ISSN 2052-8604

Applying social network analysis to security

Elizabeth Phillips[1], Jason Nurse[2], Michael Goldsmith[2] and Sadie Creese[2]
[1]Oxford University Centre for Doctoral Training in Cyber Security
elizabeth.phillips@cybersecurity.ox.ac.uk

[2]Cyber Security Centre, Department of Computer Science,
University of Oxford, Oxford, UK
{*firstname.lastname*}@cs.ox.ac.uk

Abstract

In this paper, we set out to explore some of the many ways in which Social Network Analysis (SNA) can be applied to the field of security. In particular, we investigate what information someone (e.g., an attacker) could infer if they were able to gather data on a person's friend-groups or device communications (e.g., email interactions) and whether this could be used to predict the "hierarchical importance" of the individual. This research could be applied to various social networks to help with criminal investigations by identifying the users with high influence within the criminal gangs on DarkWeb Forums, in order to help identify the ring-leaders of the gangs. For this study we conducted an initial investigation on the Enron email dataset, and investigated the effectiveness of existing SNA metrics in establishing hierarchy from the social network created from the email communications metadata. We then tested the metrics on a fresh dataset to assess the practicality of our results to a new network.

Introduction

The Internet has transformed the way in which people communicate with each other within society. With the increase in communications, comes an added exposure associated with this additional traffic. This paper aims to focus on the specific test case of inferring hierarchy from such communications. The technique that we are specifically interested in is Social Network Analysis (SNA), i.e. a set of approaches that allow for the study of social links between elements (e.g. people, devices or things).

Social networks have been an attractive resource to analyse dating as far back as 1930[1]. Freeman in 1979 highlighted the initial works of Moreno, Jennings, Warner and others in investigating the social networks within schools, prisons and workplaces. However, the Real World Experiment of Travers et al. in 1969[2] was the first to highlight how connected our own social networks are with the “small world phenomenon”. The directed nature of communication allows SNA to be used to help create comprehensive network graphs that can be assessed visually and mathematically (through a range of SNA metrics) to help identify influential nodes and/or clusters within the network.

Email is widely accepted by the business community as the first broad electronic communication medium and was the first ‘e-revolution’ in business communication. Typically, email is used for alerting, archiving, task management, collaboration, and interoperability. According to Radicati’s 2014 Surveys[3], 108.7 billion business emails are sent and received daily (up from 89 billion in 2012 [4]). This accounts for 55.4% of the total email communication globally (196.3 billion). By 2018, this is expected to increase by 28.2% to 139.4 billion. Within an organisation, emails may be used to send messages regarding the latest football score or to discuss the latest draft of a report[5]. The diverse interactions that email mediates allow researchers a unique insight into the everyday workings of an organisation and may help reveal informal hierarchies that may not be evident to an individual outside of the organisation[6], [7].

Since the revelations of metadata collection exposed by Edward Snowden in June 2013 [8], the importance of metadata from emails is gaining awareness. In the light of these revelations, organisations are investigating the current risk exposure of their own data[9] and the extent to which the US surveillance schemes may affect their organisation. In order to collect a sufficiently large dataset along with the associated ground truth, we decided to focus on email communication networks. As these techniques are improved, it may be possible to apply these techniques in order to identify influential players within DarkNet forums or other criminal networks in order to help with criminal convictions.

Research Question and Approach

In this paper we set out to investigate the effectiveness of existing SNA techniques when applied to hierarchical analysis based upon the metadata from email communications. As there has been research on this topic in the literature (e.g., the specific objective here will be towards enhancing the accuracy of inferring these relationships and using fewer metadata elements to complete the inference. In particular, we aim to answer “***To what extent can SNA techniques be used to assess email communications metadata to identify known, but also hidden social groups***”.

