CO461 Data Warehousing and Data Mining Mining
Mini Project Report

Anomaly Detection in Social Networks

Submitted in partial fulfillment of
the requirements for the award of the degree of

Bachelor of Technology
in
Computer Science and Engineering

Submitted by

Roll No	Names of Student
14CO103	Ananda Rao H
14CO142	Sheetal Shalini

Under the guidance of
Dr. M Venkatesan

Department of Computer Science and Engineering
NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
Surathkal, Karnataka, India – 575 025

Odd Semester 2017-18

Department of Computer Science and Engineering

National Institute of Technology Karnataka

Certificate

This is to certify that this is a bonafide record of the project presented by the students whose names are given below during the Odd Semester 2017-18 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Computer Science and Engineering.

Roll No	Names of Student
14CO103	Ananda Rao H
14CO142	Sheetal Shalini

Dr. M Venkatesan
(Course Instructor)

Date:

Abstract

Anomalies are typically defined in terms of deviation from some expected behaviour. Anomalies in online social networks can signify irregular, and often illegal behaviour. Detection of such anomalies has been used to identify malicious individuals, including spammers, sexual predators, and online fraudsters. In this project we implement the OddBall algorithm for ego-Facebook dataset of SNAP (Stanford Network Analysis Project).

Contents

1 Introduction **1**

1.1 Problem Statement 1

1.2 Proposed Solution 1

2 Method Used **2**

2.1 Data Set 2

2.2 Technique Used 2

2.2.1 Calculation of weights 2

2.2.2 Detection of outliers 2

2.2.3 Types of Outliers 3

3 Results and Conclusion **4**

3.1 Results 4

3.2 Conclusion and Future Work 4

References and Resources **5**

Chapter 1

Introduction

1.1 Problem Statement

Anomalies arise in online social networks as a consequence of particular individuals, or groups of individuals, making sudden changes in their patterns of interaction or interacting in a manner that markedly differs from their peers. The impacts of this anomalous behaviour can be observed in the resulting network structure. For example, fraudulent individuals in an online auction system may collaborate to boost their reputation. Because these individuals have a heightened level of interaction, they tend to form highly interconnected subregions within the network. In order to detect this type of behaviour, the structure of a network can be examined and compared to an assumed or derived model of normal, non-collaborative interaction. Regions of the network whose structure differs from that expected under the normal model can then be classified as anomalies (also known as outliers, exceptions, abnormalities, etc.).

1.2 Proposed Solution

Since the social network can be modelled as a graph, link analysis can be used to detect anomalies in the graph. A specific algorithm called the Oddball which uses the neighborhood sub-graphs (called egonets) has been implemented. Oddball uses properties of the egonet like density, weights and eigenvalues to detect anomalies in the social network.

Facebook data obtained from SNAP is an unweighted graph. Weights to the edges is given by determining the similarity between the features of the nodes. Now this weighted graph is used as the input for the Oddball algorithm.

Chapter 2

Method Used

2.1 Data Set

The dataset is obtained from Stanford Network Analysis Project (SNAP) Datasets. Social network analysed for this work is Facebook. The dataset contains the follows :

- *facebook_ combined.txt :* The entire unweighted graph.
- *node.feat :* Feature matrix of the egonet of node.
- *node.egofeat :* Features of the node.
- *node.featnames :* Anonymised Feature names of the egonet of the node.

2.2 Technique Used

2.2.1 Calculation of weights

The input graph of Facebook is unweighted. In order to convert this to a weighted graph features of the end nodes of each edge are compared and matching features determine the weight of each edge.

2.2.2 Detection of outliers

Oddball algorithm is used on the weighted graph obtained from the procedure mentioned above. Based on the different power laws stated in the Oddball algorithm outlierness score is generated for each power law. Average of the outlierness gives the total outlierness. A certain threshold is set for the total outlierness and the nodes having total outlierness score greater than the threshold is classified as an outlier.

