Analysis of E-mail Communication Activities for Detecting Patterns of Pathological Behaviour

Michael Negnevitsky, Mark Jyn-Huey Lim
School of Engineering and ICT (Hobart)
University of Tasmania,
Hobart, Tasmania 7001, Australia

Jacky Hartnett
School of Engineering and ICT (Launceston)
University of Tasmania,
Hobart, Tasmania 7001, Australia

***Abstract*— E-mail is one of the most popular and widely used form of electronic communication used today. The patterns in the social interactions or contacts between people by e-mail can be analysed using social network analysis and user behaviour analysis. In this paper we provide a review of the work related to the areas of dynamic modelling and link prediction of social networks, and anomaly detection for detecting changes in the behaviour of e-mail usage. We then discuss about the benefits of applying artificial intelligence techniques to these fields.**

I. Introduction

One of most common forms of electronic communication in use today is electronic mail or e-mail. The use of e-mail has made a large impact on society in the way people communicate with each other, because it is easy to write, quick to send, and allows a single message to be sent to large groups of people. The result of these features of e-mail has made it a popular and wide-spread form of electronic communication that people use to communicate and socialise with each other. This has provided a suitable environment for researchers to study the social interactions of individuals over e-mail, part of a field of study called social network analysis [1], [2].

The observation of e-mail communication social networks can be represented as a type of complex network, where each vertex of the network represents a person or individual and each edge represents the interaction or contact between people, displaying a type of graph. A simple diagram of e-mail communications involving the authors of this paper, shown in Fig. 1, provides an example of how an e-mail communication social network could be represented. In the diagram, weights can be assigned to the graph edges to indicate the strength of the communication link, which in this case is used to describe the frequency of e-mail transmissions between particular individuals in the e-mail communication social network.

The focus of this research will be to monitor the e-mail communication links between individuals or groups of people for any changes in the social network. The types of changes that are to be considered are changes in the communication activity between particular individuals (e.g. changes in rate of e-mail transmissions) and predicting the additions of new links to the e-mail communication social network. Much of this research will look at e-mail communications that is not confined within the boundaries of an organisation such as a company or academic institution. Rather, the type of e-mail communications considered will be of a global nature, where the individuals making up the social network are spread out across large geographic areas and are not necessarily part of any structured organisation.

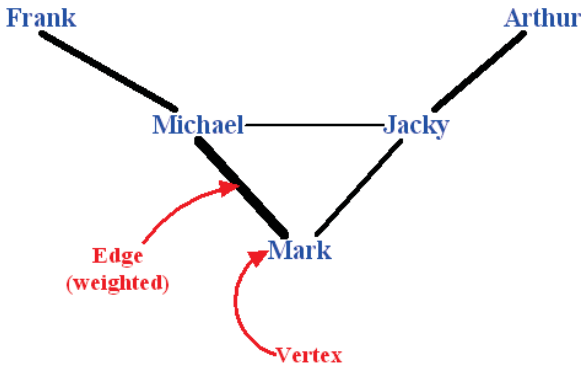

Fig. 1. Example of an e-mail communication social network

There are several applications where our research could be useful. Firstly, for the homeland security and intelligence gathering community, the techniques of monitoring for changes in e-mail communication activities could be useful for tracking the electronic communication of terrorist suspects for any clues of an upcoming terrorist attack. The detection of increased e-mail traffic between particular parts of the terrorist social network over a particular period of time can be helpful in determining when and where an upcoming event is likely to take place. A second useful application is for law enforcement agencies, where the same techniques used for monitoring terrorist suspects is also applied for tracking the electronic communications of criminal organisations and predicting the likely occurrence and location of an upcoming event. Another third application is for tracking and monitoring the spread of e-mail-borne computer viruses through e-mail social networks. The detection of unusual variations in e-mail communication activities can be used to locate individuals in the social network who are currently affected by an e-mail-borne computer virus. This can then be used to determine how quickly the e-mail virus will spread via the social contacts of the infected individual and how many 'hops' it will take for the e-mail virus to spread globally via e-mail social networks.