We will split the research into four main tasks, namely:-

ISSN 2052-8604

- **Initial investigation**: This task focuses on implementing several of the existing SNA methods and metrics, and applying them to a communication dataset to see how well they perform in identifying groups and their structures (i.e. hierarchies of individuals). We put special emphasis to the number of data elements required to define structures and the accuracy with which these structures can be identified. For this experiment we use the Enron email communications dataset given the availability of ground truths to evaluate the methods and support our findings, and also its large size.
- **Enhancing the discovery of groups and social structures:** Having investigated the effectiveness of existing SNA techniques, we will aim to enhance the accuracy of these techniques in predicting the "hierarchical importance" of an individual. We will also introduce new methods through which groups and social structures can be identified. For an initial evaluation of these new approaches, we again use the Enron dataset.
- **Collecting a new email communications corpus**: To test our enhanced inference techniques, we collect a new communications corpus from willing volunteers and use our techniques established above to compare our predicted hierarchy with the true hierarchy in the dataset. We use the metrics identified as useful from the first two experiments.
- **Evaluating the enhanced inference methods**: At this stage, we evaluate our SNA proposals and the level of accuracy with which they can identify the known social groups (as documented in the sample's ground truth). As we are using an organisational dataset for our analysis, we are also interested in discovering whether our approaches can discover the organisational hierarchies.

Methodology

In order to address the research question aims, we began by collecting the emails from the dataset of interest. From the email collection we were able to extract the metadata from each email from which we can build our network. Once we have extracted the data from the email communication network, we then created a graph of the new social network where each node will represent an employee and each directed edge ***a* → *b*** represents an email sent from **a** to **b**. The weight of each edge corresponds to the number of emails sent from **a** to **b**.

Once we have created our graph, we then set out to identify metrics on our network that may be useful in helping to determine the relative "***importance***" of an individual within it. Once these metrics have been calculated for each node of our network, our next task is to apply Supervised Machine Learning (SML) to identify the metrics that are useful when determining hierarchy within the organisation. SML allows us to create a model which links the metrics to a corresponding hierarchical job "***category***" within the organisation as well as allowing us to exclude particular metrics from future experiments due to their lack of contribution. After performing SML we identified a number of useful metrics that can be used to determine the relative importance of an individual. Once our model was created, we tested the validity of our results on a real dataset. We apply our trained model to this new dataset in order to determine how accurate it is at identifying the senior management in the group.

Link Analysis and SNA

Complex interactions between entities can be modelled as networks. These networks include the Internet [10], food webs [24] and biochemical networks [15]. Each of these networks consists of a set of nodes or vertices (e.g. computers or routers on the Internet or people in a social network), connected together by links or edges, representing data connections between computers, friendships between people etc.

Link Analysis (LA) is the analysis of relationships and information flow between a network of individuals, groups, organizations, servers and other connected entities, and has been a topic of study for several decades[10], [11]. A Social Network (SN) is defined as the representation of 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. We can perform SNA on our newly created Enron social network in order to determine the hierarchical structure of the organisation. Within a group's social network, we define the "hierarchical importance" of an individual as the seniority of the individual within the group.

SNA Metrics

Within the field of SNA, there are a range of metrics that can be used to assess a network and the nodes (individuals) within it. In this experiment we aim to assess whether these (or enhanced variations of them) could be used to determine the importance of an individual simply through a broad set of Email-Communications data.

ISSN 2052-8604

Attribute Name	Description
Sent Messages (SM)	The number of emails sent by an employee.
Received Messages (RM)	The number of emails received by an employee.
Degree Centrality (DCS)	The number of distinct employees within the network that an employee has sent emails to.
Betweenness Centrality Score (BCS)	The betweenness centrality measure for an employee.[11]
Pagerank Score (PRS)	The PageRank score an employee.[27]
Markov Ranking (MR)	The markov ranking of an employee. [20]
HITS Authority Score (HAS)	The authority score for an employee (if several users with high hub weights send an email t the user then they will have a higher authority score). [18]
HITS Hub Score (HHS)	The hub score for an employee (if the user sends emails to users with high authority scores then they will have a higher hub score). [18]
Clique Score (CS)	The number of cliques (maximal subgraphs) an employee is in using the Bron and Kerbosch algorithm.[6]
Weighted Clique Score (WCS)	The weighted clique score for each user, weighted by the number of users within each clique.
Average Distance Score (ADS)	The average distance between the user and all other users in the graph.
Clustering Coefficient (CC)	The extent to which vertices in a graph tend to cluster together. [35]

TABLE 1:- DESCRIPTION OF OUR CHOSEN SNA METRICS

Our assumption that $\boldsymbol{p_2}$ in Figure 1 plays a central role is due to the proportion of the network that they connect with. This is formally known as the Degree Centrality of the node and is one of many SNA metrics that may be of use in our analysis. Table 1 contains the metrics that we decided to investigate as part of our analysis. The metrics were chosen based on a literature review of previous research and their ability to identify nodes of influence within a SN[12]. We present these in terms of their use with our Enron dataset where the nodes represent employees and the graph edges represent email communications between employees.

