Outlier Detection in Ego-Centric Networks

Bharatraj Telkar[1], Chirag Ramesh[2], and Daniel Issac[3]

***Abstract*— Anomalies in social networks arise due to irregular activities or sudden changes in interaction by individuals or a group of individuals in the network. These irregularities make such individuals different from the rest of their nodes in the network. Detection of such anomalies can be used to identify fraudulent and malicious individuals. Anomalies have a significant impact on the network structure, as several malicious entities often have distinct patterns of interaction. These patterns can be learned via several techniques in an attempt to detect such anomalies.**

I. INTRODUCTION

We intend to detect anomalous interactions or behavior in social networks, which may be indicative of even larger problems in the network. Depending on the specific domain or context, these interactions could indicate the presence of fraudulent individuals, spammers, or subversive activities like interactions of terrorist groups on social networks.

There are various forms of anomalous behavior; a major one that we are interested in being outlier detection where abnormal behavior or characteristics in a dataset might often indicate that the person perpetrates suspicious activities. These could be of two types:

Data Set Level Where the behavior of a person/instance does not comply with the overall behavior.

Data Item Level Behavior of a person/instance does not comply with normal behavior of that person/instance.

II. LITRATURE SURVEY

In their paper on anomalous user behavior in social networks, Bimal Vishwanath et. Al. try to use unsupervised anomaly detection techniques to search for suspicious behavior patterns, thus lowering the dependence on labeled examples. The technique used is principal component analysis (PCA), which searches for deviation from a normal standard of behavior on social networks. This use case for detecting anomalous behavior is a focus we adopt in our paper as well, though we think trained examples give us greater accuracy in our learning model. This paper also explicates the usage of anomalous behavior to take advantage of the crowd-sourced nature of interactions on social networks, which is a scenario where ego-centric networks could be formed which our model would be particularly well-placed to detect.

Leman Akoglu et. Al. have proposed the OddBall algorithm which spots anomalies in weighted graphs. They discuss the power laws based on density, weights, ranks and eigenvalues and how they can be used as rules to identify neighborhoods and anomalies. New features are developed which are applied to further detect anomalies. We have taken inspiration from this paper in our approach towards detecting anomalies and developing new features.

In their paper on moving towards fully supervised anomaly detection, Nico Gornitz, Marius Kloft et Al. demonstrate how unsupervised anomaly detection often fails to meet the required detection rates and thus demonstrate the need for labeled data to guide model generation. They say that though anomaly detection is seen to be an unsupervised task as these anomalies are unpredictable and have unknown distributions, they say that a semi-supervised algorithm can be grounded on unsupervised anomaly detection methods and devised to overcome this problem. Additionally, they also propose an active learning strategy to label candidates. They are able to observe that this methodology requires much less examples than other state-of-the-art techniques, while providing higher detection rates. As our solution also attempts to implement semi-supervised anomaly detection, we used this paper to understand the development of such a model.

III. METHODOLOGY

Algorithm: Fraudulent Community Detection Approach Our proposed solution is loosely based on the GotchAll methodology for fraud detection in social networks. The algorithm calculates scores for the individual nodes as well as the detected cliques and uses them to form a model that can be used to detect anomalies in a social network. The algorithm is divided into three steps.

- Individual Scoring. Initially, the edges are not weighted. We assign each edge a weight based on the number of common characteristics between the nodes connected by the edge. This helps us find ego-centers. We discretize the features of these nodes and use it to calculate the individual scores of the nodes.
- Clique Detection. In order to detect complete cliques, we search for the exact match for the clique in the community. Then, we delete the original clique. To detect partial cliques, we find partial overlaps and keep the original clique.
- Clique Scoring. Clique scoring is done using Intrinsic Features, Relational Features, Clique Based Features.

IV. IMPLEMENTATION

We have implemented this algorithm using Python on Jupyter Notebook. Our implementation attempts to extract features from the dataset, which will then be used to cluster the data into a model which will give us information about possible outliers, in addition to the nodes that are already

known as definite outliers and then train a semi-supervised model based on these predictions for future outliers.

We have initially used three power laws, the edge weight power law, the edge density power law and the Eigen value power law, to form three features that allow us to test for particular types of anomalies (bursty edges, almost star sub-networks). We then iterated these similar to the Page Rank algorithm with separate initial weights as the initial outlierness scores, to determine whether these converge to a certain level. We eyeballed certain thresholds for all these features to determine outliers on graphs.

We then formed thresholded features for each of these, which would have not one but two thresholds, thus leading to three sets, with the idea being that we would be able to separate the nodes into safe, possibly outlier and definitely outlier sets. The final set of features was formed by calculating a score for each node from the weights of its edges raised to the power of the thresholded normal outlierness score for the edge target. What this would return is a higher value for those nodes near many outliers, which is an important characteristic for prediction.

This gave us our list of features which we used to create a dataset for analysis. At first, from a heat map of the different features, some interesting properties were observed. The iterated outlierness scores from the modified PageRank algorithm were all perfectly correlated even though they corresponded to different metrics, thus implying that they resulted in the same data for us, allowing us to discard 2 of them. Other sets also displayed high degrees of correlation, though these could be disregarded as the minute change in values represented high amounts of information in an essentially sparse dataset.

A. Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled. Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

$$\alpha + \beta = \chi \tag{1}$$

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use (1), not Eq. (1) or equation (1), except at the beginning of a sentence: Equation (1) is . . .

B. Some Common Mistakes

- The word data is plural, not singular.
- The subscript for the permeability of vacuum ?0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter o.
- In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an inset, not an insert. The word alternatively is preferred to the word alternately (unless you really mean something that alternates).
- Do not use the word essentially to mean approximately or effectively.
- In your paper title, if the words that uses can accurately replace the word using, capitalize the u; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones affect and effect, complement and compliment, discreet and discrete, principal and principle.
- Do not confuse imply and infer.
- The prefix non is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the et in the Latin abbreviation et al..
- The abbreviation i.e. means that is, and the abbreviation e.g. means for example.

V. USING THE TEMPLATE

Use this sample document as your LaTeX source file to create your document. Save this file as **root.tex**. You have to make sure to use the cls file that came with this distribution. If you use a different style file, you cannot expect to get required margins. Note also that when you are creating your out PDF file, the source file is only part of the equation. *Your TEX → PDF filter determines the output file size. Even if you make all the specifications to output a letter file in the source - if you filter is set to produce A4, you will only get A4 output.*

It is impossible to account for all possible situation, one would encounter using TEX. If you are using multiple TEX files you must make sure that the "MAIN" source file is called root.tex - this is particularly important if your conference is using PaperPlaza's built in TEX to PDF conversion tool.

A. Headings, etc

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next

level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named Heading 1, Heading 2, Heading 3, and Heading 4 are prescribed.

B. Figures and Tables

Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation Fig. 1, even at the beginning of a sentence.

TABLE I
AN EXAMPLE OF A TABLE

One	Two
Three	Four

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an document, this method is somewhat more stable than directly inserting a picture.

Fig. 1. Inductance of oscillation winding on amorphous magnetic core versus DC bias magnetic field

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity Magnetization, or Magnetization, M, not just M. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write Magnetization (A/m) or Magnetization A[m(1)], not just A/m. Do not label axes with a ratio of quantities and units. For example, write Temperature (K), not Temperature/K.

VI. CONCLUSIONS

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

APPENDIX

Appendixes should appear before the acknowledgment.

ACKNOWLEDGMENT

The preferred spelling of the word acknowledgment in America is without an e after the g. Avoid the stilted expression, One of us (R. B. G.) thanks . . . Instead, try R. B. G. thanks. Put sponsor acknowledgments in the unnumbered footnote on the first page.

References are important to the reader; therefore, each citation must be complete and correct. If at all possible, references should be commonly available publications.

REFERENCES

[1] G. O. Young, Synthetic structure of industrial plastics (Book style with paper title and editor), in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 1564.
[2] W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123135.
[3] H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
[4] B. Smith, An approach to graphs of linear forms (Unpublished work style), unpublished.
[5] E. H. Miller, A note on reflector arrays (Periodical styleAccepted for publication), IEEE Trans. Antennas Propagat., to be publised.
[6] J. Wang, Fundamentals of erbium-doped fiber amplifiers arrays (Periodical styleSubmitted for publication), IEEE J. Quantum Electron., submitted for publication.
[7] C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style), IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].
[9] M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
[10] J. U. Duncombe, Infrared navigationPart I: An assessment of feasibility (Periodical style), IEEE Trans. Electron Devices, vol. ED-11, pp. 3439, Jan. 1959.
[11] S. Chen, B. Mulgrew, and P. M. Grant, A clustering technique for digital communications channel equalization using radial basis function networks, IEEE Trans. Neural Networks, vol. 4, pp. 570578, July 1993.
[12] R. W. Lucky, Automatic equalization for digital communication, Bell Syst. Tech. J., vol. 44, no. 4, pp. 547588, Apr. 1965.
[13] S. P. Bingulac, On the compatibility of adaptive controllers (Published Conference Proceedings style), in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 816.
[14] G. R. Faulhaber, Design of service systems with priority reservation, in Conf. Rec. 1995 IEEE Int. Conf. Communications, pp. 38.
[15] W. D. Doyle, Magnetization reversal in films with biaxial anisotropy, in 1987 Proc. INTERMAG Conf., pp. 2.2-12.2-6.
[16] G. W. Juette and L. E. Zeffanella, Radio noise currents n short sections on bundle conductors (Presented Conference Paper style), presented at the IEEE Summer power Meeting, Dallas, TX, June 2227, 1990, Paper 90 SM 690-0 PWRS.
[17] J. G. Kreifeldt, An analysis of surface-detected EMG as an amplitude-modulated noise, presented at the 1989 Int. Conf. Medicine and Biological Engineering, Chicago, IL.
[18] J. Williams, Narrow-band analyzer (Thesis or Dissertation style), Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.
[19] N. Kawasaki, Parametric study of thermal and chemical nonequilibrium nozzle flow, M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.
[20] J. P. Wilkinson, Nonlinear resonant circuit devices (Patent style), U.S. Patent 3 624 12, July 16, 1990.