Factorization (NMF) technique [5]. The rationale behind this task is that the sub-graphs identified through community detection are considered to be the best candidates for detecting anomalies in graphs. As listing out all possible connected sub-graphs of a large graph leads to combinatorial explosion, the same is achieved through community detection. Then, each one of these sub-graphs is subjected to the power law relationship based anomaly detection procedure [1].

The following section gives a brief summary of various methods developed for anomaly detection in graph data. Section 3, describes the novel method being proposed for mining anomalous sub-graphs. Details on the experimental evaluation of the proposed method along with the results obtained are furnished in Section 4. Finally, Section 5 concludes the paper with a discussion on the proposed method along with a few directions for future work.

2 Related Work

There have been some research efforts made in the recent past for detecting various types of anomalies in graph data. The method proposed in [8] is an early one

using the Minimum Description Length (MDL) principle. The main idea of this method was that sub-graphs containing many common sub-structures are generally less anomalous than sub-graphs with few common sub-structures. Thus, it is suitable mainly for applications involving many common sub-structures such as the graphs describing the atomic structure of various chemical compounds. Similarly, the method proposed in [9] is meant for anomalous link (edge) discovery defining a novel edge significance measure. More recently, a method named *OddBall* was proposed [1] for detecting anomalous sub-graphs using the egonet notation, as depicted in Fig. 1(a). According to this method, the number of nodes N_i and the number of edges E_i of the egonet G_i of a node i follow a power law relationship, named the Egonet Density Power Law (EDPL), defined as

$$E_i \propto N_i^{\alpha}, \quad 1 \leq \alpha \leq 2. \tag{1}$$

Consequently, anomalous egonets (near cliques and near stars) are determined by measuring the amount of deviation from the power law relationship.

As discussed in the previous section, detecting communities in graph data is an important related graph mining task. According to the literature, communities in a graph represent groups of nodes with dense connectivity within each group [4]. Among the existing methods for detecting communities in graph representation of network data, the method proposed in [11] is a recent one employing the Non-negative Matrix Factorization (NMF) technique due to its powerful interpretability and close relationship with the clustering methods. The NMF algorithm [5] was basically proposed for finding factors of a matrix as $G \approx WH$, highlighting the philosophy of 'parts-based representation of objects'. Sparseness of the data is another important aspect managed effectively by the NMF technique. Thus, the same is employed here for identifying various connected sub-graphs of the input graph.

3 Proposed Method

As stated earlier, the objective is to identify various sub-graphs with anomalous characteristics in a given graph. This is achieved through first determining various candidate sub-graphs in the input graph and then putting them through an anomaly detection procedure as per the scheme shown in Fig. 2.

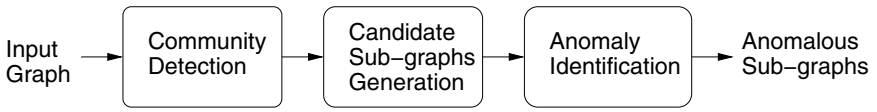


Fig. 2. Proposed scheme for detecting anomalous sub-graphs

As mentioned in Section 2, non-negative matrix factorization (NMF) [5] is a promising method for community detection [4] in graph data. Accordingly, the

adjacency matrix (G) corresponding to the input graph is subjected to community detection using the NMF procedure [5] by assigning a user specified value for the number of communities (indicated by k) to be detected. This results in a mapping W (given by $G \approx WH$) of the N nodes in G to various communities ($\{G_1, G_2, ..., G_k\}$) with varying degree of membership. Here, a membership value $w_{ij} \in W$ indicates the membership of the node n_i in G to the community G_j (j^{th} column of W). Applying a suitable threshold value (indicated by t) on the membership values (elements of W), crisp communities (sub-graphs) are determined. Let N_i be the number of nodes and E_i be the number of edges present in the sub-graph G_i. We now define the following two constraints to prune the set of crisp sub-graphs producing the candidate set of l ($0 \leq l \leq k$) sub-graphs.

(A) *Sub-graph non-triviality constraint*: Minimum number of nodes required in a non-trivial sub-graph is set as 3, i.e., $N_i \geq 3$.
(B) *Sub-graph connectivity constraint*: Minimum number of edges present in a connected sub-graph, i.e., $E_i \geq (N_i - 1)$.

For exploring the anomalous characteristics of the l candidate sub-graphs satisfying the above constraints, the power law relationship given in Equation 1 is considered here. In the context of mining anomalous arbitrary sub-graphs, this relationship is renamed as Sub-graph Density Power Law (SDPL) to reflect the property satisfied by the arbitrary sub-graphs of the input graph. Accordingly, a scatter plot with E_i versus N_i on the log-log scale is produced depicting the l sub-graphs along with the least squares fitting line. Thus, the outlierness score of a sub-graph G_i is computed as the distance to the fitting line as defined in [1].

$$out\_score(G_i) = \frac{max(E_i, CN_i^{\alpha})}{min(E_i, CN_i^{\alpha})} log(|E_i - CN_i^{\alpha}| + 1) \tag{2}$$

Finally, the anomalous sub-graphs are indicated in the scatter plot showing their deviation from the fitting line. A summary of various computational steps involved in the proposed method is furnished in Algorithm 1.

Algorithm 1. Mining anomalous sub-graphs in graph data.

Input: A graph with N nodes and E edges given by its adjacency matrix G.
Output: List of top ranked anomalous sub-graphs.

1: Factorize the adjacency matrix using the NMF procedure as $G \approx WH$.
2: Determine k communities $\{G_1, G_2, ..., G_k\}$ using W matrix.
3: Apply the membership threshold t on W to generate crisp communities.
4: Apply the sub-graph constraints A and B to determine the candidate set.
5: Produce the E_i versus N_i scatter plot along with least squares fitting line.
6: Measure the deviation of each candidate sub-graph G_i using Equation 2.
7: Obtain a ranked sequence of the sub-graphs as per their outlier scores.

