A Survey on User Profiling Model for Anomaly Detection in Cyberspace

Arash Habibi Lashkari, Min Chen and Ali A. Ghorbani

Canadian Institute for Cybersecurity (CIC), University of New Brunswick (UNB)
Fredericton, Canada
Email: a.habibi.l@unb.ca; mchen5@unb.ca; ghorbani@unb.ca

Received 03 March 2017; Accepted 26 September 2018;
Publication 31 October 2018

Abstract

In the face of escalating global Cybersecurity threats, having an automated forewarning system that can find suspicious user profiles is paramount. It can work as a prevention technique for planned attacks or ultimate security breaches. Significant research has been established in attack prevention and detection, but has demonstrated only one or a few different sources with a short list of features. The main goals of this paper are, first, to review the previous user profiling models and analyze them to find their advantages and disadvantages; second, to provide a comprehensive overview of previous research to gather available features and data sources for user profiling; third, based on the deficiencies of the previous models, the paper proposes a new user profiling model that can cover all available sources and related features based on the cybersecurity perspective. The proposed model includes seven profiling criteria for gathering user's information and more than 270 features to parse and generate the security profile of a user.

Keywords: User Profiling, Cybersecurity Profiling, Big Security Data, Security Data Source, Security Profiling Features, Anomaly Detection, Cybersecurity forewarning system.

Journal of Cyber Security and Mobility, Vol. 8_1, 75–112. *River Publishers*
doi: 10.13052/jcsm2245-1439.814

1 Introduction

Based on the technical reports related to cybersecurity, it is clear that cybersecurity professionals are focused on keeping data secure but fail to prioritize the user experience. Researchers estimate that almost half (49 percent) of all security breaches are caused by lack of user compliance. According to this research, emails, external websites, and the Internet activities from workstations are the most challenging end user applications to secure. These three activities cover more than 80 percent of end users' daily work and can cause many attacks such as Denial-of-Service (DOS) attacks, website defacement, access to sensitive information and attacks on critical infrastructure [70].

Generally, among user related attacks, the inside attacks in which the attackers already have access to an organization's network cause significantly more damage than outside attacks. According to the national fraud survey, in the United States alone, internal attacks cost about $400 billion per year and $348 billion can be tied directly to privileged users [71]. Therefore, monitoring and managing privileged user actions is paramount for cybersecurity and compliance reporting. One of the most useful and robust techniques is profiling users and creating a user model to monitor and detect anomalies.

A profile describes the common pattern of a user which contains the user's behavior tendencies and preferences. On the one hand, the knowledge acquired from user profiles provides a strong indication of his/her internal thinking and predict his/her intentions. On the other hand, it is easy to discover users with similar behaviour if their user profiles are similar. Therefore, practically, it is possible to predict a user's behaviour tendencies based on the existing behavioural models.

The content of a user profile reflects different aspects of a user. User security-related behaviours [58] indicate the probability of a user to being able harm or have negative effects on security. Also, various factors have been studied for their association with security [3]. For example, it is interesting to look at the profiles of an end user's sophistication which is a factor for insider threat prediction [35]. Another example is that profiling social interactions in social networks such as Facebook and Twitter can predict user personality [47]. Therefore, it is important to study user security-related profiles, which provide the security analyst with useful information to make security-related decisions.

Our Contribution: Our contribution in this research is threefold. First, we study the current user profiling models based on the data sources, extracted features and profiling techniques to find their advantages and disadvantages. Second, the related works in the user profiling area were reviewed to collect

all features and data sources which have been found by researchers. Third, we select the security-related features and propose a new model for collecting all useful data from different sources and generating an user security profile for anomaly detection.

2 Available User Profiling Models

User profiling, especially from a security perspective, became one of the most important research domains in the last decade. Since 2000, different models have been proposed and implemented in this domain. This section reviews the models and analyzes them to find their advantages and disadvantages.

Pannel [49] proposed and implemented a prototype of an intrusion detection system based on the browser's history files and Windows OS audit logs. At first, different types of user profiles, such as the profile of the website viewed, the profile of the application's performance, and the profile of the applications running, were constructed in the system. Different types of user's anomalous behaviors from unauthorized data can be detected by these profiles. Then, a combined user profile based on the authorized data will be generated by these single-behavior profiles. However, it is hard to conclude that unauthorized data, which cause single-behavior anomaly will definitely be useless for the combined profile. Also, the proposed user profiling system does not use the other user behaviors, e.g. social network activities.

Bradley *et al.* [7] developed a case-based user profiling system for content personalization. The system aims to retrieving a set of jobs that meet a user's job search requirements. The proposed profile captures the user's interests and preferences that come from the user's jobs that visited online visited websites. The profiling method lacks of other data sources for user profiling, such as demographic information that could reflect the user interests.

Sugiyama *et al.* [59] presented an adaptive search system based on user profile. In their system, user profiles are constructed from the user's browsing history. The user profiles are updated whenever there are changes in the user's browsing web pages. The system can return search results that adapt to users with different information needs. But user's browsing history only reflects user interests and preferences from one respective. User profiling from more data sources can be more accurate in inferring a user's information needs.

Grčar *et al.* [19] presented a user profiling system that is implemented as a toolbar in Internet Explorer. They maintain user profiles in the form of interest-focused topic ontology. In the topic ontology, each topic is associated with a set of the user-viewed Web pages. They discover topic-related Web

Table 1 Evaluation of available models

Paper	Data Source	Extracted Features	Profiling Tech.	Advantages	Disadvantages
[59]	Logs	UF1.1	Neighborhood-based approach	User profiling for adaptive search	User profiling of Web search interests based on viewed history webpages
[19]	Logs	UF1.1	Statistical model	User profiling with the support of topic ontology	Profiling user interests based on viewed web page content
[28]	Logs	UF1.1	Collaborative approach	User profiling support user identification and enrichment of neighbors' patterns	Profiling user interests based on clicked, collected, bookmarked content
[13]	Logs	UF1.4	Collaborative filtering	Collaborative filtering for Google new recommendation	Profiling user interests based on clicked items
[7]	Human resource	UF3.15–UF3.17	K-nearest neighbor	Two-stage user profiling: server-side and client-side	Learning user preference based on demographic data
[49]	Log	UF5.2.1.1, UF5.2.1.3, UF5.2.1.10, UF5.2.1.11, UF5.1.2.2	Statistical model	User profiling on system usage and browser history to anomaly detection	Build user profiles for host-based anomaly detection
[11]	Twitter website	UF6.1.1, UF6.1.6–UF6.1.8	Statistical model	User interests profiling for URLs recommendation for Twitter user	User interests profiling based on user interaction in Twitter
[21]	Twitter website	UF6.1.1–UF6.1.5	Collaborative filtering	User profiling for recommending Twitter users to follow	Learn user interests from user activities in Twitter
[47]	Facebook website	UF6.2.1-UF6.2.7	Machine learning	Profiling user interactions in Facebook	Prediction of user personality based on social interaction in Facebook
[8]	Facebook website	UF6.2.8–UF6.2.47	K-means clustering	Profiling user personality traits correlated with buying behavior	User interaction in Facebook needs to be proven in the consumer behavior discipline
[12]	Logs	UF5.2.1.20–UF5.2.1.22	Statistical model	User profiling for anomaly detection	User profiling only based on system logs

pages by analyzing the content of the viewed web pages. Since the generated user profiles group viewed Web pages, the system can quickly guide a user to navigate to one of his interested Web pages. The limitation is that the topic ontology can only include viewed Web pages and cannot recommend

non-visited Web pages. If the system includes more factors to build user profiles, it can include potential-interested topics in the topic ontology.

Das *et al.* [13] proposed a Google News recommendation system. Their system builds and updates each user's profile based on his/her click history of the stories. When a user sends a request of Google news to the system, the system will generate personalized Google news for the user. But their user profile does not consider other data sources that can better describe user interests.

Kim *et al.* [28] proposed a recommender system based on a collaborative user modelling. In the system, they built user profiles not only from user interested items, i.e. the content of movies, but also based on the interested items of the user's neighbors who have interests in common. They believe that feedbacks from a user's neighbors can provide useful information to predict the user's interests. However, their system can consider more information sources to better describe user interests and preference.

Hannon *et al.* [21] presented a Twitter user's recommender system. When a user sends a query to look for Twitter users, the system recommends a list of Twitter users based on the user's profile. They focused on the user tweets and social relation with his followees and followers to build the user profile. The proposed profiling system is only based on user activities on Twitter rather than on multiple social networks.

Chen *et al.* [11] proposed a URL recommendation model for Twitter user to better direct them to their interested content stream. Their model provides recommendations and suggestions based on Twitter user profiles, while their user profiles are based on popular URLs, users' tweets, users' followees' tweets, and social voting within users' neighborhoods. Users' behavior traces outside Twitter, e.g. browser history, can also provide useful information for the Twitter user recommendation.

