Identifying anomalous activity in the Enron email corpus using machine learning and glocal multivariate graph statistics

Disa Mhembere (dmhembe1, dmhembe1@jhu.edu), Kunal Lillaney (klillan1, lillaney@jhu.edu)

1 Abstract

Graphs are quickly emerging as a leading abstraction for the representation of data flow and interactions within networks. One high-interest area of study deals with the use of graph theory to detect attacks and anomalous activity in networks [1], [2], [3]. Additionally, the use of machine learning to detect threats within networks has become a topic of particular interest [4], [5], [6], [7].

The use of glocal multivariate graph statistics [8] as a supplementary source of multi-dimensional features for anomalous network detection, has not been explored *to our knowledge*. Such statistics can expose otherwise-hidden topological attributes within networks that have the potential to improve anomalous activity detection within a network. We propose the use of such statistics to generate novel features for use within classification tasks for the Enron corpus of emails [9]. We hope to show that the use of such statistics can help to identify periods of anomalous activity and associated actors.

2 Methods

We intend to use the following statistics as supplementary attributes: Degree, Triangle Count, Clustering-coefficient, Scan statistics, K-means clustering, Connected Components, Spectral Decomposition, Adhesion and possibly more. These statistics are chosen based on their ability to efficiently distill graph topological properties into discrete values that are usable by ML techniques. We will use the igraph [10] package to create the time-series graphs and compute statistics. We will parse and transform the time series graphs into an ML ingestible format.

Since we can correlate events in the news with time, we hope to generate classification labels based on events surrounding reported Enron fraudulent activities. Literature [7], [5] has led us to believe that SVMs may be suitable from such a classification task. We propose the use of a weighted kNN to identify malicious actors in the network at different time points since we conjecture the features of malicious users will have high relative 'similarity'. Please this may change as we research further into the topic.

3 Resources

We will use the following resources to generate our dataset: (i) `http://www.cs.cmu.edu/ ./enron/` – Contains 0.5M emails from senior management of Enron, (ii) `http://cis.jhu.edu/ parky/Enron/enron.html` – Scan Statistics on Enron graphs. We will use custom scripts to create ML-ingestible instances/features.

4 Milestones

4.1 Must achieve

Able to correlate specific reported events in the news with anomalous email activity.

4.2 Expected to achieve

Identify the different activities as anomalous or benign events.

4.3 Would like to achieve

Determine specific actors as anomalous at varying time points through hybrid weighted kNN and SVM techniques.

5 Final Writeup

Our final writeup will include (i) A survey of graph statical algorithms pertinent to classification tasks, (ii) A survey of machine learning techniques used to detect anomalies in networks, (iii) A cost-benefit analysis between methods in *(ii)* and our approach, (iv) Detailed description of our methods and associated novelty, (v) The classification results of anomalous activity on the Enron dataset, (vi) The classification of individual actors during times of anomalies, (vii) Future work and Conclusion.

6 Bibliography

References

[1] C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park, "Scan statistics on enron graphs," *Computational & Mathematical Organization Theory*, vol. 11, no. 3, pp. 229–247, 2005.

[2] Y. Park, C. Priebe, D. Marchette, and A. Youssef, "Anomaly detection using scan statistics on time series hypergraphs." SDM, 2009.

[3] Y. Park, C. E. Priebe, and A. Youssef, "Anomaly detection in time series of graphs using fusion of graph invariants," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 1, pp. 67–75, 2013.

[4] M. V. Mahoney, "A machine learning approach to detecting attacks by identifying anomalies in network traffic," Ph.D. dissertation, Florida Institute of Technology, 2003.

[5] T. Shon, Y. Kim, C. Lee, and J. Moon, "A machine learning framework for network anomaly detection using svm and ga," in *Information Assurance Workshop, 2005. IAW'05. Proceedings from the Sixth Annual IEEE SMC*. IEEE, 2005, pp. 176–183.

[6] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Security and Privacy (SP), 2010 IEEE Symposium on*. IEEE, 2010, pp. 305–316.

[7] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," *Information Sciences*, vol. 177, no. 18, pp. 3799–3821, 2007.

[8] D. Mhembere, W. G. Roncal, D. Sussman, C. E. Priebe, R. Jung, S. Ryman, R. J. Vogelstein, J. T. Vogelstein, and R. Burns, "Computing scalable multivariate glocal invariants of large (brain-) graphs," *arXiv preprint arXiv:1312.4318*, 2013.

[9] W. Cohen. (20014) Enron email dataset. [Online]. Available: http://www.cs.cmu.edu/~enron/

[10] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: http://igraph.org