



A Naïve Algorithmic Approach for Detection of Users' with Unusual Behavior in online Social Networks

Pran Dev^a, Jyoti^b, Dr. Kulvinder Singh^c, Dr. Sanjeev Dhawan^d

^{1,2}M.Tech, Computer Engineering, Dept. of CSE, U.I.E.T., Kurukshetra University, Kurukshetra (136119), Haryana, India

^{3,4}Faculty of Computer Science and Engineering, Dept. of CSE, U.I.E.T., Kurukshetra University, Kurukshetra (136119), Haryana, India

Abstract: Unusual users' behavior in online social networks can be detected by analyzing the activities of particular individuals or a group of individuals who make frequent changes in their order of communication and that considerably differ from rest of the group. This type of behavior can be detected by examining the structure of the network and comparing it with the presumed model of the normal network. The regions of the network where architecture is different from the normal network structure can be stated as outliers, exceptions, abnormalities etc. Today, with the convulsive use of the online social networking such as Facebook, Flickr, LinkedIn etc., there has been a gigantic hike in the availability of online social network data, and also the exponential growth in the number of users on online social networks is a severe issue. Here an analysis of Facebook Egonet dataset is carried out which presents various aspects of social network analysis like graph structure, density, degree distribution, modularity etc. Then, on the basis of in degree and out degree of nodes an algorithm is presented which detects the users' with abnormal behavior and the algorithm is tested on the Facebook Egonet dataset.

Keywords: Users' Behavior; Social Network Analysis; Graph Structure; Degree Distribution;

I. Introduction

Emerging technologies and online social networks have become a manifold for the young generation. The number of users is also increasing very rapidly on the online social networks like Facebook, Twitter, LiveJournal etc. But, the large size of online social networks is also attracting the malicious users for various purposes like rumors, data theft, spams etc. So, the problem of detecting the hidden relationships in large online social networks is very important from the perspective of Internet Service Provider, Social Network Provider and also from users' perspective. For this purpose, it is necessary to find out the communication patterns of users to distinguish the malicious nodes from the legitimate nodes. This encourages researchers to find the unusual activities occurring these days on online social networks. Here also, an attempt has been made to find an algorithm to detect unusual behavior of users over the web.

The paper is divided among 5 sections. Section 2 describes the tools used for graph analysis. Section 3 gives a detailed description of the analysis of Facebook dataset^{1, 2} which includes degree distribution, graph distance report and graph density details. Section 4 represents the workflow of the proposed algorithm's execution and the concrete results. Section 5 finally concludes the paper in nutshell and provides the scope of present work in near future.

II. Tools Used

Gephi is an open source tool for the analysis of graphs and network structures downloadable at www.gephi.org. Gephi has its own file extension *.gephi*. It can also read excel and *.csv* files. There are three major sections within Gephi: Overview Tab, Data Laboratory and Preview Tab. It also provides various layouts for representation of network structures and graphs. It is a very helpful tool for social network analysis³.

Eclipse⁴ is used to test the algorithm designed on the Facebook dataset. It is a java based integrated development environment. The results of algorithm are stored in excel sheet through java code in eclipse.

III. Analysis of Facebook Dataset

Input is a set of 4039 vertices and 88234 edges. This is a real world anonymous dataset of Facebook online social network available at <http://snap.stanford.edu/data/egonets-Facebook.html>. The graph structure of the above Facebook dataset is analyzed for further research. The graphical representation using Fruchterman Reingold Layout of dataset is shown below:

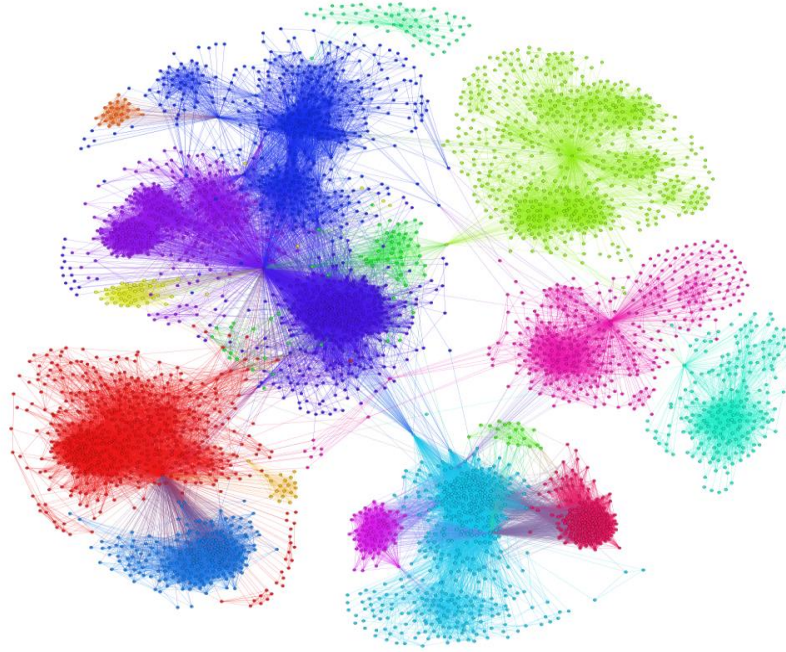


Fig 1: Graph of Facebook dataset

A. Degree Distribution

Average Weighted Degree: 21.846

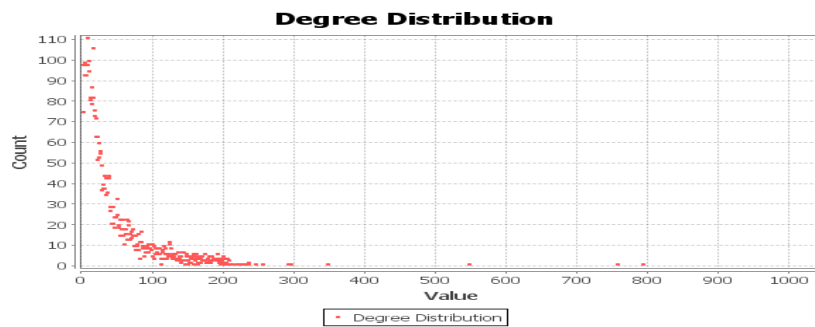


Fig 2: Total Degree Distribution

B. Graph Distance Report

Diameter: 17
Average Path length: 4.3377
Number of shortest paths: 2508102

B.1. Betweenness Centrality Distribution

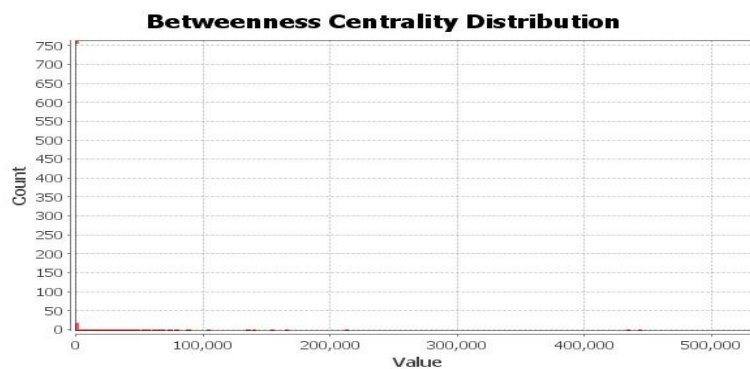


Fig 3: Betweenness Centrality Distribution

B.2. Closeness Centrality Distribution

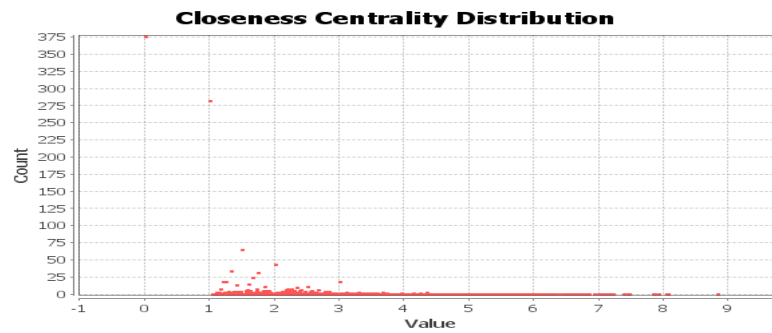


Fig 4: Closeness Centrality Distribution

B.3. Eccentricity Distribution

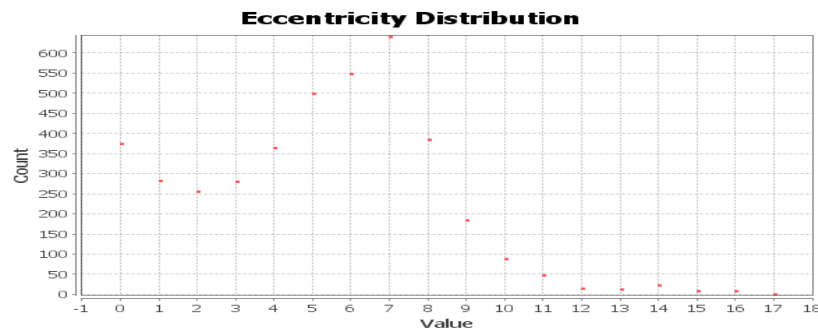


Fig 5: Eccentricity Distribution

C. Graph Density and Modularity

Parameters: Directed
 Density: 0.005
 Randomize: On
 Use edge weights: On
 Resolution: 1.0

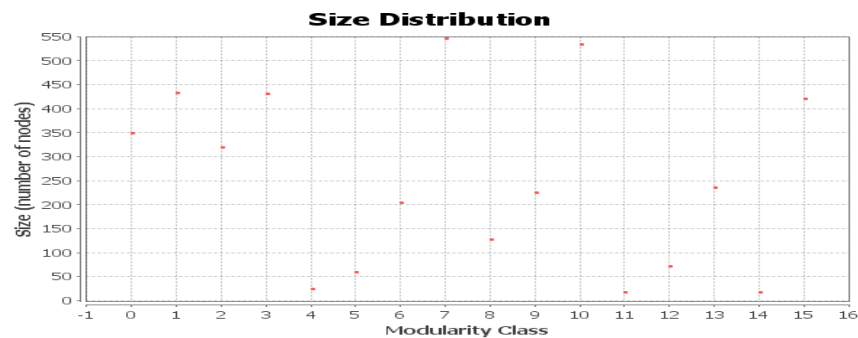


Fig 6: Size Distribution

IV. Designed Algorithm and Results

There are four categories of nodes: one having high in degree and high out degree, second, high in degree and low out degree, third, low in degree and low out degree and the last low in degree and high out degree. Nodes of first three categories are considered to be normal as of their communication patterns but nodes of fourth category may not be normal as of their high out degrees. That means, they are trying to connect with a large number of nodes but as we are having low in degree which makes them dubious of having very few incoming connections. So, it makes them behave different from the other three above mentioned categories. Although, we cannot say that these nodes are spams, but still, it can be said that such communication patterns in the network are not usual or normal. The naive algorithm aims to find nodes with such unusual behavior. This algorithm takes an excel file as input which is having three attributes: node Id, In degree and Out degree. This algorithm finds out the nodes with unusual behavior and gives another excel file as the output which is having three attributes: node Id, In degree and Out degree.