4 Experimental Evaluation

For the purpose of evaluating the proposed method, we have considered certain real life graph data sets from the SNAP repository [6]. As the basic task here is

to detect the communities determined by the connectivity structure, the graphs employed in this experimentation are undirected and without any weights on the edges, as per the details furnished in Table 1.

Table 1. Details of the experimentation with the SNAP graph data sets

Input Details			Parameter Values		Observations	
Data Set Name	# Nodes (N)	# Edges (E)	# Comm. (k)	Membership threshold (t)	# Candidate sub-graphs (l)	Highest score
Facebook-107	1,034	53,498	100	0.01	100	4.1549
AS20000102	6,474	13,233	500	0.001	50	2.6012

The initial experiments were carried out on the Facebook graph [7] corresponding to the node '107', the largest node graph. This graph was subjected to the proposed sub-graph anomaly detection method resulting in the scatter plot as shown in Fig. 3(a), as in [1]. This plot represents the anomalous sub-graphs in the form of triangle points, with the area of each triangle point representing the amount of deviation of the corresponding sub-graph, as measured using Equation 2. Triangle points below the fitting curve (solid line) indicate near tree anomalies, while the ones above the fitting curve indicate near clique anomalies.

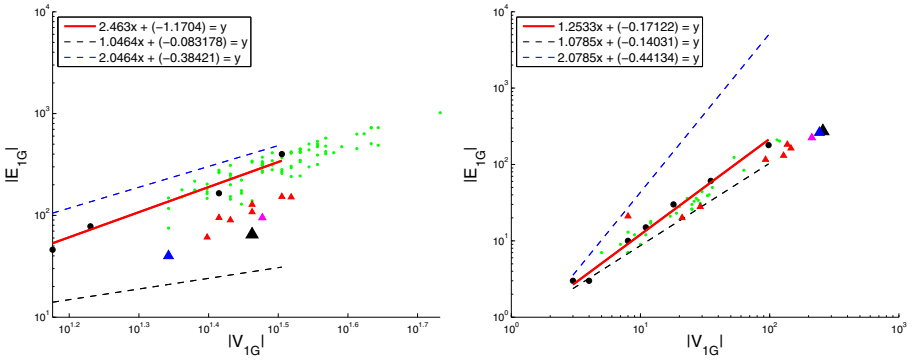
Similar experimentation was carried out on the 'AS20000102' graph from the Autonomous Systems collection [6] producing the results as shown in Fig. 3(b).

Table 1 provides the details on the values set for the parameters of the proposed algorithm (k and t) in this experimentation. It also indicates the number of candidate sub-graphs surviving after applying the proposed two constraints on the crisp communities resulted through the threshold application on the community membership values. Also shown is the highest anomaly score obtained corresponding to each graph data set. Referring to the input graph structure, it was found that many of the anomalous sub-graphs detected in this process are not egonets structurally, demonstrating the merit of the proposed method.

5 Conclusion and Future work

A novel method for mining anomalous sub-graphs in graph data has been proposed here through community detection by employing the NMF technique. The proposed method identifies anomalous sub-graphs by subjecting the detected communities to the power law constraints defined in the OddBall method. While the OddBall method can detect anomalous sub-graphs that are egonets structurally, the novel method can detect any arbitrary sub-graph of the input graph having anomalous characteristics. Thus, the proposed method has been established as a more generic one in its applicability for mining anomalous sub-graphs in graph representation of network data.

Further work in this direction could be on improving the candidate sub-graphs generation procedure and employing a more subjective anomaly identification methodology. One can even try with a different set of parameter values.



(a) Facebook-107 graph (b) AS20000102 graph

Fig. 3. Results of detecting anomalous sub-graphs on benchmark graph data sets

Acknowledgment. The authors would like to thank Director, CAIR for supporting this work. The authors also thank Dr. L. Akoglu for providing the source code of her paper [1].

References

1. Akoglu, L., McGlohon, M., Faloutsos, C.: Oddball: Spotting anomalies in weighted graphs. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS, vol. 6119, pp. 410–421. Springer, Heidelberg (2010)
2. Albanese, A., Pal, S.K., Petrosino, A.: Rough sets, kernel set and spatio-temporal outlier detection. IEEE Trans. on Knowledge and Data Engineering (2012) (online)
3. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys 41(3), 15.1–15.58 (2009)
4. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. PNAS 99(12), 7821–7826 (2002)
5. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
6. Leskovec, J.: Stanford network analysis platform, SNAP (2013), `http://snap.stanford.edu/data/index.html`
7. McAuley, J., Leskovec, J.: Learning to discover social circles in ego networks. In: NIPS, Nevada, USA, pp. 548–556 (2012)
8. Noble, C.C., Cook, D.J.: Graph-based anomaly detection. In: Proc. SIGKDD, Washington, DC, USA, pp. 631–636 (August 2003)
9. Rattigan, M.J., Jensen, D.: The case for anomalous link discovery. SIGKDD Explorations 7(2), 41–47 (2006)
10. Suri, N.N.R.R., Murty, M.N., Athithan, G.: Data mining techniques for outlier detection. In: Zhang, Q., Segall, R.S., Cao, M. (eds.) Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications, ch. 2, pp. 22–38. IGI Global, New York (2011)
11. Wang, F., Li, T., Wang, X., Zhu, S., Ding, C.: Community discovery using nonnegative matrix factorization. DMKD 22(3), 493–521 (2011)