Initial Investigation

For our first investigation, we used the Park et al.[13] dataset for our analysis. This was based on the original dataset of Adibi and Shetty in ISI[14], but has been modified to delete extraneous duplicate emails and fix some anomalies in the data. Our final dataset consisted of 184 email addresses corresponding to 147 employees and a total of 517,431 emails. The ground truth was obtained by investigating information available from the original dataset[14], previous papers [15], articles available online[16], [17] and the request for immediate managers issued by FERC1 which contains the job role and the immediate supervisor of 480 Enron employees[18].

In total, we chose 7 categories which reflect the hierarchical level of each employee from their organisational role based upon the generalisation of the key roles described in the official FERC report [18]. These categories are similar in nature to previous research articles [14]. Below we present the 7 categories.

- **Chief Officer (CO):-** The 11 senior C-Suite Officers in their divisions e.g. CEO etc.
- **Vice President (VP)**:- 24 employees with divisional control of 100+ employees.
- **Director**:- 24 employees who control larger teams (>60)
- **Manager**:- 29 employees with control of up to 10 employees.
- **Trader**:- 37 low-level employees who perform the day-to-day trading.
- **Specialist**:- 17 employees with specialist roles (such as IT administrator).
- **Assistant**:- 5 personal assistants to senior VPs and CO's.

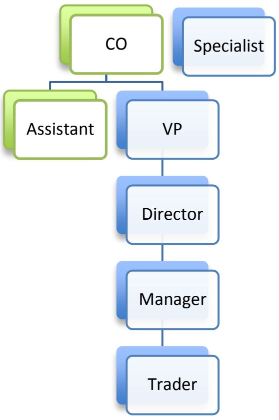


FIGURE 1:- HIERARCHY OF THE ENRON CORPORATION

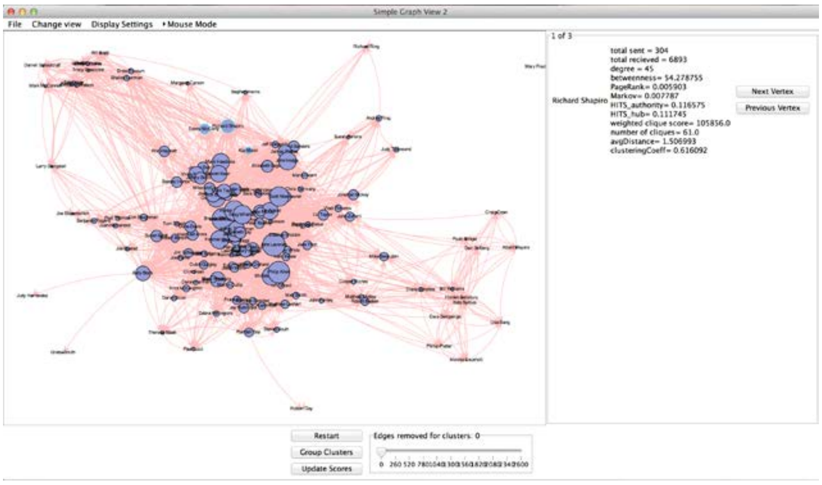
Figure 1 shows the visual representation of the categories. We leave the "Specialist" category separate from the main chain of

ISSN 2052-8604

hierarchy as these individuals interact with all members of the organisation at the different levels of hierarchy and move between groups within the organisation.

Tool Support

Over the last few years several SNA tools have been developed for different purposes such as Gephi[19], GraphViz[20], VisOne[21], Netlytic[22], UCINet[23] and Socilyzer[24]. Whilst these are all ideal for their own purposes, none provided us with all the analysis that would be needed in order to calculate the selected metrics. As such, we decided to create our own tool that would allow us to calculate all the metrics identified in the previous section in the same software. Figure 2 shows a representation of our social network with our new tool.



```
total sent = 304
total recieved = 6893
degree = 45
betweenness= 54.278755
PageRank= 0.005903
Markov= 0.007787
HITS_authority= 0.116575
HITS_hub= 0.111745
weighted clique score= 105856.0
number of cliques= 61.0
avgDistance= 1.506993
clusteringCoeff= 0.616092
```

