Applying Social Network Analysis to Security

Elizabeth Phillips supervised by Michael Goldsmith and Jason Nurse

Oxford University CDT in Cyber Security

UNIVERSITY OF OXFORD — CENTRE FOR DOCTORAL TRAINING *in* CYBER SECURITY

EPSRC Engineering and Physical Sciences Research Council

Introduction

In this project, we set out to explore the many ways in which Social Network Analysis (SNA) can be applied to the field of security, and especially, investigating what interactions could someone (e.g., an attacker) infer if they were able to gather data on a person's friend groups or technology devices (e.g., email interactions) and whether this could be used to predict the "hierarchical importance" of the individual.

For this experiment we conducted our initial investigation on the Enron email dataset [SA04] and then investigated the effectiveness of existing SNA metrics on establishing hierarchy from the social network created from the email communications from our research group. We also created tool-support to allow us to analyse the discovered social networks.

Social Networks

Link Analysis (LA) is the analysis of relationships and information flow between a network of individuals, groups and other connected entities and has been a topic of study for several decades [GD05], [W94]. A Social Network (SN) is defined as the representation of communication networks with people as nodes and relationships between them as links in a graph. Social Network Analysis (SNA) is defined as the application of Link Analysis to a social network. Within an organisation's social network, we define the "hierarchical importance" of an individual as the seniority of the individual within the organisation.

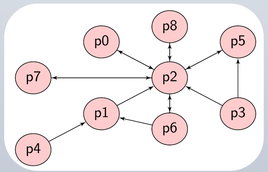

Figure 1:- An example of a social network of 9 nodes V = $\{p_0,...p_8\}$ and the edges (links) between the individuals.

Methodology

For each of our two experiments, we followed a three step method in order to analyse our results as outlined in figure 2.

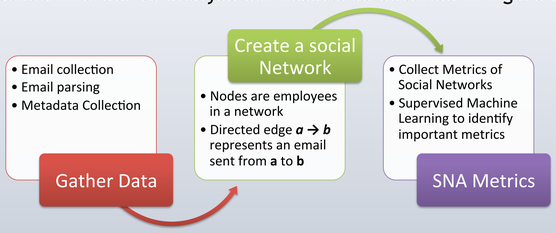

Figure 2:- A flowchart outlining our methodology for each experiment.

Metrics

When performing SNA on our network we calculated the following metrics for each user based on an extensive literature survey and their ability to identify importance within graphs :-

- Number of emails sent
- Number of emails received
- Degree Centrality
- Betweenness Centrality
- PageRank
- Markov Centrality
- HITS Authority Score
- HITS Hub Score
- Clique Score
- Weighted Clique Score
- Average Distance
- Clustering Coefficient

Experiment 1

In this experiment we used the publicly available Enron dataset and analysed over 600,000 emails from 149 employees email mailboxes. We then calculated the metrics specified for each employee. Each employee was assigned 1 of 7 tiered Hierarchical roles based on the ground truths available. This was eventually split into "Boss" and "Not Boss" roles due to the incomparable roles due to varying team sizes.