Ortigosa *et al.* [47] proposed and designed a Facebook application to predict user personality on social networks. Their prediction is based on user profiles on Facebook which describes the user interactions within Facebook, e.g. the number of friends, the number of posts, the number of posts written by friends, etc. They build user profiles on the hypothesis that users with similar personality always exhibit similar behavior patterns. But Facebook is just one source of social network, a user may exhibit a little different behavior pattern in other social networks, e.g., Twitter. Also, other online behaviors, e.g. surfing behavior on the Internet, are good sources for studying user personality.

Baik *et al.* [8] proposed a personality-traits prediction model based on user behaviors on social networks. They investigated user behaviors on Facebook, e.g., personal profile in Facebook, like posts, links, status, tagging the other

users in photos, and etc. User profiles are built by the k-means clustering method. The contribution is that the prediction model based on user profiles is verified to correlate with user buying behavior. However, the model limits the study of user personality traits only based on the activities in Facebook not all commonly used social networks.

Corney *et al.* [12] focused on the computer system logs to build a user profile. They created a user profile to capture user usage pattern of running processes and programs on a computer system. Events that deviate from the user profile can trigger alerts to notify the administrator for further actions. In order to reduce the number of false positive alerts, they use an application grouping technique. Prototype softwares are built and test the anomaly detection system based on user profiles. However, their research is only focused on system logs that are very restrictive to user's behaviors. Normally more information should be considered for the construction of user profiles.

Table 1 shows the results of this evaluation and presents the data sources, features and profiling techniques of each model. In this table to make it easy to read and undertsand, we defined a unique ID for each feature (Tables 3, 4, 5, 6, and 7) and list the extracted features for each research in the "extracted features" column. As the table shows, all proposed model focused on one or maximum two different datasources to define their user profile.

3 Available Features and Data Sources

In this section we review the previous research on three categories, namely: user information, user behavior and network traffic. Subsequently we describe all related features for user profiling. Also, we will identify the primary data sources based on the previous studies.

3.1 Previous Works on Feature Extraction

In this subsection, we review the previous research on three categories, i.e., user information, user behavior and network traffic which have been summarized in Table 2. For each research we list their extracted features on abbreviation format which defined and explained in Tables 3, 4, 5, 6, and 7 (in Subsection 3.2).

3.1.1 User information

In the existing work on user profiling, some research focused on user information including user interests, knowledge and skills, intention & motivation, and demographic information.

Table 2 Summary of profiling criteria (based on the related research)

Profiling Criteria			References
UF1	Users' interests		[2], [13], [19], [28], [29], [39], [43], [55], [59], [62], [66]
UF2	Knowledge, skill		[4]
UF3	Demographic information		[4], [7], [30], [54]
UF4	Intention		[4], [17], [23]
UF5	Behavior	Online behavior	[4], [5], [48], [51], [58], [67]
		Offline behavior	[12], [14], [16], [18], [32], [33], [35], [41], [44], [46], [48], [49], [50], [53], [57], [60], [65], [68], [69]
UF6	Social media activity		[8], [11], [21], [22], [34], [36], [47]
UF7	Network traffic		[1], [6], [9], [10], [15], [25], [26], [27], [38], [40], [42], [45], [56], [60], [63], [64]

User interests:

Li and Yan [43] presented a dynamic model for user profiling. They developed a new approach to describe user profiles as the random sets based on feature UF1.1.

Sugiyama *et al.* [59] created a personalized Web search strategy based on user profiles constructed from Web browser history. The authors believed that user interests, discovered from the browser history, can better guide search engines to filter useful information to satisfy user needs. They used feature UF1.1 in their research.

Grčar *et al.* [19] presented a system to develop and maintain user profiles. The browser history has been used to build a user-interest topic ontology. After extracting terms from the content of web pages, visited web pages have been categorized in accordance with the topic ontology and the extracted terms. Therefore, user profiles consist of web page topics. But only calculating the frequency of visited web pages in terms of user-interest topics cannot be adequate. Features UF1.1 have been used in this research.

Tebri *et al.* [62] proposed a new incremental profile learning approach. The profiling approach worked on the several user-selected documents to learn the user's interests. The experiments showed the effectiveness of the incremental profiling approach based on feature set UF1.1.

Ahmed *et al.* [2] defined a statistical user profiling framework to model user interest. Different from previous work, this research considered a historical user activity as a key factor to build user profiles. The authors believed that a user profile is an online navigational pattern that will change over time. A Bayesian approach has been employed in this research in modelling user's interests by using feature sets UF1.6 and UF1.7.

Au Yeung *et al.* [66] proposed a method to develop user profiles of multiple interests from users' self-defined tags. In a social bookmarking website,

e.g. delicious website, a user can create multiple tags for a bookmarked website. The authors believed that user-defined tags provide a way for a user to describe his/her interested resources. The preferred clustering technique has been used in this research to create user profiles from these tags by using the feature set UF1.2.

Michlmayr [39] created user profiles from tagging data. Co-occurrence of tags was considered rather than occurrence of tags. He believed that a user always likes to use several tags to denote one bookmark in a social bookmarking system. The profiling system used the feature set UF1.8.

Kim *et al.* [28] presented a personal user profile via text mining technique. Multiple data sources were selected from the web content, user's clicked, viewed, and bookmarked to build the user profile. Also a user profile via collaborative characteristics expressed was enriched by the neighborhood users, who exhibited similar interests to the user. Feature set UF1.1 is used in this personal profiling.

Kim *et al.* [29] proposed a collaborative user modelling by leveraging users' rating and social tagging information. The Naive Bayes approach has been chosen to profile a user's interest as a set of tags according to the positive and negative items rated by the user. Feature sets UF1.2 and UF1.3 were chosen in this research.

Das *et al.* [13] introduced a hybrid profiling strategy based on the links that users clicked on the Google news website. The research assumed that users' clicks indeed indicate users' interests by using the feature set UF1.4.

Semeraro *et al.* [55] submitted construct user sense-based profiles. The Word Sense Disambiguation (WSD) procedure has been applied to extract the context words for each word of a user's selected document. A user sense-based profile is represented by these context words for using the feature set UF1.5.

Knowledge and skills:

Murugan [4] targeted developing an AI-based behaviour model to profile employee Web usage. This study shows that surfing nonwork-related websites during work hours not only reduces employees' productivity, but also increases security issues in the workplace. A hybrid method that uses Artificial Neural Networks (ANN) and Genetic Algorithms (GA) was proposed to implement the user profiling system. During training the ANN, GA was used to select the optimal weights for the next iteration. The results indicated that one of the ANN models, Simple Recurrent Network (SRN), is a superior classification method with a high accuracy rate of 89.7% regarding this behaviour profiling problem.

They used feature sets UF2.2.1.1–UF2.2.1.4, UF2.2.2.1–UF2.2.2.3, UF3.1, UF3.2, UF3.7, UF3.8, UF3.10, UF3.13, UF4.2.1, UF4.2.2 and UF5.1.2.1.

Intention and motivation:
Hernández *et al.* [23] aimed to discover user intent behind web queries. Terms extracted from user queries have been suggested for query classification. Besides the frequency of extracted terms, Figueroa [17] proposed to consider linguistic attributes in the query to find user web search intent. Both of them used features UF4.1.1–UF4.1.3 in their queries.

Demographic information:
Krulwich [30] designed an intelligent agent to interact with users on the Internet and recommend web pages according to user profiles. Demographic information has been collected through received survey questions by using feature sets UF3.1–UF3.12.

Silvia *et al.* [54] proposed a hybrid approach that engages case-based reasoning with Bayesian networks for user profiling in an incremental way. The authors believed that representing user preference by user profiles is the learning task for agents to assist users. The proposed hybrid approach consists of two components. In the first component, the case-based reasoning performs an action according to acquired knowledge about users. The action which reflects his/her habits and preferences provides inputs for the second component. In the second component, Bayesian networks models the relationships between items of interest in terms of providing cases. The proposed hybrid technique suits, particularly users' interests that vary over time. Demographic information such as feature sets UF3.10 and UF3.14 have been selected.

Bradley *et al.* [7] explained a two-stage user profiling approach. First, they performed similarity-based user profiling on the server side to filter out useful information. Then, they presented a case-based profiling strategy on the client side to provide personalized service for the user. They argued that client-side profiling can maintain privacy and does not need to submit user-privacy-related data to the server side. The feature sets UF3.15–UF3.17 were used in their proposed approach.

3.1.2 User behavior

User behaviour profiling includes modelling common behaviour, e.g. online behaviour to view web-pages, offline behaviour to leave execution traces on a personal computer, and activities in social networks.

Common behavior:
Denning and Dorothy [14] profiled a user's behaviour on a computer by monitoring the computer's system. Statistical models have been selected to profile a user's normal behaviour. The main purpose for building a user's profile is to detect an abnormality. Their study used feature sets UF5.2.1.1, UF5.2.1.2, UF5.2.1.12, UF5.2.1.18, UF5.2.1.19, UF5.2.4.1, and UF5.2.4.2.

Early in 1997, Lane and Brodley [32] proposed a machine learning approach to anomaly detection. The user profiles constructed based on command sequences and an intruder is supposed to behave quite differently from normal users in terms of command traces. The empirical results demonstrated that the command sequence learning is a promising technique to anomaly detection contingent on using feature set UF5.2.3.1.