2.2.3 Types of Outliers

CliqueStar

CliqueStar detects anomalies having to do with near-cliques and near-stars. Near-cliques and near-stars disobey the Egonet Density Power Law (EDPL) which states that the number of nodes N_i and the number of edges E_i of G_i follow a power law :

$$E_i \propto N_i^{\alpha}, 1 \leq \alpha \leq 2 \tag{2.1}$$

HeavyVicinity

HeavyVicinity detected *heavy egonets*, with considerably high total edge weight compared to the number of edges. Egonets in this case disobey the Egonet Weight Power Law (EWPL) which states that the total weight W_i and the number of edges E_i of G_i follow a power law :

$$W_i \propto E_i^{\beta}, \beta \geq 1 \tag{2.2}$$

DominantPair

DominantPair checks whether there is a single dominant heavy *bursty* edge in the egonet. Egonets in this case disobey the Egonet λ_w Power Law (ELWPL) which states that the principal eigenvalue $\lambda_{w,i}$ of the weighted adjacency matrix and the total weight W_i of G_i follow a power law :

$$\lambda_{w,i} \propto W_i^{\gamma}, 0.5 \leq \gamma \leq 1 \tag{2.3}$$

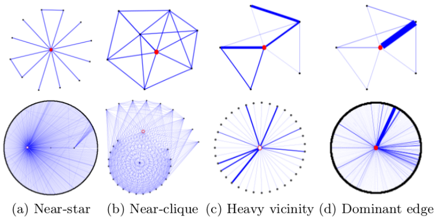
(a) Near-star (b) Near-clique (c) Heavy vicinity (d) Dominant edge

Figure 2.1: Types of Outliers

Chapter 3

Results and Conclusion

3.1 Results

53 outliers were detected among 4039 nodes. Threshold was set to 0.5. *outlierness* ranged from 10^{-5} to 0.74.

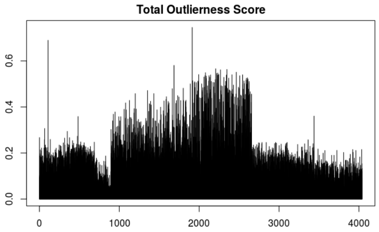

Figure 3.1: Total Outlierness Score

Attribute	Value
# of Nodes	4039
# of Features	1238
# of Edges	88234
# of Outliers	53
% of Outliers	1.3%

Table 3.1: Results in Tabular form

3.2 Conclusion and Future Work

OddBall is a fast un-supervised method to detect abnormal nodes in weighted graphs. It uses egonet attributes like number of nodes, number of edges, edge weights and principal eigen values to assign an *outlierness* score to each node.

This work can be extended to include the following

- Anomaly detection using Trust Rank Algorithm in directed weighted graphs like Twitter and Instagram.
- Use semi-supervised data mining techniques for anomaly detection based on graph methods using proximity, clustering and classification.
- Detect static labelled anomalies to analyze honest/fraudulent behaviour, needed for product review on social networks.

References and Resources

[1] Dataset : https://snap.stanford.edu/data/egonets-Facebook.html

[2] Presentation : https://docs.google.com/presentation/d/1s4bU3eP51VSH7rYaT93D-SDiHvVRBloSSPLWuOMlhHU/edit?usp=sharing

[3] R Documentation : https://www.rdocumentation.org/

[4] igraph Documentation : http://igraph.org/r/doc/

[5] Jure Leskovec and Andrej Krevl, SNAP Datasets, Stanford Large Networks Dataset Collection, http://snap.stanford.edu/data, June, 2014

[6] Leman Akoglu, Mary McGlohon, Christos Faloutsos, *OddBall: Spotting Anomalies in Weighted Graphs*, LNCS, June, 2010

[7] David Savage, Xiuzhen Zhang, Xinghuo Yu, Pauline Chou, Qingmai Wang, *Anomaly Detection in Online Social Networks*, Social Networks, Volume 39, October 2014, Pages 62 to 70