This paper reviews two main fields of research that have applications for analysing e-mail communications and discusses how the results from these areas could be assisted with the use of artificial intelligence. Firstly we highlight some of the previous work conducted in the area of complex

ISSN 2305-7254

networks, a field that provides an overall view of a system and enables the analysis of various properties and characteristics of the system's network topology. The second part reviews some of the studies conducted in user behaviour analysis, which examines the on-line behaviour of an individual or group of people on the network. The paper then discusses about how the use of artificial intelligence could add value to the study of complex networks and user behaviour analysis in e-mail communications.

II. Complex networks

The study of social networks focuses on finding patterns in the way people interact or contact with each other. This is just one of the many types of networks being studied under the broad field of complex networks [1]. Complex networks examine the statistical properties of large-scale networks, which may contain millions or billions of vertices. Although this field is primarily in the area of physics, its applications are interdisciplinary in that they can be applied to any discipline that requires the understanding of complex interactions or processes. The various disciplines that complex networks can be applied to include computer science (e.g. peer-to-peer computer networks, Internet, World Wide Web, e-mail communications), electrical engineering (e.g. power system grids, telephone networks), geology (e.g. river networks), or biology (e.g. metabolic pathways, food web of predator-prey interactions), just to name a few disciplines. Complex networks is a field that has been receiving growing interest in recent years and an in-depth survey about some of the recent work in complex networks is given by [1].

A. Statistical properties

There are several statistical properties of complex networks described by [1]. Properties of interest for e-mail communication social networks include the following:

1) The Small-World Effect: The small world effect is a network property that describes the average shortest path (mean geodesic path) between any two vertices in a network. This shows on average the shortest distance required for traversing from one vertex to another in a network [1]. An investigation of the small-world effect using e-mail has been studied by [3], where they carried out an experiment in which each of the participants were given a message that needed to be delivered to a randomly selected individual. The message then had to be delivered via people whom the participant were acquainted with. On average the study found it took at least five to seven steps to reach the targeted individual. This type of property in social networks is often termed "six degrees of separation" and is useful for e-mail social networks in that it describes how quickly information can pass via one individual to the next in the network.

2) Transitivity or Clustering: The network property of transitivity or clustering, describes the ratio of connectedness between sets of vertices, described as triples, in a complex network [1]. Essentially the property describes that if vertex A is connected to vertex B, there is an increased probability that vertex A will be connected to vertex C. In terms of a social network, it describes the probability that a friend of your friend is likely to be also your friend. Studies have been conducted by [4] and [5] that investigate the clustering effect in several real-world networks, such as the Hollywood actors collaboration network, language (word-to-word semantic) network, World Wide Web, the interdomain/autonomous system level Internet, the router level Internet and power grid. For e-mail social networks, clustering describes how closely connected each individual is to their neighbours, in terms of how "closely-knit" each individual is to one another.

3) Community Structure: The community structure network property seeks to describe how vertices in a network are clustered together into groups of vertices with a high density of edges between them [1]. In the context of a social network, this property describes how people naturally divide into social groups based on shared attributes such as for example: common interests, age or occupation. A study by [6] of the e-mail collected from a university in Spain demonstrates how community structures can be extracted from the e-mail data to provide visual representations of the different groups in the e-mail social network. This network property is useful for analysing e-mail social networks in that it allows one to examine where various individuals in the network form social groups and to determine the likely paths for messages to pass between individuals.