FIGURE 2:- IMAGE OF OUR NEWLY CREATED TOOL SUPPORT

Results from our initial experiment

In our first experiment, we evaluated the effectiveness of our metrics by their ability to distinguish between the 7 categories defined previously. **Error! Reference source not found.** shows the breakdown of the metrics on a category-by-category basis. A full breakdown of our results can be seen in

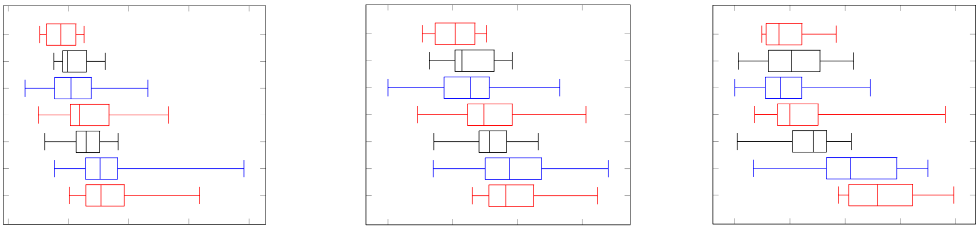


FIGURE 3:- DISTRIBUTION OF OUR TOP THREE SNA METRICS

ISSN 2052-8604

our paper[1].

The Markov Centrality scores were capable of separating off the VPs and COs and the remaining categories, but provided little distinction between other categories. Similarly, the PageRank Scores were able to provide some level of distinction between senior employees and the other categories.

As the HAS and HHS are closely related, we would have expected similar performance from both. Our initial results confirm this and showed that the HAS was significantly greater on average for VPs and Cos when compared to the other categories, leading us to believe that HAS may be a useful indicator of seniority within the network, however there are still some VPs with a low HAS. Our results also showed that the WCS outperformed CS. If an individual has a Weighted Clique Score greater than 200,000, then they have a high likelihood of being in one of the more senior categories. Conversely, all of our traders had a score less than 200,000. This leads us to believe that there may be a stronger correlation between the WCS and the employee category than between the CS and employee category.

Both the NSM and NRM were useful in identifying assistants, managers and directors as they sent comparatively fewer messages, but was not able to help distinguish further. The DCS and the BCS were useful in distinguishing some of the categories. The DCS metric proves effective at distinguishing between COs/VPs and other categories, and was useful in identifying senior employees. The BCS was able to help highlight COs and other senior members within the organisation, but several mid- seniority Managers also had high scores and these outliers may restrict the metric’s utility.

The ADS were noticeably good at distinguishing between the COs and the other categories (with the exception of Assistants) as COs tended to have an ADS of 1.5 or greater whereas those that were not in a position of authority had a lower ADS. It was less good, however, at distinguishing between the employees of lower seniority. The CCS proved ineffective when attempting to find a correlation with the employee category. Alone, it gave little insight into the difference in employee categories.

[1] http://www.cs.ox.ac.uk/people/elizabeth.phillips/

ISSN 2052-8604

Summary

Many of the conclusions from our initial analysis coincide with some real world assumptions. The ADS, for example, was expected to provide a good distinction between COs and other categories as most employees would not contact the CO directly but would communicate through their line manager.

Similarly, due to the nature of the HHS and HAS metrics, a higher HAS for senior management is expected as lower hubs (i.e. employees of lower seniority) would send several messages to them and they would also send numerous messages to lower-seniority employees. The WCS was expected to be useful as many COs would be the critical nodes in the graph and as such, would be part of many more complete sub-graphs (and in turn, gain a higher WCS). From the initial investigation, it emerged that there are a number of potentially useful metrics that can aid in identifying individuals of hierarchical importance within an organisation or group. We therefore decided to test these metrics in order to assess their effectiveness in a more rigorous manner.

Enhancing discovery of social groups and hierarchies

In order to calculate the social structure, we applied a Machine Learning approach to associate the metrics with the role Category. This would allow us to use the metrics obtained above and the ground truths to train a model that would predict the employee’s category based only on the SNA metrics of the employee.