Eskin and Lee [16] proposed an entropy modeling method and a probability modeling method to build user profiles based on system calls. Feature set UF5.2.2.1 was used for this research.

Somayaji and Buntwal [57] proposed an abnormal system call detector by profiling the behaviour of user applications (e.g. Netscape on Unix). Two methods were offered to profile system calls, the sequence and the look-ahead pair method. The system-call monitoring system was implemented and showed that detection can be performed efficiently in real time. The authors picked feature set UF5.2.2.1 for their proposed detection system.

Yeung and Ding [65] constructed two types of behavioural models to profile the user normal behaviour using Unix shell commands. The purpose is to detect anomalies from normal user behaviour using the feature set UF5.2.3.1.

Li *et al.* [68] utilized statistical characteristics of N-grams for system calls to profile the normal behavior of a process. The experiments showed a high flexibility and efficiency of anomaly detection by using the feature set UF5.2.2.1.

Pepyne *et al.* [50] targeted very specialized groups of users and profiled user behaviours on the computer. The authors believed that specialized groups of users, e.g. accountants, would use computers in very similar and regular ways due to the nature of their work. User profiles, generated based on three sets of features recorded at each session, are compared with their groups to detect anomalies. Feature sets UF5.2.1.9 and UF5.2.1.12–UF5.2.1.16 were used for proposed profiling.

Stanton *et al.* [58] proposed a taxonomy of end-user security-related behaviour. The level of the technical knowledge and the intentionality of the behaviour have been used to develop the taxonomy. A dataset collected

from 110 individuals was created to prove the effectiveness of the taxonomy. Feature sets UF5.1.1.1–UF5.1.1.3 have been recommended for this taxonomy.

Magklaras and Furnell [35] in 2005 proposed a methodology of leveraging computer system usage and application execution audit to measure end user IT sophistication. The authors used statistical models to build user profiles based on the proposed metrics. For the evaluation, 60 users' audit records and system usage were used to verify the validity of the proposed methodology by using the feature sets UF5.2.1.1, UF5.2.1.3, and UF5.2.1.4.

Baeza *et al.* [5] targeted profiling user behaviour in a query session. The authors believed that user behaviour in query sessions can reveal how users search and use search engines and provide useful information for query recommendation systems. Feature set UF5.1.3.1 was selected for this profiling.

Qiu and Cho [51] learned user search interests from user history-click data. In this research topic-sensitive page rank has been nominated rather than a simple page rank algorithm to profile user's interests. For the evaluation of the proposed profiling, feature set UF5.1.3.2 is the best choice.

Li and Song [33] profiled the Windows NT operating system (OS) users' system behaviours. The Processes of Windows OS has mostly been investigated during a user's login and logout activities. A one-class neural network classifier and Support Vector Machines (SVM) were selected to model user profiles for the purpose of masquerade detection. The authors selected four feature sets UF5.2.1.1, UF5.2.1.10, UF5.2.1.13, 5.2.1.17, UF7.

Ochoa [46] proposed a user profiling method based on process usage of a computer system. The authors believed that users in a department of a company might solve similar tasks and therefore have similar process usage. The users' profiles were generated to classify users into groups based on feature sets UF5.2.1.3–UF5.2.1.9.

Pannell and Ashman [48] proposed a user modelling method based on basic statistics for anomaly detection. The authors selected feature sets UF5.1.2.2, UF5.2.1.1, UF5.2.1.3, UF5.2.1.10, and UF5.2.1.11.

Pannell and Ashman [49] profiled user behaviour based on system usage and visited websites. The user profiles are used for host-based anomaly detection. The results showed that the chosen feature sets UF5.1.2.2, UF5.2.1.1, UF5.2.1.3, UF5.2.1.10, and UF5.2.1.11., can significantly reduce intruder detection time.

Salem and Stolfo [53] proposed a model to profile a user command search behavior in the Windows operating system for masquerade detection. The research is based on the assumption that users are familiar with their own file system while masqueraders are not. Also a Windows command taxonomy has

been designed to classify user commands. The authors build user profiles based on user command search behavior in different categories. The experimental results proved the great performance of the proposed approach with selected features UF5.2.5.1–UF5.2.5.3.

Corney *et al.* [12] targeted the study of the identification of anomalous events in computer system logs. They implemented a prototype software to build user profiles and identify anomalous events with user profiles. User profiles capture user routines of program usage in a computer system. Any event that deviates from user profiles is regarded as suspicious and triggers an alert. Since lots of false positive alerts will be triggered by suspicious events an application grouping technique was introduced to reduce the number of false positive alerts. The results proved that the number of false positive alerts is greatly reduced with feature sets UF5.2.1.20–UF5.2.1.22.

Yu *et al.* [67] built user profiles based on semantics and user's browsing behaviours. A user's browsing sequence of viewing web pages in a session is considered as a meaningful transaction to achieve a navigation goal. A graph-based structure combined with a probability model has been selected to capture the semantics and relations embedded in user browsing sessions. However, the experiment is a simple case study to trace only one user's browsing content with feature UF5.1.2.3.

Fu *et al.* [18] proposed a system log profiling method to detect the execution anomaly of system components in a distributed system. This research converted unstructured text log files (free format) into log keys. Then the Finite State Automata (FSA) technique is applied to log keys to profile normal execution behavior of system components. The profiling method extracted feature set UF5.2.1.23 from a system log.

Nousiainen *et al.* [41] analyzed system log data which are obtained by monitoring servers. The study made a few observations about traffic features. Afterwards, anomaly detection based on Self-Organizing Maps (SOM) was proposed followed by several sample user cases. The feature sets mentioned in this study are UF5.2.1.17 and UF5.2.1.24–UF5.2.1.27.

Li [44] targeted automated log analysis in an effective and efficient way. This study investigated new features as well as different combinations of machine learning algorithms to mine the log data. Finally, a systematic learning process was built with the feature sets UF5.2.1.28–UF5.2.1.34 for the proposed automated system.

Zwietasch [69] tried to include context in the system logs for studying the anomaly detection problem with machine learning technology. The study proposed three feature representation methods to capture the context

information in a log file from different perspective. Then, a position-based anomaly detection algorithm was developed. The final data representation method proposed by the author is a feature vector representation by feature set UF5.2.1.35.

Social network activities:
With the increasing number of netizens, social networks have become extremely popular. Social networks provide a platform for people to interact with others in virtual communities. People could use social networks to make friends, chat, share and exchange information, give ratings and reviews, and make comments.

Hung *et al.* [24] proposed a tag-based user profiling approach. They used tags indicated in Flickr and Delicious to build user profiles. These tags are the user's interests, as well as the interests of the user's social contacts by using feature UF6.1.14.

Hannon *et al.* [21] proposed a Twitter user profiling technique based on the tweets and the relationships between Twitter users, i.e. followers and followees. The objective is to develop a followee recommendation system. They analyzed and evaluated both content-based and collaborative filtering profiling approaches. The evaluation of real-use data suggested that the recommendation system with features UF6.1.1–UF6.1.5. is able to deliver some meaningful followee suggestions.

Chen *et al.* [11] proposed a URL recommendation system based on users' tweets and the tweets from followees. The authors believed that the followees of a user normally have something in common that attracts the user. Hence, the more URL is mentioned by a followees and followees-of-followees, the more likely user is interested in the URL. They used four feature sets UF6.1.1, UF6.1.6, UF6.1.7, UF6.1.8.

Lu *et al.* [34] proposed a re-rank tweets approach in order to recommend the tweets that a user is most interested in. A user's profile consists of the user's interested topics and his/her affinity with other users. The user's interested topics expanded via linking with the knowledge base constructed from Wikipedia by using feature sets UF6.1.9 and UF6.1.10.

Hannon *et al.* [22] presented a multi-faceted user model for Twitter users. The tags associated with Twitter lists have been used to profile the users. The tags represent the core interests of the target user. Also a user's tags contributed to self-interested tags and common interested tags with friends and followers. Feature sets UF6.1.11, UF6.1.12, and UF6.1.13 were chosen for the model.

Ortigosa *et al.* [47] profiled user activities on Facebook to discover user personality traits. This profiling assumed that users with similar personality tend to exhibit common behavioural patterns when interacting with others through social network websites, such as Facebook. Results showed the proposed user profiling approach has a high level of accuracy for predicting user personality with feature sets UF6.2.1–UF6.2.13.

Baik *et al.* [8] profiled user behaviours on Facebook to predict user personality traits. The authors assumed that user buying behaviour can be modelled from the user interaction in Facebook website. The experiments showed the effectiveness of the proposed approach. However, the validity that user behaviour in a Facebook parallel consumer behaviour needs to be proved. Feature sets UF6.2.8–UF6.2.47 were used in this research.