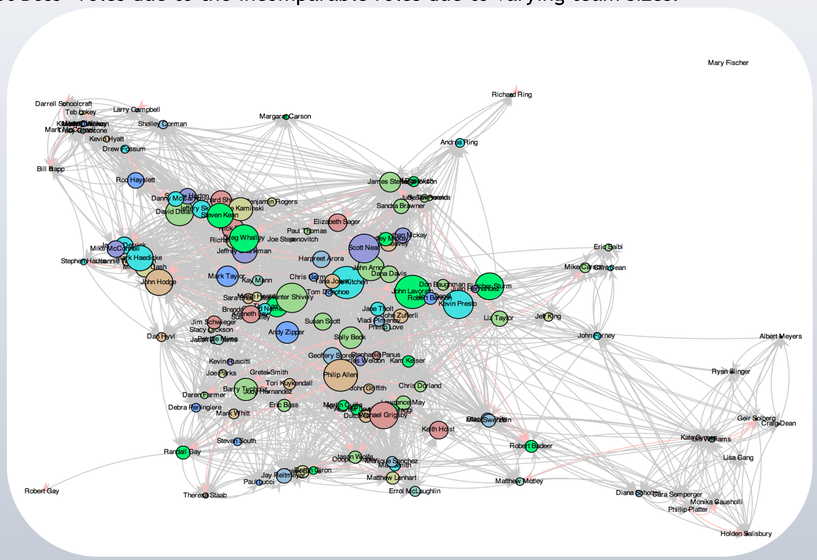

Figure 3:- A copy of the social network graph corresponding to the social network of the Enron data. The size of each node represents the HITS authority score of the employee.

Experiment 2

We then collected a secondary dataset with which to verify the effectiveness of the results discovered from the Enron dataset, from the 10 employees within our group. For this experiment we wanted to predict whether an employee was categorised as a "Boss" or "Not Boss".

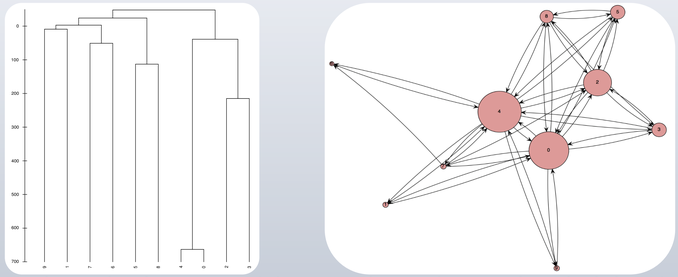

Figure 4: The Dendogram hierarchical structure of the research group's network to the left and the social network representation of the same network to the right.

Results

Experiment 1:-

In order to predict whether an employee was in the "Boss" Category, we compared the results of several different supervised learning algorithms including MultiLayer Perceptron, K-Star, Naïve Bayes and Bayesian Networks. MultiLayer Perceptron provided the best results with an ROC Area of 0.939.

MultiLayer Perceptron						
TP_Rate	FP_Rate	Precision	Recall	F-Measure	ROC Area	Class
0.956	0.273	0.977	0.956	0.967	0.939	Not_Boss
0.727	0.044	0.571	0.727	0.64	0.939	Boss

Figure 5:- A breakdown of the results from our MultiLayer Perceptron model.

Experiment 2:-

Our second experiment identified 5 metrics which helped to separate the 2 "Boss" employees from the other employees; namely number of emails sent, number of messages received, pagerank score, betweenness centrality and clustering coefficient. Our dendogram also helped identify strong relationships which could help build a stronger hierarchical structure of the network.

Future Work

One direction our research can take is to investigate how the hierarchical structure of the network changes over time and to test whether this can help identify anomalies to help with insider threat detection. We also hope to apply these techniques within Dark Web Forums in order to identify influential members.

At a higher level, the accuracy and ease of which the SNA metrics can be performed may help us ask questions about the risk exposure an individual may have online based upon the availability of the metadata from email communication.

References

[SA04] Shetty, Jitesh and Adibi, Jafar. ***"The Enron email dataset database schema and brief statistical report"***. *Information Sciences Institute Technical Report*, University of Southern California , 4, 2004. http://foreverdata.com/1009/Enron_Dataset_Report.pdf

[GD05] L. Getoor and C. P. Diehl, ***"Link Mining: A Survey,"*** *SIGKDD Explorer Newsletter*, vol. 7, no. 2, pp. 3-12, Dec. 2005.

[W94] S. Wasserman, ***"Social Network Analysis: Methods and Applications"***. Cambridge University Press, 1994.

Contact Details

Address:- CDT in Cybersecurity
Robert Hooke Building
Parks Road
Oxford
OX1 3PR

Email:- elizabeth.phillips@cs.ox.ac.uk

Website:- http://www.elizabethphillips.co.uk