Actual Category	Classified as						
	CO	VP	Director	Manager	Trader	Specialist	Assistant
CO	9	1	0	0	0	0	1
VP	10	4	6	1	3	0	0
Director	1	4	6	0	13	0	0
Manager	2	3	4	0	20	0	0
Trader	1	2	7	1	26	0	0
Specialist	0	3	2	0	11	1	0
Assistant	1	0	1	0	1	1	0

Table 2:- Confusion Matrix for initial Categories

To test the ability of the supervised learning algorithm to predict the employee category, we began by testing the dataset using a Bayesian Network Classifier. In order to validate the created models, we used 10-fold cross validation. Table 2 shows the classification results of the Bayesian Network model in a confusion matrix. The table revealed that 20 of the 29 Managers were incorrectly classified as Traders. This discrepancy could be due to the structure of the underlying network. Within the Enron corporation, many individuals were assigned the role of a manager but were only managers of small teams and were performing the role of a trader. This problem is exacerbated further due to the discrepancies between the ground truth sources.

In order to address this problem, we reduced the number of categories from seven to two, as we were primarily interested in identifying the senior employees. The new "Boss" category corresponded to the previous CO and VP categories whilst the "Not_Boss" corresponded to the remaining five categories. Despite the lower level of granularity of the employer's category that we were now able to predict, it allowed us to focus on highlighting the employees of greatest interest within the organisation.

Breakdown of reclassified data

Table 3 shows the statistical breakdown of the network once they have been reclassified using the 2 new categories while Table 3 shows a breakdown of some of the most useful metrics. From the analysis of the figures, we were able to identify the metrics that have a different distribution of values for each category, which in turn makes them potentially useful contributors to the Machine Learning algorithm in order to distinguish between the two categories. In particular, ADS, DCS, HAS, WCS and MCS all showed a distinction between the two categories and hence they may be useful metrics.

Attribute	Category	REC	DEG	BC	PR	MAR	HAS	HHS	WCS	CS	ADS	CC
Max	Boss	6893.00	132.00	1889.20	0.0196	0.0170	0.20	0.28	369852.00	490.00	1.70	0.70
Max	Not Boss	2972.00	92.00	1507.12	0.0133	0.0153	0.19	0.27	338456.00	360.00	1.67	1.00
Min	Boss	216.00	22.00	26.81	0.0051	0.0065	0.05	0.02	692.00	14.00	1.39	0.24
Min	Not Boss	0.00	1.00	0.00	0.0014	0.0000	0.00	0.00	1.00	1.00	1.29	0.00
Mean	Boss	1414.20	56.94	335.43	0.0089	0.0101	0.12	0.11	106370.29	114.14	1.52	0.48
Mean	Not Boss	530.18	28.80	128.95	0.0057	0.0066	0.05	0.04	13218.01	37.08	1.43	0.56
StdDev	Boss	1505.63	24.30	383.09	0.0035	0.0029	0.04	0.06	111219.20	109.60	0.06	0.11
StdDev	Not Boss	583.48	16.91	208.03	0.0021	0.0029	0.03	0.04	45734.44	51.01	0.07	0.17

TABLE 3:- RESULTS FROM RECLASSIFIED DATA

Once we had created our two new categories, we tested the effectiveness of our new model using a variety of different Machine Learning Methods. In total we selected seven models, namely Naive Bayes (NB), Bayesian Network (BN), Multi-Layer Perceptron Model (MLP), IB1, K-Star and SMO, and compared them to random guessing. The overall best performing classifier is the MLP, with the NB and BN close behind by providing a greater True Positive (TP) rate for the Boss category and producing a greater Receiver Operating Characteristic (ROC) curve area and F-Score. A higher F-Score and ROC curve area is an indication of a good classifier.

Random Guessing						
TP	FP	Precision	Recall	F-Score	ROC Area	Class
0.522	0.455	0.934	0.522	0.67	0.5	Not_Boss
0.545	0.478	0.085	0.545	0.146	0.5	Boss

Multi-Layer Perceptron						
TP	FP	Precision	Recall	F-Score	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

TABLE 4:- ANALYSIS OF OUR MLP MODEL VS. RANDOM GUESSING

Table 4 shows a breakdown of the results for our MLP model vs. random guessing. Our results show us that by categorising the Enron dataset into two categories and by introducing the new metrics and categorisations, we have been able to predict whether an individual is a Boss with an F-Score of 0.64 and an ROC Area of 0.939 compared to random guessing which achieved 0.146. It also identified five critical attributes, namely

WCS, ADS, HAS, HHS and DCS. This has enabled us to improve on existing metrics, which are accurate to only 82.37% [25]and 87.58%[26] respectively.