B. Models of complex networks

Studies in complex networks often involve building models of networks (e.g. Fig. 2) to compare the simulated statistical properties of the model to those found in real-world networks (e.g. Internet, World Wide Web, paper citation network). Two main approaches for modelling complex networks are:

1) Static Modelling: The static modelling of complex networks aims to create models of networks with particular structural properties such as the small-world effect, clustering and community structure [1]. There are several models that are useful for comparisons with empirical studies of e-mail social networks [7] and these include the small-world model [8], scale-free model [9] and community structures model [10], [11]. Static modelling is useful in that one can easily build a complex network, containing more than a thousand vertices, without having to collect large amounts of experimental data in order to observe particular properties of complex networks. It is also useful in that the model of the network can be constructed quickly and viewed in a complex network visualisation programs such as Pajek [12].

2) Dynamic Modelling: A disadvantage with the static modelling approach is that it is unable to provide information on how the network topology has evolved over time to obtain a particular network structure. Dynamic modelling or growth modelling techniques take into account the influence that network topology evolution has on the structure of the network [1]. There have been many techniques developed for modelling evolving network topology, some of which includes the growth of transitivity/clustering by increasing links between existing vertices [13], and growth using preferential attachment where new vertices added to the graph are attached preferentially to high degree vertices (vertices with a large

number of connections) [5]. At this stage, there has not been any techniques found that closely model the evolution of e-mail social networks.

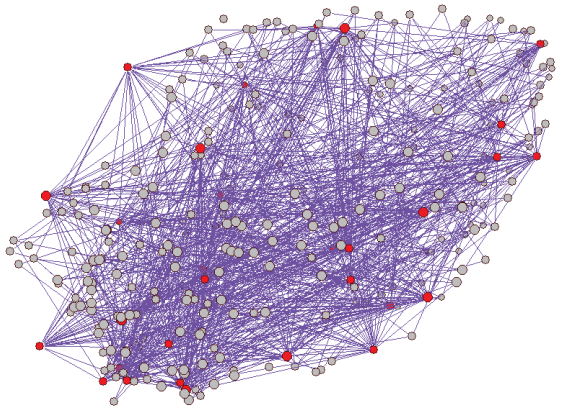

Fig. 2. A complex network generated with 300 vertices and 853 randomly chosen edges

C. *Link prediction*

The current approaches in dynamic modelling of complex networks still do not model the intuition used by individuals for selection of new links in growing social networks [14]. A different approach called link prediction, attempts to model the intuition used by individuals in social networks in order to predict the likelihood of new links being created. In [14], prediction is based on the proximity or 'closeness' of neighbouring individuals in the social network. For example if two people have similar interests, share colleagues in common and travel to similar locations, then there is a higher likelihood they will become acquainted with each other in the near future. [14] reviews several different mathematical techniques for link prediction. Some of the prediction techniques reviewed are methods based on node neighbourhood (prediction of links between two vertices based on shared neighbouring vertices), methods based on the ensemble of all paths (considers all the shortest paths in the network), and "higher-level" approaches (mathematical methods that can be used in conjunction with the previous two methods mentioned). For e-mail social networks, link prediction will be useful in that it will enable a better understanding about how external influences (e.g. occurence of external events such as conferences) affect the formation of new links and how it changes the topology of the network over time.

III. User behaviour analysis

User behaviour analysis examines the on-line actions of an individual or group of users on a computer network. It belongs to the field anomaly detection, which forms part of intrusion detection systems [15]. Anomaly detection analyses computer network traffic for network activity that is considered different from 'normal' network activity. The difference in activity is determined by building a profile of 'normal' activity from historical data collected over a period of 'normal' operation, then comparing the profile of current network activity to the 'normal' profile. If the profile of current activity deviates significantly from the 'normal' profile, then there is likely to be 'abnormal' activity taking place.

There have been a variety of methods used to monitor the on-line activities of users. One approach has been to monitor the command line sequences or keystrokes of users to determine the presence of masqueraders, who are computer users that use another person's computer account [16], [17]. In [16] they used statistical and probabilistic techniques (Uniqueness, Bayes 1 Step Markov, Hybrid Multi-Step Markov, Compression, IPAM and Sequence-Match) to determine masqueraders from legitimate users, while [17] used Naive Bayes classifiers.

