A Decay-Based Approach to Anomaly Detection in Temporal Graph Data

Wen Cao

I. INTRODUCTION

Outlier detection for temporal data has become an increasingly difficult task due to the broad range of applications where time related data is utilized. For virtually all domains, it is of interest to discover abnormal behavior within data sets. Unfortunately, there is no one-size-fits-all model; depending on the nature of the task, there are various methods that can be applied to find abnormality in the data. For example, one has to identify the structure of the data set (whether it is a continuous series, discrete series, etc.) as well as the type of anomaly that is of interest (an unusual data point, an abnormal pattern within the data, etc.) and so forth. Gupta et. al [1] explains this extensively in their survey of outlier detection.

Of the various formats that temporal data can take, it is of particular interest to study graphs as they can accurately model relational data which would be hard to represent otherwise. While there already exists several methods used for anomaly detection in temporal graph data, it is worthy to note that simply having an approach that is functional is not sufficient. Computations on graphs are often very expensive and this becomes evident as more data (in the form of edges and nodes) are added to an expanding graph.

The research community has proposed many solutions for this computation problem, and the overall idea is fairly simple; there exists a series of graph data over time and it is of interest to identify items, events or observations that differ from the majority of the data. This typically involves the computation of a specific value and the comparison of that value with a threshold. The majority of proposed solutions target different ways to compute this comparison value.

II. PROPOSAL AND METHODOLOGY

I propose to apply the decay-based method described by Wu et. al [2] towards the problem of outlier detection in temporal graph data. This approach also allows for different time granularities by choosing a λ value that corresponds to the time frame for detecting irregular behavior in the data. The specific data set that will be used is the city bike data provided by the US cities of Philadelphia and Atlanta. The main aspects of this project will cover the following:

- The city bike data from Philadelphia and Atlanta will be configured into the appropriate format to be represented as a graph. Generally speaking, a graph is represented by a set of nodes and a set of edges. This will be the desired structure for the city bike data sets. The nodes shall be the bike stations in each city and the edges shall be the corresponding bike paths between stations. All edges are undirected and there shall be a graph for each instance of time in the data set.
- For each node in a graph at a time instance *t*, a decay-based product matrix will be produced. This matrix will be of size *n x n*, where *n* is the number of nodes adjacent to the current node. Each value in the matrix will be a decay-based frequency product between the edges of the current node and the corresponding adjacent nodes.
- A characteristic vector and characteristic value is then produced from the product matrix by way of eigenvectors and eigenvalues. These are used to compute the Half-life Magnitude Change value that is used in a normalized Z-score calculation.
- A lazy update approach will be applied towards the decay-based product matrices. The values of the product matrix for a node will be recomputed only when a new node is added to the node's neighborhood. This approach is desirable because eigenvector and eigenvalue calculations are computationally expensive.
- Several values will be tested for the algorithm parameters depending on the scope of outlier detection. For example, different values of λ will detect anomalies across different time scales. This will also affect the half-life value $(\frac{1}{\lambda})$ used in the algorithm.

III. EVALUATION

The metrics for evaluation of the method is still under debate. If time permits, an implementation of another outlier detection algorithm may be compared with the decay-based method. Wu et. al [2] noted that the main bottlenecks within their experimental results came from the eigenvector/eigenvalue calculations and the memory space required. Therefore, it is reasonable to assume that the evaluation criteria may focus on the overall performance and space complexity of the algorithm.

REFERENCES

[1] M. Gupta, J. Gao, C. Aggarwal, J. Han, "Outlier detection for temporal data: A survey", *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250-2267, Sept 2014.

[2] W. Yu, C. C. Aggarwal, S. Ma, H. Wang, "On anomalous hotspot discovery in graph streams", *2013 IEEE 13th InternationalConference on Data Mining (ICDM'13)*, pp. 1271-1276, 2013.