Summary

From our analysis using our new role categories, we were able to identify five metrics that have a different distribution for Bosses than ordinary employees, which in turn can make them useful contributors to our model to predict the employee's role category. In order to quantify how effective each metric was, we decided to use machine learning metric evaluators. In particular we used the Relief-F evaluator[27] which was chosen for its consistency and its ability to cope with the dependence between our attributes.

Experiment 2

For our second experiment, our new dataset was considerably smaller than the Enron dataset and represented the communications amongst a single group. For this group, we collected a total of 6,936 emails sent amongst the ten members of the group over a twelve-month period from 20 June 2013 to 20 June 2014. Each email was sent to an average of 1.97 recipients. As our data-collection scripts hide the identity of email recipients of emails sent outside of the group, the actual number of recipients in an email may well have been much higher than this.

After establishing our initial network, we then proceeded to collect the ground truth for the actual hierarchical structure of the network. Within this network, there was one official Boss for the research group (Employee #0) who acted as the main supervisor for many (but not all) of the projects. Employee #4 was also in a unique position as they had worked on a variety of different projects with various members of the group in the past. They are considered a senior member in the group because of the various interactions across projects (often simultaneously) and we therefore categorised employee #4 as a Boss as well.

From the initial network, we discovered that the graph was almost fully connected, with 84 out of the 90 possible edges between the ten employees established based on their email communication which led to some of our SNA metrics being ineffective as they were unable to differentiate important connections from insignificant ones. For each email sent, we add 1 to the thickness of each graph edge. The distribution of weights was expected given the small size of the group and the interaction between members for non work-related purposes associated with a close research group.

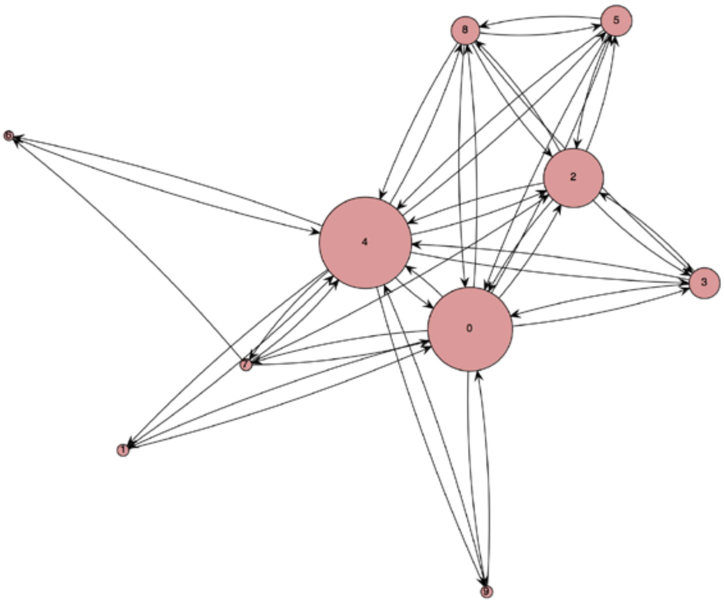


FIGURE 4:- GRAPH REPRESENTATION OF OUR NEW

In order to overcome this, we decided to only consider edges of weight 30 or more in order to only identify strong ties between members. Whilst this pruning might lead to us potentially missing some important connections, it is more important to prune the edges that may not have been central to the work-focused network. 17 shows the structure of the new network with nodes sized according to their Authority score and edges of weight 29 or less removed from the network and is laid out using the force-directed layout of Fruchterman[28] and uses the notion of “force” and connectivity between nodes and their edges to determine where they should be placed.

The graph immediately identifies employee #4 and employee #0 as strongly connected nodes due to their close positioning in the graph (with 1095 emails sent between the 2 employees). It also identified employee #9 as an employee that is linked to only a few members in the group; this reflects the fact that employee #9 only worked on one project with the 2 senior members of the group and as such, had little collaboration with other members. Similarly, Employee #6's distance from the cluster reflects the fact that they had only recently joined the group (March 2014).