Another approach for monitoring the on-line activities of users has been to examine e-mail traffic data for changes in traffic activity. The studies by [18], [19] use methods for differentiating 'normal' traffic generated by computer users from traffic generated by e-mail based computer viruses. The study by [18] simulates the behaviour of e-mail users and computer viruses using the 'specification-based anomaly detection' technique [20], which combines state-machine specifications of network protocols with statistical machine learning. In their specification-based model, they simulate the protocol interactions between email clients and email servers belonging to the same organisation, by manually mapping the interactions onto a state-machine model consisting of three states: INIT, RCVD and DONE. The anomaly detector then uses statistical machine learning to monitor the statistical properties of the state-machine transitions for any variations from the 'normal' statistical properties. Detection of significant variations indicates that e-mail viruses are causing changes to the e-mail traffic activity.

On the other hand, the study by [19] developed an on-line "behaviour-based" security system called the Malicious Email Tracking (MET) system. The MET system is assisted by an off-line analysis system called the Email Mining Toolkit (EMT), which is used to develop behaviour-based profiles of users for the MET system by analysing e-mail data collected from e-mail servers. The EMT system was designed with several analysis tools, two of which are related to anomaly detection and social network analysis: "Account Statistics and Alerts" and "Group Communication Models". The "Account Statistics and Alerts" mechanism uses histograms of e-mail usage to build profiles of users and examine the frequency of e-mail transmissions over short-term (i.e. over 24 hours) and long-term (e.g. up to a month) periods. Anomalies in the e-mail account behaviour are detected by using a weighted Mahalanobis distance function to compare the difference between the recent short-term histogram profile of the user to their long-term histogram profile. In the "Group Communication Model" mechanism, they use a clique finding algorithm [21] to find e-mail accounts that form a social group. These social groups are then profiled by creating a frequency table of all e-mails sent by each group member, to monitor the overall behaviour of the group. The anomaly detector then monitors e-mail traffic for any deviations from the 'normal' frequency table of each group to identify the presence of e-mail-borne viruses.

IV. Potential applications for AI

A. AI in complex networks

There are some potential applications for the use of AI in complex networks in the areas of dynamic modelling and link prediction. In dynamic modelling, the current approaches lack the mechanism for incorporating the intuitive reasoning used by individuals when forming new links in social networks. This is an area where fuzzy systems or neural networks could be applied [22], to account for some of the intuition mechanisms missing in current dynamic modelling approaches. Neural networks have already been applied for growth modelling by [11], where they use it to more precisely model the growth of real-world networks, such as the World Wide Web. They achieved this by allowing their neural network to learn from the changing topology of real World Wide Web data, in order to more closely model the changes in network topology.

For link prediction, the current approaches reviewed by [14] only use mathematical techniques and found that the best predictor, Katz clustering (ensemble of all paths method), was only correct for up to 16% of the predictions. This shows that there could be further improvement in the rate of correct predictions by the use of fuzzy systems or neural networks. Again, neural networks have already been applied to this area by [11] where they use it to predict new links based on real network topology data modelled from the World Wide Web.

B. AI in user behaviour analysis

The approaches for anomaly detection described by [16], [17], [18], [19] only use mathematical techniques to detect 'abnormal' variations in the behaviour of users. The problem with using mathematical techniques is that they impose strict boundaries around the profiles of what is considered 'normal' and 'abnormal' behaviour. The slightest deviation from the 'normal' profile while the actual network activity is still operating normally will cause the anomaly detector to give off a false alarm. Incorrect identification can also occur if the actual network activity is operating abnormally, but does not have the profile that the anomaly detector considers to strictly match an 'abnormal' profile, resulting in missed detection. These problems with imposing strict boundaries on the 'normal' and 'abnormal' profiles can be seen from the results of the false alarm and miss rates of [16], [17], [19], where a summary of their best results are shown in Table I.