Network traffic:
Wang and Stolfo [64] proposed a payload-based network intrusion detector. Since payload is just byte streams and does not have a fixed length, larger payloads may indicate non-printable or media format data, e.g. pictures, videos, and executable files. The authors made use of the port number and the lengths of bytes in a payload to complete a profile using a clustering method. According to the experimental results, the technique works surprisingly well, it achieves around a 60% detection rate with a false positive rate lower than 0.1% for every port by using feature sets UF7.5.11 and UF7.5.12.

Tabia and Benferhat [60] profiled user's behaviour in using computer and network resources where the built profiles represent normal behaviours. The proposed profiling detected attacks by comparing system activities with normal behaviours with the decision tree algorithm. The network traffic related feature sets UF5.2.1.1, UF5.2.1.2, UF7.1.1.1–UF7.1.1.7, UF7.1.1.10, UF7.1.1.11, UF7.5.11, UF7.5.12, used in the classifier are directly extracted from network packets [52].

Imbert [26] explored network traffic for the identity assurance of user activities. Based on a case study, he developed a user profile from network behavior by analyzing the intercepted network packets. The objective of the user profiling is to assign a confidence level of user activities on the network. The feature sets UF7.3.1–UF7.4.8, UF7.6.1–UF7.6.4 were selected in the user activities model.

Singh *et al.* [56] proposed a network traffic profiling technique on the basis of feature reduction and sample reduction processes. Feature reduction filtered irrelevant features and sample reduction reduced the size of the training set. The experiments were performed on the NSL-KDD dataset [61] and Kyoto

University benchmark dataset [31]. The authors built a network profile with neural network for the purpose of intrusion detection by using feature sets UF7.3.3 and UF7.1.2.1–UF7.1.2.8.

Iglesias and Zseby [25] proposed a multi-stage feature selection method for network traffic based on anomaly detection. The authors analyzed the contribution of every feature for the anomaly detection task. Then, a feature reduction method to reduce the number of network traffic relevant features was proposed for anomaly detection. The original network traffic selected features include UF7.3.3, UF7.3.8–UF7.3.11, UF7.1.2.1–UF7.1.2.10, UF7.1.2.20–UF7.1.2.26, and UF7.5.1–UF7.5.10 for 2 seconds as well as UF7.5.1–UF7.5.10 for the last 100 connections.

Cao *et al.* [9] proposed a new LDA-Based (Latent Dirichlet Allocation) network intrusion detection method. In order to apply the LDA model to network traffic data, network traffic kept in the form of packets with tcpdump are first parsed into documents which are expressed by words. Then, the LDA model built the behavior pattern of normal traffic for intrusion detection. During the parsing phase from tcpdump packets of documents, each packet is represented by a feature vector of 16 feature set, including UF7.2.13–UF7.2.26.

Mantere *et al.* [38] investigated network traffic features for machine learning based anomaly detection in a specific industrial control system network. The research analyzed the feasibility of traffic features presented in [37]. The network traffic has been captured from a living industrial system networks, which is functional in a restricted environment. The authors discussed about the feature sets UF7.2.1–UF7.2.6, UF7.6.4, and UF7.1.2.15 for the proposed detection technique.

Chang *et al.* [10] proposed a two-stage flow-based anomaly detection method to improve the reliability of networks. In the first stage, the normal network traffic profiles are constructed. In the second stage, anomaly is detected by means of entropy-based distance measurement. The experimental result demonstrated its high accuracy and low complexity. This study chose three network traffic feature sets UF7.1.2.11, UF7.1.2.12, and UF7.1.2.13 for anomaly detection.

Lakhina *et al.* [42] studied the distribution of packet features and used entropy as a tool to analyze the feature distribution. Normal traffic profiles are built by the entropy-based clustering method. The experiment on data from two backbone networks validated the high sensitivity of the feature distribution-based method for anomaly detection. The authors used features UF7.1.2.12 and UF7.1.2.13 in their research.

Bereziński *et al.* [6] presented a case study for entropy-based network traffic anomaly detection. The authors studied the performance of several entropy-based anomaly methods on a number of anomalous network traffic traces. The case study proved that combining entropy with a set of selected feature distribution performs better than the traditional just entropy-based method. The researchers calculated the distribution of feature sets UF7.1.2.1, UF7.1.2.3, UF7.1.2.4 and UF7.1.2.13.

Kind *et al.* [27] proposed a feature-based anomaly detection approach on the basis of the construction for histograms of different traffic features. The proposed histogram-based anomaly detection approach modeled histogram patterns and then identified deviations from the constructed models. The experimental results showed the effectiveness of the approach in identifying a wide range of anomalies. The research profiled histogram patterns using features UF7.1.2.1, UF7.1.2.11–UF7.1.2.14, UF7.2.19, and UF7.2.26.

Agarwal and Mittal [1] proposed a hybrid approach that combines the entropy of network features and Support Vector Machines (SVM) for detection of anomalous network traffic. The experimental result demonstrated that the hybrid approach outperforms entropy-based approach and SVM-based approach. This research extracted feature sets UF7.2.7–UF7.2.11 for the proposed approach.

Dokas *et al.* [15] proposed several data mining-based schemes to identify anomalies and detect attacks. Several outlier detection schemes and support vector machines were investigated to model normal network traffic behaviors for detection of attacks. From the content of this research, derived time-based feature sets UF7.1.2.19, UF7.2.13, UF7.5.1, and UF7.5.13 are developed.

Thatte *et al.* [63] developed parametric methods for network anomaly detection. By adopting statistical models, the proposed method achieved real-time estimation of model parameters with background traffic for training. The experimental results demonstrated that the statistical models are able to detect artificial attacks in varying real traffic environments. The research used both feature sets UF7.2.1 and UF7.2.26 to calculate traffic statistics for statistical models.

Lu and Traore [45] proposed an anomaly detection framework for detecting network attacks. The framework includes the feature extraction and outlier detection modules. The feature extraction achieved four-dimensional feature space by some transformation on the features of network packets and flows. The outlier detection worked on the four-dimensional feature space to detect anomalies of network traffic. The transformation process of feature space worked on packet based feature sets UF7.2.14, UF7.2.16, UF7.2.18, UF7.2.24,

UF7.2.27–UF7.2.33 and flow based feature sets UF7.1.2.12, UF7.1.2.16, and UF7.1.2.18.

Münz *et al.* [40] applied the k-means clustering algorithm to the network traffic for anomaly detection from traffic. The algorithm classified network flows in clusters of normal and abnormal traffic. When new network traffic was captured, the algorithm calculated the distance between newly captured traffic flows with the centroid of the cluster for the identification of normal or abnormal traffic. These researchers extracted feature sets UF7.1.2.2, and UF7.1.2.11–UF7.1.2.14 from network flow.

3.2 Available Features

In this section we describe all features that have been extracted from the previous research and categorize them. Tables 3, 4, 5, 6, and 7 list user profiling features (denoted as UF in short) appearing in the literatures in Subsection 3.1.

Interests (UF1): User interests are reflected by user behavior that constantly concentrates on their interests. For example, we could deduce a user's interested topics from his or her browsing history. A user always visiting sport-related websites could indicate her/his interest in sports.

Knowledge and Skills (UF2): Knowledge and skills are two related factors that show how well a user understands the theory and can apply the theory in practice.

- UF2.1 (Knowledge): Users knowledge is obtained by perception and learning. It is considered as a measurement to gauge a user's understanding of a domain, which can be classified into three levels.
  - UF2.1.1 (Breadth of knowledge): This refers to the varieties of knowledge and the fact that a user may have knowledge covering several fields.
  - UF2.1.2 (Depth of knowledge): This refers to how well a user masters the knowledge.
  - UF2.1.3 (Finesse): This refers to the ability to solve a particular problem.
- UF2.2 (Skills): This refers to the ability of a user in a certain domain to solve a problem.
  - UF2.2.1 (Web experience): This refers to how well a user uses the Web as a tool to serve his/her own purpose.
  - UF2.2.2 (Formal training or self-training): It refers to different types of training.