A. Flowchart of the Proposed Algorithm

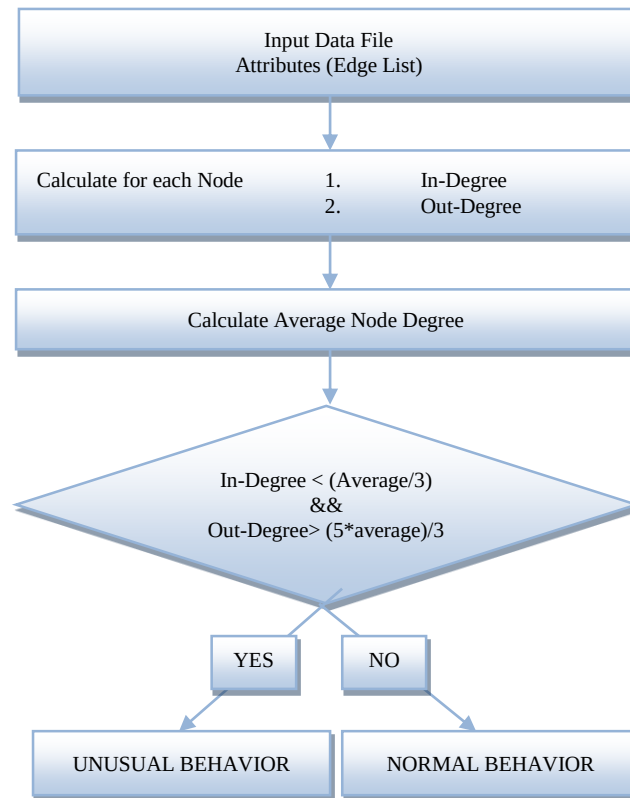


Figure 7: Flow Chart of Purposed Algorithm

B. Input

Input is an excel file with the attributes node_Id, In Degree and Out Degree. An instance of the input file is shown below.

The screenshot shows a Microsoft Excel spreadsheet with the following data:

Id	In-Degree	Out-Degree
0	0	347
1	1	16
2	2	9
3	3	16
4	4	9
5	5	12
6	6	5
7	7	19
8	8	7
9	9	55
10	10	9
11	11	0
12	12	0
13	13	30
14	14	14
15	15	0
16	16	8
17	17	12
18	18	0
19	19	2
20	20	12
21	21	62
22	22	9
23	23	16

Fig 8: Instance of Input

C. Output

Id	Degree Out-Degree
1	347
2	55
3	62
4	65
5	62
6	39
8	1043
9	225
10	100
11	93
12	93
13	42
14	123
15	65
16	37
17	44
18	76
19	42
20	146
21	124
22	104
23	44
24	163
25	71

Fig 9: Instance of Output

Output file contains the entry for the nodes having unusual behavior and it contains their In Degree and Out Degree along with their node id, but only of the nodes which are not normal behaving. So, finally we scrutinized the Nodes in the Facebook data set which are having Unusual Behavior. Here is a snapshot of the output file. Initially, there were total 4039 Nodes in the available dataset, but after applying the designed algorithm for detection of unusual behaving users; we found that there are only 123 Nodes of this category.

V. Conclusion and Future Work

With advent in interactions between the entities in the network, the behavior of users in online social network has become an enigma. Discovering countless communities in the networks is of vital importance. There is a dire need of an affirmative study to have a deep understanding of social networks and detecting hidden patterns. In this research the attempt is made to propose a naive algorithm for the analysis of Facebook with different parameters. This algorithm is based on the in degree and out degree of the nodes. This algorithm takes as input, a list of nodes with in degree and out degree and after execution; it gives output, a list of users with unusual behavior. In future, some of the problems in hidden relationship detection as discussed above can be tackled by making small modules of large data over small intervals of time and analysis can be done on these small units and then they can be integrates them back to solve the whole problem. So, a divide and conquer strategy can be developed, and feature extraction can be applied on small sets of data. Performance of the proposed algorithm can also be enhanced by applying fuzzy methods, neural network approach or genetic algorithm. For the future, it is important to test the proposed algorithm more intensively in a larger network using several test groups.

References

- [1]. J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.
- [2]. <http://snap.stanford.edu/data/egonets-Facebook.html> accessed on 12 January 2015 at 4:50 pm.
- [3]. <http://www.gephi.com> accessed on 20 January 2015 at 12:35 pm.
- [4]. http://en.wikipedia.org/wiki/Eclipse_%28software%29, accessed on 19 May 2015 at 01:00 pm.