TABLE I: The best false alarm and miss rates of anomaly detectors using mathematical techniques

Technique	False Alarm	Miss Rate
Uniqueness [16]	1.4%	60.6%
Naive Bayes [17]	1.3%	38.5%
Histogram/Statistical [19]	4%	41%

Because of the uncertainty in defining the boundaries between what is 'normal' and 'abnormal' user behaviour when analysing command sequences or e-mail traffic behaviour of users, this is an ideal area for fuzzy systems to be applied [22]. Some studies have already been conducted using fuzzy systems for anomaly detection, such as in [23]. In [23], they applied fuzzy logic and data mining techniques for anomaly detection of TCP-level network traffic.

Other studies where AI has been applied for anomaly detection have been the use of neural networks to learn and identify the profiles of 'normal' and 'abnormal' behaviour [24]. The advantage of using neural networks in anomaly detection is that features of 'normal' and 'abnormal' behaviour can easily be learned by the neural network, as opposed to applying mathematics to describe the features of the data to the anomaly detector. The applications of either fuzzy systems or neural networks will be ideal for analysing e-mail communication traffic activity.

V. Overview of e-mail traffic analyser system

In this project, we developed an e-mail traffic analyser system that will enable us to extract different types of e-mail traffic behavioural patterns to obtain information on the behaviour of e-mail users. The different types of e-mail traffic behavioural patterns we are analysing are: the social connections between e-mail users, the level of e-mail usage by e-mail users (e.g. number of e-mails sent per day, number of e-mails received per day), and the level of interaction between different e-mail users (e.g. how often a user sends e-mail to a particular individual, how quickly a user responds to e-mails received from other users). These e-mail traffic behavioural patterns are being analysed by the e-mail traffic analyser system, to allow us to explore the data and extract "interesting" behavioural patterns from each e-mail user. Examples of "interesting" e-mail traffic behavioural patterns could be where an e-mail user suddenly starts sending more e-mails to a particular individual, where an e-mail user stops communicating with a particular individual, or a period of time where there is a change in the level of interactions between particular email users. The methods for extracting the "interesting" e-mail traffic behavioural patterns will be based on artificial intelligence techniques, which forms the main part of our investigations with the e-mail traffic analyser system.

In the e-mail traffic analyser system, e-mail data is processed through several stages to ensure that the data is clean and is in a suitable format before information is extracted [25]. The diagram in Fig. 3 provides an overview of the e-mail traffic analyser system and shows how the e-mail data from the e-mail system is processed to enable for behavioural patterns to be extracted for different e-mail users. There are two main stages for processing the e-mail data in the e-mail traffic analyser system: the "Data Acquisition and Filtering Stage" and the "Information Extraction Stage".

At the first stage, "Data Acquisition and Filtering Stage", e-mail data is collected from the e-mail system and cleaned/filtered to fill in missing values, remove noise, and fix up inconsistent data. Once the e-mail data is filtered and cleaned it is stored into the e-mail traffic database.

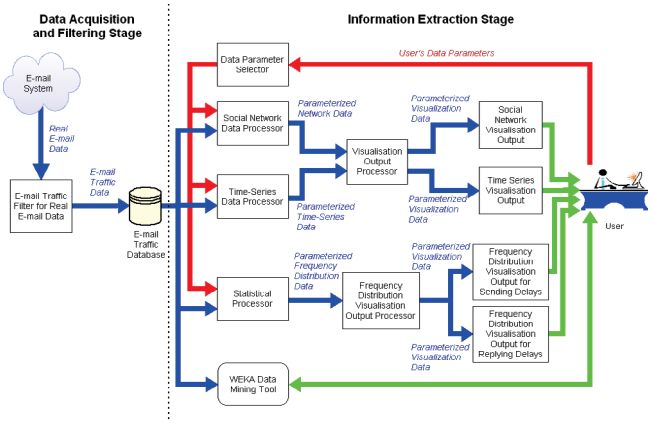

Fig. 3. Overview of e-mail traffic analyser system

At the second stage, "Information Extraction Stage", information from the e-mail traffic data is extracted and processed through different components to provide details on the e-mail traffic behavioural patterns of social connections ("Social Network Data Processor" component), level of e-mail usage ("Time-Series Data Processor" component), and level of interaction between e-mail users ("Statistical Processor" component). After these e-mail traffic behavioural patterns have been extracted and processed, they are visualised and presented to the user for analysis.