Table 3 User information features (UF1–UF4)

Factor	UF1	UF2	UF3	UF4
Criteria	Interests	Knowledge, Skill	Demographic Information	Intention, Motivation
Sub-criteria	(No sub-criteria)	2.1 Knowledge 2.2 Skill 2.1.1 Breadth of knowledge 2.1.2 Depth of knowledge 2.1.3 Finesse 2.2.1 Web experience 2.2.2 Formal training or self-training	(No sub-criteria)	4.1 Web search intent 4.2 Motivation (No sub-criteria)
Feature	F1.1 Extracted terms from interested content F1.2 Tags associated with bookmarks F1.3 Ratings F1.4 Clicked items F1.5 Context words of each terms extracted from a document F1.6 Topic usages at a certain time F1.7 Historic topic usages at previous epochs F1.8 Combinations of social tags associated with bookmarks	F2.1.1.1 Number of types of IT tools used F2.1.2.1 The level of mastery of a particular IT application or IT knowledge F2.1.2.2 Years of training F2.1.2.3 Years of hands-on experience F2.1.3.1 The efficiency to solve particular IT problems F2.1.3.2 To solve particular IT problems in innovative ways or not F2.2.1.1 Using Internet search engines F2.2.1.2 Downloading files from the Internet F2.2.1.3 Creating Web pages F2.2.1.4 Accessing the Internet F2.2.2.1 In-house company courses F2.2.2.2 Trained by a fellow worker F2.2.2.3 Self-study or self-taught	F3.1 Gender F3.2 Age F3.3 Marital status F3.4 City F3.5 Country F3.6 Number of children F3.7 Education F3.8 Income F3.9 Hobbies F3.10 Career F3.11 Preference F3.12 Zip code F3.13 Size of business F3.14 Department F3.15 Job type F3.16 Salary F3.17 Key skills	F4.1.1 Informational queries F4.1.2 Navigation queries F4.1.3 Transaction queries F4.2.1 Perceived usefulness F4.2.2 Perceived enjoyment

Table 4 User behavior features (UF5)

Factor	UF5	
Criteria	Behavior	
Sub-criteria	5.1 Online behavior 5.1.1 Password-related behavior 5.1.2 Web usage behavior 5.1.3 Web search behavior	5.2 Offline behavior 5.2.1 Computer system behavior 5.2.2 System calls 5.2.3 User commands 5.2.4 User Logins 5.2.5 Command search
Feature	F5.1.1.1 Create weak password F5.1.1.3 Frequency of change password	F5.1.1.2 Sharing password
	F5.1.2.1 Visited non-work related websites F5.1.2.3 Sequences of viewed web pages in a session	F5.1.2.2 visited web pages
	F5.1.3.1 Clicked item in a query session F5.1.3.2 Clicked item in each topic category in a query session	
	F5.2.1.1 CPU usage F5.2.1.2 I/O usage F5.2.1.4 Average instances per process in a certain period F5.2.1.5 Standard deviation of instance in a session F5.2.1.6 Total instances of process from the whole session F5.2.1.7 Average elapse time in a session F5.2.1.8 Standard deviation of elapsed time in a session F5.2.1.10 Number of Windows opened in a session F5.2.1.11 Number of simultaneously running application in a session F5.2.1.12 The time elapsed since the end of the previous session (interval) F5.2.1.13 Number of operating system commands generated during the session F5.2.1.14 The mean of command rate during the session F5.2.1.15 An integer indicating the day of the week when the session began F5.2.1.16 An integer indicating the hour of the day when a session began F5.2.1.17 Number of processes F5.2.1.18 Number of attempts to execute unauthorized programs during a day F5.2.1.19 Number of programs terminates abnormally during a day F5.2.1.20 The hours of the day an application was started F5.2.1.21 The day of the week an application was started F5.2.1.22 Whether or not the application had been run by the user previously F5.2.1.23 Word sequence of log message F5.2.1.24 Number of processes in the kernel run queue F5.2.1.25 Amount of free memory F5.2.1.27 Percentages of CPU time spent on user processes and system processes F5.2.1.28 Total execution time in a log file F5.2.1.29 The variance of time difference between adjacent log messages F5.2.1.30 The maximum time difference between adjacent log messages F5.2.1.31 The average time different between adjacent log messages F5.2.1.32 Number of occurrence of character sequence in a log file F5.2.1.33 Number of occurrence of word sequence in a log file F5.2.1.34 Number of occurrence of a single word in a log file F5.2.1.35 Number occurrence of each event in a log file	F5.2.1.3 Memory usage F5.2.1.9 Total elapsed time for a session F5.2.1.26 Swap space

(Continued)

Table 4 Continued

Factor	UF5	
Criteria	Behavior	
Feature	F5.2.2.1 Sequence of system calls	F5.2.3.1 Unix shell command sequences
	F5.2.4.1 Login frequency	F5.2.4.2 Number of unsuccessful login
	F5.2.4.3 Maximum duration of staying login	F5.2.4.4 Minimum duration of staying login
	F5.2.4.5 Average duration of staying login	
	F5.2.5.1 Number of search-related actions	F5.2.5.2 Number of file accessed
	F5.2.5.3 Percentage of file system navigation user actions during an epoch	

Table 5 User behavior features (UF6)

Factor	UF6	
Criteria	Social Network Activity	
Sub-criteria	6.1 Activity on Twitter	6.2 Activity on Facebook
	6.3 Activity on Youtube	6.4 Other
Feature	F6.1.1 Users tweets	F6.1.2 Tweets of users followees
	F6.1.3 Tweets of users followers	F6.1.4 User ids of users followees
	F6.1.5 User ids of users followers	F6.1.6 Users followees tweets
	F6.1.7 URLs posted by users followees	F6.1.11 Tags associated with users tweets
	F6.1.8 URLs posted by followees of followees	F6.1.9 Extracted concepts from tweets
	F6.1.10 Number of tweets that reply, retweet, or mentioned between a user and a followee	
	F6.1.12 Tags associated with users tweets are of interest to user and his friends	
	F6.1.13 Tags associated with users tweets are of interest to both a user and his followers	
	F6.1.14 Tags associated with user and his social contacts	
	F6.2.1 Number of posts the user has in his wall	F6.2.2 Number of friends
	F6.2.3 Number of different friends that have written in the users wall	
	F6.2.4 Number of posts written in users wall in a day	
	F6.2.5 Number of months since the user started using Facebook	
	F6.2.6 Mean of different friends that have written in the users wall during a certain period	
	F6.2.7 Mean of posts written in the users wall during a certain period	
	F6.2.8 Gender F6.2.9 Age F6.2.10 Blood type	F6.2.11 Relationship status
	F6.2.12 Frequency to be a friend of other users	
	F6.2.13 Frequency being a friend of acquaintances	
	F6.2.14 Frequency of total likes	F6.2.15 Frequency of like photos
	F6.2.16 Frequency of like status	F6.2.17 Frequency of like post
	F6.2.18 Frequency of like link	F6.2.19 Frequency of like check-in
	F6.2.20 Frequency of like own photos	F6.2.21 Frequency of like own status
	F6.2.22 Frequency of like own post	F6.2.23 Frequency of like own link
	F6.2.24 Frequency of like own check-in	F6.2.25 Frequency of like own post on
	F6.2.26 Frequency of tag users in own photo	F6.2.27 Frequency of like page
	F6.2.28 Frequency of tag oneself in own photo	
	F6.2.29 Frequency of being tagged in other users photo	
	F6.2.30 Frequency of joining group	
	F6.2.31 Frequency of photo upload on his feeds with GPS	
	F6.2.32 Frequency of photo upload on his feeds without GPS	

Table 5 Continued

Factor	UF6	
Criteria	Social Network Activity	
Feature	F6.2.33 Frequency of total feeds about status	
	F6.2.34 Frequency of feeds about status with GPS	
	F6.2.35 Frequency of feeds about status without GPS	
	F6.2.36 Frequency of total feeds about check-in	
	F6.2.37 Frequency of feeds about check-in with GPS	
	F6.2.38 Frequency of feeds about check-in without GPS	
	F6.2.39 Frequency of status update	
	F6.2.40 Frequency of privacy setting of feed	
	F6.2.42 Frequency of privacy setting of feed (self)	
	F6.2.43 Frequency of privacy setting of feed (network friends)	
	F6.2.44 Frequency of privacy setting of feed (everyone)	
	F6.2.45 Frequency of privacy setting of feed (all friends)	
	F6.2.46 Frequency of privacy setting of feed (friends of friends)	
	F6.2.47 Frequency of privacy setting of feed (custom)	
	F6.3.1 Number of uploads	F6.3.2 Number of watches
	F6.3.3 Number of channel views	F6.3.4 System join date
	F6.3.5 The time elapsed between the join date and the last login	
	F6.3.6 The interconnection between a user and his neighbors	
	F6.3.7 The probability of mutual subscription	
	F6.3.8 Number of subscriptions made by a user	
	F6.3.9 Number of subscriptions received by a user	

Table 6 Network traffic features (UF7 – part 1)

Factor	UF7	
Criteria	Network Traffic	
Sub-criteria	7.1 Flow based features	7.2 Packet based features
	7.3 Login behavior	7.4 Traffic volume
	7.5 Traffic related to the same host or the same service or the same port	
	7.6 Other statistical features	
	7.1.1 http connections	7.1.2 Other connections
Feature (part 1)	F7.1.1.1 Request length	F7.1.1.2 URI length
	F7.1.1.3 Request method	F7.1.1.4 Type of requested resource
	F7.1.1.5 Number of parameters	F7.1.1.6 Number of arguments
	F7.1.1.7 A request method	F7.1.1.8 Respond code to http request
	F7.1.1.9 Number of requests with same URL	
	F7.1.1.10 Time elapsed since the corresponding http request	
	F7.1.1.11 Number of requests requesting different URLs	
	F7.1.2.1 Duration of the connection	F7.1.2.2 Type of protocol
	F7.1.2.3 Network service on the destination	
	F7.1.2.4 Number of data bytes to destination	
	F7.1.2.5 Number of data bytes to source	F7.1.2.6 Normal or error status
	F7.1.2.7 Number of bad checksum packets in a connection	