Results from experiment 2

Employee	WCS	HAS	DCS	ADS	MCS	Class
0	6.168	0.459	0.8	1.900	0.195	Boss
1	1.149	0.175	0.2	1.563	0.043	Not_Boss
2	3.870	0.409	0.55	1.692	0.128	Not_Boss
3	1.320	0.314	0.35	1.600	0.088	Not_Boss
4	7.316	0.471	0.9	2.000	0.249	Boss
5	1.320	0.314	0.4	1.643	0.087	Not_Boss
6	1.149	0.144	0.15	1.529	0.032	Not_Boss
7	2.380	0.175	0.3	1.643	0.044	Not_Boss
8	2.639	0.302	0.45	1.692	0.092	Not_Boss
9	1.149	0.175	0.2	1.563	0.043	Not_Boss
Average	2.846	0.294	0.43	1.682	0.100	
StdDev	2.252	0.124	0.254	0.153	0.072	

Table 5:- Results from Experiment 2

Table 5 presents the distribution of employees and their metric scores. The results of our new analysis support our previous theory as both employee #4 and #0 are the true Bosses and have notably higher scores than the other employees. All other employees' scores are less then 1 standard deviation above the mean for each of the top 5 metrics. This finding strengthens our initial belief that these metrics are a good measure of "hierarchical importance" within an organisation.

The results of our second experiment demonstrated that the metrics identified in Experiment 1 performed as expected and were reasonably effective at distinguishing between the two employee categories. This confirmed the utility of using the 5 metrics (especially the Weighted Clique Score) in allowing the inference to be made from email-communication metadata to the hierarchical structure of a group.

The work assumes that supervisors and bosses are active users of email in order for the communication network to reflect the true communications within the network. Whilst some management styles prefer to use other tools (such as phone calls) to communicate, if we were able to collect this form of data, then our abstraction of the email communications to a social network would allow it to be incorporated into our network by increasing the edge weight based on the type of communication, so as to create a new network which better reflects the underlying hierarchy, on which we can perform the same SNA analysis.

Conclusions and Future Work

Our results have identified five SNA metrics which have proved effective in distinguishing between the employees that are assigned a Boss category and those who are assigned to a Not_Boss category based only on the email communications between them; namely Weighted Clique Score, HITS Authority Score, Average Distance, Markov Centrality Score and Degree Centrality Score.

The primary value of our research is the improvement in selecting and improving on existing metrics whilst using the minimum amount of data, so as to enable the methods to be applied to any generic communications network including Dark Net Forums, Social Networking Sites as well as phone records and other offline communication networks such as face-to-face meetings.

One direction of future research is to apply our metrics to a communications network established from other sources such as the 2012 dataset extracted from the ISI-KDD Challenge of the Dark Web forums 2[2]. This should allow us to identify the most influential contributors to the forum which may help identify the ring-leaders of criminal groups that use the forums. Another direction our research could take is within Insider Threat Detection within organisations. This in turn could be a feature of Machiavellianism, which as one of the Dark Triads personality traits [29]could be a potential predictor for a malicious insider. Further research would be required to investigate to what extent uncharacteristically high influence relates to Insider Threat Detection.

References

[1] L. C. Freeman, “***Centrality in social networks conceptual clarification***,” *Soc. Netw.*, vol. 1, no. 3, pp. 215–239, 1979.

[2] J. Travers, S. Milgram, J. Travers, and S. Milgram, “***An Experimental Study of the Small World Problem***,” *Sociometry*, vol. 32, pp. 425–443, 1969.

[2] Available at http://128.196.40.222:8080/CRI Indexed new/datasets/ansar1.txt

[3] Sara Radicati, “***Email Statistics Report, 2014 - 2018***,” Radicati Group, Apr. 2014.

[4] Sara Radicati, “***Email Statistics Report, 2012 - 2016***,” Radicati Group, Apr. 2012.

[5] Chron, “T***he Use of Email in Business Communication***,” *Small Business - Chron.com*. [Online]. Available:http://smallbusiness.chron.com/use-email-business-communication-118.html. [Accessed: 22-Jun-2014].

[6] Atul Kachare, “***Analysis and Visualization of E-mail Communication Using Graph Template Language***,” *SAS Glob. Forum*, 2013.