However, it is a difficult task for the user to pick out the "interesting" patterns from the visualisation outputs, because the user is presented with so much visual information. So, to assist with the task of finding the "interesting" e-mail traffic behavioural patterns, the WEKA Data Mining Tool program [26] is being used in the "Information Extraction Stage" of the e-mail traffic analyser system, to assist the user in finding the "interesting" patterns from the e-mail traffic data. After the user has found some interesting patterns through the use of the WEKA Data Mining Tool, they can then focus their attention on the areas in the e-mail data where the interesting e-mail traffic behavioural patterns were found. This can be done through the use of the "Data Parameter Selector" component, which allows the user to "zoom-in" on the details of the interesting patterns by the specifying a combination of different data selection parameters (e.g. a specific period of time, a specific group of e-mail users, selecting all e-mail messages being received a particular user). The process of extracting information from the e-mail traffic data, presenting the information to the user, and allowing the user to focus on particular details in the e-mail traffic data, provides a more interactive way for the user to analyse the e-mail traffic data for interesting behavioural patterns. To help develop and test the e-mail traffic analyser system, a conceptual simulation model of the e-mail system has been developed to generate different types of behavioural traffic patterns for the analyser system to examine. The simulated e-mail system is used to substitute for the data provided by the real e-mail system, as shown in Fig. 4.

VI. Conclusion

In this paper we have provided a brief review of some of the work done in complex networks and user behaviour analysis, related to the analysis of e-mail communications. Previous works in these fields have concentrated on using mathematical approaches. However we have shown that the analysis of e-mail communications is also suitable for applying AI techniques.

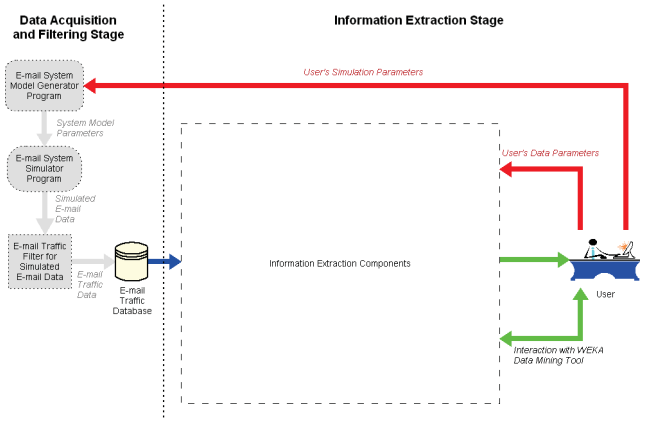

Fig. 4. Overview of how the e-mail system simulator is used with the e-mail traffic analyser system

References

[1] M. E. J. Newman, "The structure and function of complex networks," *Siam Review*, vol. 45, no. 2, pp. 167–256, 2003.

[2] M. E. J. Newman and P. Juyong, "Why social networks are different from other types of networks," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 68, no. 3, pp. 36122–1–8, 2003.

[3] P. S. Dodds, R. Muhamad, and D. J. Watts, "An experimental study of search in global social networks," *Science*, vol. 301, no. 5634, pp. 827–829, 2003.

[4] E. Ravasz and A. L. Barabasi, "Hierarchical organization in complex networks," *Physical Review E*, vol. 67, no. 2, pp. 26112–1–7, 2003.