(Continued)

Table 6 Continued

Factor	UF7	
Criteria	Network Traffic	
	F7.1.2.8 Number of urgent packets	F7.1.2.9 Connection status
	F7.1.2.10 If source and destination IP addresses and port numbers are equal	
	F7.1.2.11 The source address	F7.1.2.12 The destination address
	F7.1.2.13 The destination port	F7.1.2.14 The source port
	F7.1.2.15 Average duration of flows between endpoints	
	F7.1.2.16 Number of packets	F7.1.2.17 Starting time of flow
	F7.1.2.18 time window for the flow	F7.1.2.19 Number of connections
	F7.1.2.20 Sum of not found error appearances in a connection	
	F7.1.2.21 If the root gets the shell or not	
	F7.1.2.22 If the “su ”command has been used or not	
	F7.1.2.23 Sum of operations performed as root in a connection	
	F7.1.2.24 Sum of file creations in a connection	
	F7.1.2.25 Sum of operations in control files in a connection	
	F7.1.2.26 Sum of outbound commands in a ftp connection	
	F7.2.1 Number of data packets	F7.2.2 IP-port pairs
	F7.2.3 Average size of packets	F7.2.5 TCP session length
Feature	F7.2.4 Average interval between packets	
(part 1)	F7.2.6 Networking protocol (IPv4 or IPv6)	
	F7.2.7 Number of distinct source port in a time slice	
	F7.2.8 Number of distinct destination addresses in a time slice	
	F7.2.9 Number of distinct destination port in a time slice	
	F7.2.10 Number of distinct packet types (ICMP, TCP and UDP) in a time slice	
	F7.2.11 Number of distinct packets with same packet size in a time slice	
	F7.2.12 Number of services in a time slice	
	F7.2.13 Total length	F7.2.14 Type of service
	F7.2.15 Fragment flags	F7.2.16 Time to live
	F7.2.17 IP destination	F7.2.18 TCP flag
	F7.2.19 TCP checksum	F7.2.20 TCP URG pointer
	F7.2.21 TCP option	F7.2.22 UDP checksum
	F7.2.23 Destination port	F7.2.24 ICMP checksum
	F7.2.25 Packet size F7.2.26 Timestamp	F7.2.27 Source IP
	F7.2.28 Source port	F7.2.29 Length of IP header
	F7.2.30 Offset of fragment data	F7.2.31 TCP header length
	F7.2.32 Data location of the TCP segment	
	F7.2.33 Number of data transfered to the destination	

Demographic information (UF3): This feature set contains personal information such as gender, age and marital status. Demographic information provides a useful source to better understand a user’s behavior.

Intention and motivation (UF4): Intention is the purpose for a user’s behavior while motivation is the motive that initiates a user’s behavior.

- UF4.1 (Web search intent): Web search intention is the purpose of a users’ search behavior.

Table 7 Network traffic features (UF7 – part 2)

Factor	UF7	
Criteria	Network Traffic	
Feature (part 2)	F7.3.1 The time of day user access the network	
	F7.3.2 Login frequency	
	F7.3.3 Number of unsuccessful login attempts in a connection	
	F7.3.4 Attempted login in within or outside normal working hours	
	F7.3.5 Attempted login in from other location	
	F7.3.6 Average login duration	
	F7.3.7 The frequency of login at different locations	
	F7.3.8 Successfully logged in or not	
	F7.3.9 Number of logins of normal users	
	F7.3.10 If the user is accessing as root or admin	
	F7.3.11 If the user is accessing as guest, anonymous or visitor	
	F7.4.1 Download bytes per second	F7.4.2 Upload bytes per second
	F7.4.3 Download bytes in a specific time slice	
	F7.4.4 Upload bytes in a specific time slices	
	F7.4.5 The total bytes downloaded per day	
	F7.4.6 The total bytes uploaded per day	
	F7.4.7 The volume of data transferred per day	
	F7.4.8 High volume traffic to a destination IP address over a short time period	
	F7.5.1 Sum of connections to the same destination IP address	
	F7.5.2 Sum of connections to the same destination port number	
	F7.5.3 The percentage of connections that have activated the flag (UF7.1.10) s0, s1, s2 or s3, among the connections aggregated in count (UF7.5.1)	
	F7.5.4 The percentage of connections that have activated the flag (UF7.1.10) s0, s1, s2 or s3, among the connections aggregated in count (UF7.5.2)	
	F7.5.5 The percentage of connections that have activated the flag (UF7.1.10) REJ, among the connections aggregated in count (UF7.5.1)	
	F7.5.6 The percentage of connections that have activated the flag (UF7.1.10) REJ among the connections aggregated in count (UF7.5.2)	
	F7.5.7 The percentage of connections that were to the same service, among the connections aggregated in count (UF7.5.1)	
	F7.5.8 The percentage of connections that were to different services, among the connections aggregated in count (UF7.5.1)	
	F7.5.9 The percentage of connections that were to different destination machines, among the connections aggregated in count (UF7.5.1)	
	F7.5.10 The percentage of connections that were to the same source port, among the connections aggregated in count (UF7.5.1)	
	F7.5.11 Average payload length of inbound traffic for a specific port	
	F7.5.12 Average payload length of outbound traffic for a specific port	
	F7.5.13 Number of different services to the same destination address	
	F7.6.1 Number of packets per day	F7.6.2 Number of service used per day
	F7.6.3 Widely varying network usage over a short time period	
	F7.6.4 Number of connections to different IP addresses daily	

- UF4.2 (Motivation): User behaviors are normally directed and triggered by certain reasons. For instance, a person who is a hockey fan tends to read hockey-related news because he is motivated by his interests in hockey.

User Behavior (UF5): This represents repetitive patterns that a user always do and can be used by an adaptive system or an intelligent agent to assist the user according to the learned behavior.

- UF5.1 (Online behavior): This sub-criteria refers to the activities via the Internet. Online behavior falls into the following three categories.
  - UF5.1.1 (Password-related behavior): This is the behavior related to password and improper password-related behavior that would influence the security of a system.
  - UF5.1.2 (Web usage behavior): This refers to the behavior of accessing Web pages.
  - UF5.1.3 (Web search behavior): Web search behavior refers to search behavior via search engines.
- UF5.2 (Off-line behavior): This refers to the activities on the local computer systems. A user may have different types of off-line behavior:
  - UF5.2.1 (Computer system behavior): Computer system behavior refers to programs or processes running on the computer which causes the consumption of computational resource. The trace of a computer system behavior can be recorded by the system logs.
  - UF5.2.2 (System calls): A system call is an invocation of the operating system services made by an application running on an operating system such as Unix.
  - UF5.2.3 (User commands): This refers to commands that are entered by the users to perform a task.
  - UF5.2.4 (User logins): User login and logout behavior combined with the time factor can reflect users normal login and logout patterns.
  - UF5.2.5 (Command search): Command search behavior targets the search command that a user used on a computer system.

Social network activity (UF6): Social network activity refers to the user's activities on the social networks and includes:

- UF6.1 (Activity on Twitter): This refers to the activities performed by Twitter registered users through Twitter platform.

- UF6.2 (Activity on Facebook): This refers to the activities performed by Facebook users through the Facebook platform.
- UF6.3 (Activity on Youtube): This refers to the activities performed by Youtube registered users through Youtube platform.
- UF6.4 (Other): User maybe has activities on other social network websites, such as Instagram, LinkedIn, and etc.

Network traffic (UF7): This criteria refers to the data traverse through the network during a certain period of time. Network traffic is classified by six criteria.

- UF7.1 (Flow based features): This refers to features that related to a traffic flow, is also called a connection, and consists of a sequence of packets that have the same destination address, the same destination port, the same source address, the same source port, and the same type of service. A TCP connection starts with a packet having a SYN flag and ends with a packet having a FIN flag. In UDP traffic, a threshold of the time slot is used to separate packets into flows.
  - UF7.1.1 (HTTP connections): This refers to the connections related to port 80.
  - UF7.1.2 (Other connections): This refers to other features related to non-HTTP connections.
- UF7.2 (Packet based features): This refers to features extracted from each packet.
- UF7.3 (Login behavior): This refers to login behavior regarding login attempts.
- UF7.4 (Traffic volume): This refers to uploading/downloading behavior.
- UF7.5 (Traffic related to the same host or the same service or the same port): This refers to network traffic related to the same host, the same service, or the same port.
- UF7.6 (Other statistical features): This refers to another statistic about information of network work traffic.

3.3 Available Data Sources

Based on previous studies, there are different types of data for user profiling from diverse sources such as network traffic, Web, human resources, and logs. Various data sources are direct or indirect indications of a users' behaviour. In our proposed model, all available data sources have been selected and compiled. Network traffic as the first source shows the situation of network

traffic when a user is doing activities on organization network, e.g. transferring a file. The abnormal network traffic may imply that the user is under an attack, e.g. DoS attack.