[7] L. Sproull and S. Kiesler, “***Reducing Social Context Cues: Electronic Mail in Organizational Communication***,” *Manag. Sci.*, vol. 32, no. 11, pp. 1492–1512, Nov. 1986.

[8] “***Edward Snowden***.” [Online]. Available: http://www.theguardian.com/world/edward-snowden. [Accessed: 21-Jun-2014].

[9] D. Wright and R. Kreissl, “***European Responses to the Snowden Revelations: A Discussion Paper,***” IRISS, Dec. 2013.

[10] L. Getoor and C. P. Diehl, “***Link Mining: A Survey,***” *SIGKDD Explor Newsl*, vol. 7, no. 2, pp. 3–12, Dec. 2005.

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

[12] T. Coffman, S. Greenblatt, and S. Marcus, “***Graph-based Technologies for Intelligence Analysis***,” *Commun ACM*, vol. 47, no. 3, pp. 45–47, Mar. 2004.

[13] Park, “Enron employee status.” [Online]. Available:http://cis.jhu.edu/~parky/Enron/employees. [Accessed: 24-Jun-2014].

[14] J. Shetty and J. Adibi, “***The Enron email dataset database schema and brief statistical report***,” *Inf. Sci. Inst. Tech. Rep. Univ. South. Calif.*, vol. 4, 2004.

[15] G. Creamer, R. Rowe, S. Hershkop, and S. J. Stolfo, “***Segmentation and automated social hierarchy detection through email network analysis***,” in *Advances in Web Mining and Web Usage Analysis*, Springer, 2009, pp. 40–58.

[16] “***John Arnold: Ex-Enron billionaire trader retires at 38 | Mail Online,***” *Daily Mail Online*. [Online]. Available: http://www.dailymail.co.uk/news/article-2138890/John-Arnold-Ex-Enron-billionaire-trader-retires-38.html. [Accessed: 15-Jun-2014].

[17] R. Partington, “***The Enron cast: Where are they now? - Financial News,***” *Financial News*. [Online]. Available: http://www.efinancialnews.com/story/2011-12-01/enron-ten-years-on-where-they-are-now. [Accessed: 15-Jun-2014].

[18] federal energy regulatory commission subpoena duces tecum, “Request no. 11” [Online]. Available: https://raw.githubusercontent.com/diehl/Enron-GraphML-Data-

ISSN 2052-8604

Documentation/master/EnronManagerSubordinateRelationships.pdf. [Accessed: 15-Jun-2014].

[19]“***Gephi, an open source graph visualization and manipulation software.***” .

[20]J. Ellson, E. Gansner, L. Koutsofios, S. North, and G. Woodhull, “***Graphviz— Open Source Graph Drawing Tools,***” in *Graph Drawing*, vol. 2265, P. Mutzel, M. Jünger, and S. Leipert, Eds. Springer Berlin Heidelberg, 2002, pp. 483–484.

[21]***“visone.”*** [Online]. Available: http://visone.info/. [Accessed: 22-Jun-2014].

[22]“***Netlytic.org.***” [Online]. Available: https://netlytic.org/home/. [Accessed: 24-Jun-2014].

[23]S. P. Borgatti, M. G. Everett, and L. C. Freeman, ***Ucinet for Windows: Software for Social Network Analysis***. Analytic Technologies, 2002.

[24]“***An Easy-to-Use Social Network Analysis Tool - Socilyzer.***” [Online]. Available: https://socilyzer.com/. [Accessed: 24-Jun-2014].

[25]E. Gilbert, “***Phrases That Signal Workplace Hierarchy,”*** in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, New York, NY, USA, 2012, pp. 1037–1046.

[26]A. Agarwal, A. Omuya, A. Harnly, and O. Rambow, “***A Comprehensive Gold Standard for the Enron Organizational Hierarchy,***” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, Stroudsburg, PA, USA, 2012, pp. 161–165.

[27]M. Robnik-Sikonja and I. Kononenko, “***An Adaptation of Relief for Attribute Estimation in Regression***,” in *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, CA, USA, 1997, pp. 296–304.

[28]T. M. J. Fruchterman and E. M. Reingold, “***Graph Drawing by Force-directed Placement***,” *Softw Pr. Exper*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991.

[29]J. McHOSKEY, “Narcissism and machiavellianism,” *Psychol. Rep.*, vol. 77, no. 3, pp. 755–759, Dec. 1995.