[5] Alexei Vazquez, "Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations," *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 67, no. 5, pp. 056104–15, 2003.

[6] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E*, vol. 68, no. 6, 2003.

[7] Y. J. Tsai, C. C. Lin, and P. N. Hsiao, "Modeling email communications," *Ieice Transactions on Information and Systems*, vol. E87D, no. 6, pp. 1438–1445, 2004.

[8] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[9] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] M. E. J. Newman, "Properties of highly clustered networks," *Physical Review E*, vol. 68, no. 2, 2003.

[11] M. Kimura, K. Saito, and N. Ueda, "Modeling of growing networks with directional attachment and communities," *Neural Networks*, vol. 17, no. 7, pp. 975–988, 2004.

[12] V. Batagelj and A. Mrvar, *Pajek - Program for Large Network Analysis*, 2004, Viewed 11 October 2004, http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[13] E. M. Jin, M. Girvan, and M. E. J. Newman, "Structure of growing social networks," *Physical Review E*, vol. 6404, no. 4, pp. art. no.–046132, 2001.

[14] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM'03)*. 2003, pp. 556 – 559, ACM Press.

[15] R. Bace and P. Mell, *NIST Special Publication 800-31: Intrusion Detection Systems*, National Institute of Standards and Technology (NIST), 2001, Viewed 26 February 2004.

[16] M. Schonlau, W. DuMouchel, W. H. Ju, A. F. Karr, M. Theus, and Y. Vardi, "Computer intrusion: Detecting masquerades," *Statistical Science*, vol. 16, no. 1, pp. 58–74, 2001.

[17] R. A. Maxion and T. N. Townsend, "Masquerade detection augmented with error analysis," *IEEE Transactions on Reliability*, vol. 53, no. 1, pp. 124–147, 2004.

[18] A. Gupta and R. Sekar, "An approach for detecting self-propagating email using anomaly detection," in *Recent Advances in Intrusion Detection, Proceedings*, vol. 2820 of *Lecture Notes in Computer Science*, pp. 55–72. SPRINGER-VERLAG BERLIN, Berlin, 2003.

[19] S. J. Stolfo, S. Hershkop, K. Wang, O. Nimeskern, and C. W. Hu, "A behavior-based approach to securing email systems," in *Computer Network Security*, vol. 2776 of *Lecture Notes in Computer Science*, pp. 57–81. SPRINGER-VERLAG BERLIN, Berlin, 2003.

[20] R. Sekar, A. Gupta, J. Frullo, T. Shanbhag, A. Tiwari, H. Yang, and S. Zhou, "Specification-based anomaly detection: a new approach for detecting network intrusions," in *9th ACM Conference on Computer and Communications Security*, Washington, DC, USA, 2002, pp. 265–274, ACM Press.

[21] C. Bron and J. Kerbosch, "Finding all cliques of an undirected graph," *Communications of the ACM*, vol. 16, no. 9, pp. 575–577, 1973.

[22] M. Negnevitsky, *Artificial Intelligence: A Guide to Intelligent Systems*, 3rd edition, Addison Wesley, Essex, 2011.

[23] J. Guan, D. X. Liu, and T. Wang, "Applications of fuzzy data mining methods for intrusion detection systems," in *Computational Science and Its Applications - Iccsa 2004, Pt 3*, vol. 3045 of *Lecture Notes in Computer Science*, pp. 706–714. SPRINGER-VERLAG BERLIN, Berlin, 2004.

[24] A. R. Sharafat, M. Rasti, and A. Yazdian, "Neural network based anomaly detection in computer networks: a novel training paradigm," in *ISCA 16th International Conference: Computer Applications in Industry and Engineering*, Las Vegas, NV, 2003, pp. 50–53, ISCA, Cary, NC.

[25] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Series in DataManagement Systems, San Francisco, CA: Morgan Kaufmann, 2001.

[26] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, San Francisco, Calif.: Morgan Kaufmann, 2000.