From a users' behaviour on the Web as the second source, we will get to know the visited URLs and the user's social network activity. The visited URLs may disclose a user's interests and intentions. If a user reads sports news very often, a conclusion can be made that this user likes sports. We can even deduce that basketball is his/her favorite sport if this user only focuses on basketball-related news. A user's search intention is expressed in the key words that the user typed in a search engine. If the user is looking for some resource, he will type something like "purchase discount UGG boots". If a user queries a specific bank name such as "RBC", it may indicate that the user does not remember the URL of RBC and just uses the search engine for a navigational purpose.

The third source, social network activities, shows how active a user is on a social network platform, e.g. Facebook, LinkedIn, etc. If a user is sharing too much on social networks, some attacker may steal his/her information for attack purposes. For example, an attacker could make use of a user's information disclosed by the user profile on a social networking site to engineer an attack against a target, e.g. session hijacking and attacks via the malicious link technique [20].

Human resource as the fourth source can provide the demographic information about a user inside an institution. Demographic information is a good supplement to understand a user's online behaviour. For example, it is easy to identify work-related and nonwork-related websites a user visited in terms of his occupation or job title. For a software developer, GitHub website must be a work-related website.

The last source, Logs, is the information obtained by monitoring a computer system. The CPU usage, memory usage, user's working hours, and etc. can be reflected in the log files. The computer's usage pattern also can be captured by analyzing the logs.

4 The Proposed Model

Figure 1 depicts our proposed system for performing user profiling with cybersecurity perspective and analyzes the user's pattern for a forewarning system. The proposed system fetches data from four sources: network traffic, web, human resource, and logs (PC and servers). These are all data sources for usesr profiling summarized from the existing work (as presented in Section 3).

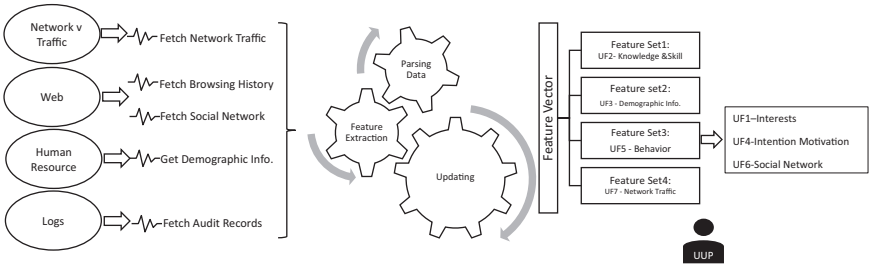


Figure 1 Unified User Profiling (UUP) system with cyber security perspective.

As far as we know, none of the previous researchers focused on cybersecurity perspective to propose a user profiling system based on all data sources; normally the user profiling systems are using only one or two data sources.

In fact, different data sources serve different purposes for user profiling in cybersecurity. The system provides a separate data fetching package for each data source. The first source, fetching network traffic, provides the information about how data traverses through the network when a user performs network activities. Anomaly in network traffic may indicate attacks occur through network traffic; a system flooded with traffic can be a potential signal of DDoS attack.

The second source, fetching web browser history, gathers a user's browsing traces to discover user's online interests and surfing habits. It may help to identify risky users from their browsing history, e.g., how often they visited potential malicious websites or how much information they downloaded from untrusted sources. Fetching social network data gathers users' activities on social network websites. Social network websites, e.g., Facebook and Twitter have become an increasingly popular communication platform. If a user shares too much information on a social network website, attackers may take advantage of this privacy-related information to clone the user's identity for malicious purposes.

The third source, getting demographic information from human resources, helps to better understand user's behavior. For example, if hockey is a user's hobby, the user may be a follower of his favorite hockey star on Twitter.

Finally, the fourth input, system logs, holds audit records which gives the system monitoring data for modeling user normal behavior. It is useful to differentiate a user's abnormal behavior from the user's normal behavior, e.g., an employee who did not work overtime but who logged into the system after working hours or on weekends.

Since data are fetched from different sources, they are obtained in different formats. After the data fetching module, the proposed system needs to parse the data into different feature sets. Parsing data is necessary for the purpose of transforming data from different sources into a uniform format for feature extraction. For example, system logs contain the system event name and event time when the corresponding events occurred. Network traffic contains traffic data composed of elements such as protocol, destination and source IP addresses, timestamp, and flags. Online user behavior includes visited websites, number of visits to a particular web page, and the amount of time spent on each web page. A human resource data source provides a user's personal information such as skills, power, and position in an organization.

After data parsing, a Unified User Profile (UUP) represented by feature vectors is created from parsed data which come from all related sources, including network traffic, web activities, human resources and logs. The system also provides a module of updating data to deal with the dynamic situation of user profiling. On one hand, the dynamic situation is caused by user feature vectors falling seven categories which are extracted in terms of different time windows. For instance, features of short-term user interests are based on a user's one-month visited web pages, recorded in a web browser. The user computer's average CPU usage is counted on a daily basis. On the other hand, the dynamic situation exists due to the continuously increasing number of generated data from data sources.

According to our research, a cybersecurity user profile vector consists of four categories with four feature sets: feature set 1 contains knowledge and skills (UF2), feature set 2 contains demographic information (UF3), feature set 3 contains user online and offline behavior (UF5), and feature set 4 contains network activities (UF7). Feature set 3 is the only one containing subcategories: interests (UF1), intention and motivation (UF4), and social network activities (UF6). Taking advantage of the construction of a user profile vector, the administrator is able to monitor them and find anomalies from any unpredictable user activities.

5 Conclusions

There are many available security tools such as virus scanners and firewalls, but finding the suspicious behaviours and having a forewarning system is one of the main issues for any organization. This paper focuses on the previous research and the current work to lay the groundwork for the new anomalies detection models. The research examines two aspects of the current issue, a

users' security-related profiling criteria and security behaviour features for analyzing the users' behaviour. In the first section, seven profiling criteria have been defined for generating a user's complete profiling. In the second part, more than 270 features have been found for defining the security-related behaviour of the user. Finally, a new model is proposed to collect all security-related data and parsing them in order to create the security profile of a user and monitoring it for giving appropriate feedback to the administrator. The future work of this research is involves an appropriate data capturing and parsing system to extract the features and create a security profile of the user and propose a new forewarning system for anomaly detection.

References

[1] Agarwal, B., and Mittal, N. (2012). Hybrid approach for detection of anomaly network traffic using data mining techniques. *Procedia Technology*, 6, 996–1003.

[2] Ahmed, A., Low, Y., Aly, M., Josifovski, V., and Smola, A. J. (2011). Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 114–122). ACM.

[3] Alfawaz, S., Nelson, K., and Mohannak, K. (2010). Information security culture: a behaviour compliance conceptual framework. In *Proceedings of the Eighth Australasian Conference on Information Security-Volume 105* (pp. 47–55). Australian Computer Society, Inc.

[4] Anandarajan, M. (2002). Profiling Web usage in the workplace: A behavior-based artificial intelligence approach. *Journal of Management Information Systems*, 19(1), 243–266.

[5] Baeza-Yates, R., Hurtado, C., Mendoza, M., and Dupret, G. (2005). Modeling user search behavior. In *Web Congress, 2005. LA-WEB 2005. Third Latin American* (pp. 10). IEEE.

[6] Bereziński, P., Pawelec, J., Małowidzki, M., and Piotrowski, R. (2014). Entropy-based internet traffic anomaly detection: A case study. In *Proceedings of the Ninth International Conference on Dependability and Complex Systems DepCoS-RELCOMEX. Brunów, Poland* (pp. 47–58). Springer, Cham.

[7] Bradley, K., Rafter, R., and Smyth, B. (2000). Case-based user profiling for content personalisation. In *International Conference on Adaptive*

Hypermedia and Adaptive Web-Based Systems (pp. 62–72). Springer, Berlin, Heidelberg.

[8] Baik, J., Lee, K., Lee, S., Kim, Y., and Choi, J. (2016). Predicting personality traits related to consumer behavior using SNS analysis. *New Review of Hypermedia and Multimedia*, 22(3), 189–206.

[9] Cao, X., Chen, B., Li, H., and Fu, Y. (2016). Packet Header Anomaly Detection Using Bayesian Topic Models. *IACR Cryptology ePrint Archive*, 2016, 40.

[10] Chang, S., Qiu, X., Gao, Z., Qi, F., and Liu, K. (2010). A flow-based anomaly detection method using entropy and multiple traffic features. In *Broadband Network and Multimedia Technology (IC-BNMT), 2010 3rd IEEE International Conference on* (pp. 223–227). IEEE.

[11] Chen, J., Nairn, R., Nelson, L., Bernstein, M., and Chi, E. (2010). Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1185–1194). ACM.

[12] Corney, M., Mohay, G., and Clark, A. (2011). Detection of anomalies from user profiles generated from system logs. In *Proceedings of the Ninth Australasian Information Security Conference-Volume 116* (pp. 23–32). Australian Computer Society, Inc.

[13] Das, A. S., Datar, M., Garg, A., and Rajaram, S. (2007). Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web* (pp. 271–280). ACM.

[14] Denning, D. E. (1987). An intrusion-detection model. *IEEE Transactions on software engineering*, 2, 222–232.

[15] Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., and Tan, P. N. (2002). Data mining for network intrusion detection. In *Proc. NSF Workshop on Next Generation Data Mining* (pp. 21–30).

[16] Eskin, E., Lee, W., and Stolfo, S. J. (2001). Modeling system calls for intrusion detection with dynamic window sizes. In *DARPA Information Survivability Conference & Exposition II, 2001. DISCEX'01. Proceedings* (Vol. 1, pp. 165–175). IEEE.

[17] Figueroa, A. (2015). Exploring effective features for recognizing the user intent behind web queries. *Computers in Industry*, 68, 162–169.

[18] Fu, Q., Lou, J. G., Wang, Y., and Li, J. (2009). Execution anomaly detection in distributed systems through unstructured log analysis. In *Ninth IEEE International Conference on Data Mining, 2009. ICDM'09.* (pp. 149–158). IEEE.

[19] Grčar, M., Mladenič, D., and Grobelnik, M. (2005). User profiling for interest-focused browsing history. In *Proceedings of the Workshop on End User Aspects of the Semantic Web* (pp. 99–109).

[20] Hadnagy, C. (2010). *Social engineering: The art of human hacking*. John Wiley & Sons.

[21] Hannon, J., Bennett, M., and Smyth, B. (2010). Recommending twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the fourth ACM conference on Recommender systems* (pp. 199–206). ACM.

[22] Hannon, J., McCarthy, K., O'Mahony, M. P., and Smyth, B. (2012). A multi-faceted user model for twitter. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 303–309). Springer, Berlin, Heidelberg.

[23] Hernández, I., Gupta, P., Rosso, P., and Rocha, M. (2012). A simple model for classifying web queries by user intent. In *Proc. 2nd Spanish Conf. Information Retrieval* (pp. 235–240).

[24] Hung, C. C., Huang, Y. C., Hsu, J. Y. J., and Wu, D. K. C. (2008). Tag-based user profiling for social media recommendation. In *Workshop on Intelligent Techniques for Web Personalization & Recommender Systems at AAAI* (pp. 49–55).

[25] Iglesias, F., and Zseby, T. (2015). Analysis of network traffic features for anomaly detection. *Machine Learning, 101*(1–3), 59–84.

[26] Imbert, Courtney (2014). "Beyond the Cookie: Using Network Traffic Characteristics to Enhance Confidence in User Identity", Available at: *https://www.sans.org/reading-room/whitepapers/authentication/cookie-network-traffic-characteristics-enhance-confidence-user-identity-35362* (accessed October 2016).

[27] Kind, A., Stoecklin, M. P., and Dimitropoulos, X. (2009). Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management,* 6(2), 110–121.

[28] Kim, H. N., Ha, I., Lee, K. S., Jo, G. S., and El-Saddik, A. (2011). Collaborative user modeling for enhanced content filtering in recommender systems. *Decision Support Systems*, 51(4), 772–781.

[29] Kim, H. N., Alkhaldi, A., El Saddik, A., and Jo, G. S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. *Expert Systems with Applications,* 38(7), 8488–8496.

[30] Krulwich, B. (1997). Lifestyle finder: Intelligent user profiling using large-scale demographic data. *AI magazine*, 18(2), 37.

[31] Kyoto University Benchmark Dataset. (2016). Available at: *http://www.takakura.com/Kyoto data/* (accessed October 2016).
[32] Lane, T., and Brodley, C. E. (1997). An application of machine learning to anomaly detection. In *Proceedings of the 20th National Information Systems Security Conference* (Vol. 377, pp. 366–380). Baltimore, USA.
[33] Ling, L., Song, S., and Manikopoulos, C. N. (2006). Windows nt user profiling for masquerader detection. In *Proceedings of the IEEE International Conference on Networking, Sensing and Control, ICNSC'06.* (pp. 386–391). IEEE.
[34] Lu, C., Lam, W., and Zhang, Y. (2012). Twitter user modeling and tweets recommendation based on wikipedia concept graph. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence* (pp. 33–38).
[35] Magklaras, G. B., and Furnell, S. M. (2005). A preliminary model of end user sophistication for insider threat prediction in IT systems. *Computers & Security,* 24(5), 371–380.
[36] Maia, M., Almeida, J., and Almeida, V. (2008, April). Identifying user behavior in online social networks. In *Proceedings of the 1st Workshop on Social Network Systems* (pp. 1–6). ACM.
[37] Mantere, M., Uusitalo, I., Sailio, M., and Noponen, S. (2012). Challenges of machine learning based monitoring for industrial control system networks. In *26th International Conference on Advanced Information Networking and Applications Workshops (WAINA),* (pp. 968–972). IEEE.
[38] Mantere, M., Sailio, M., and Noponen, S. (2013). Network traffic features for anomaly detection in specific industrial control system network. *Future Internet*, 5(4), 460–473.
[39] Michlmayr, E., and Cayzer, S. (2007). Learning user profiles from tagging data and leveraging them for personal (ized) information access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization*.
[40] Münz, G., Li, S., and Carle, G. (2007). Traffic anomaly detection using k-means clustering. In *GI/ITG Workshop MMBnet* (pp. 13–14).
[41] Nousiainen, S., Kilpi, J., Silvonen, P., and Hiirsalmi, M. (2009). *Anomaly detection from server log data*. Technical report.
[42] Lakhina, A., Crovella, M., and Diot, C. (2005). Mining anomalies using traffic feature distributions. In *ACM SIGCOMM Computer Communication Review* (Vol. 35, No. 4, pp. 217–228). ACM.
[43] Li, Y., and Yao, Y. Y. (2002). User profile model: a view from artificial intelligence. In *International Conference on Rough Sets and Current Trends in Computin*g (pp. 493–496). Springer, Berlin, Heidelberg.

[44] Li, W. (2013). Automatic Log Analysis using Machine Learning: Awesome Automatic Log Analysis version 2.0.

[45] Lu, W., and Traore, I. (2005). A new unsupervised anomaly detection framework for detecting network attacks in real-time. In *International Conference on Cryptology and Network Security* (pp. 96–109). Springer, Berlin, Heidelberg.

[46] Ochoa, E. (2007). *User and group profiling based on user process usage* (Master's thesis, Høgskolen i Oslo. Avdeling for ingeniørutdanning).

[47] Ortigosa, A., Carro, R. M., and Quiroga, J. I. (2014). Predicting user personality by mining social interactions in Facebook. *Journal of computer and System Sciences,* 80(1), 57–71.

[48] Pannell, G., and Ashman, H. (2010). User modelling for exclusion and anomaly detection: a behavioural intrusion detection system. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 207–218). Springer, Berlin, Heidelberg.

[49] Pannell, G., and Ashman, H. (2010). Anomaly detection over user profiles for intrusion detection.

[50] Pepyne, D. L., Hu, J., and Gong, W. (2004). User profiling for computer security. In *Proceedings of the American Control Conference,* (Vol. 2, pp. 982–987). IEEE.

[51] Qiu, F., and Cho, J. (2006). Automatic identification of user interest for personalized search. In *Proceedings of the 15th International Conference on World Wide Web* (pp. 727–736). ACM.

[52] Salem, B., and Karim, T. (2008). Classification features for detecting server-side and client-side web attacks. In *IFIP International Information Security Conference* (pp. 729–733). Springer, Boston, MA.

[53] Salem, M. B., and Stolfo, S. J. (2011). Modeling user search behavior for masquerade detection. In *International Workshop on Recent Advances in Intrusion Detection* (pp. 181–200). Springer, Berlin, Heidelberg.

[54] Schiaffino, S. N., and Amandi, A. (2000). User profiling with Case-Based Reasoning and Bayesian Networks. In *IBERAMIA-SBIA 2000 Open Discussion Track* (pp. 12–21).

[55] Semeraro, G., Degemmis, M., Lops, P., and Basile, P. (2007). Combining Learning and Word Sense Disambiguation for Intelligent User Profiling. In *IJCAI* (Vol. 7, pp. 2856–2861).

[56] Singh, R., Kumar, H., and Singla, R. K. (2015). An intrusion detection system using network traffic profiling and online sequential extreme learning machine. *Expert Systems with Applications,* 42(22), 8609–8624.

[57] Somayaji, A. B. (2002). *Operating system stability and security through process homeostasis* (Doctoral dissertation, University of New Mexico).

[58] Stanton, J. M., Stam, K. R., Mastrangelo, P., and Jolton, J. (2005). Analysis of end user security behaviors. *Computers & Security,* 24(2), 124–133.

[59] Sugiyama, K., Hatano, K., and Yoshikawa, M. (2004). Adaptive web search based on user profile constructed without any effort from users. In *Proceedings of the 13th international conference on World Wide Web* (pp. 675–684). ACM.

[60] Tabia, K., and Benferhat, S. (2008). On the use of decision trees as behavioral approaches in intrusion detection. In *2008 Seventh International Conference on Machine Learning and Applications* (pp. 665–670). IEEE.

[61] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009.* (pp. 1–6). IEEE.

[62] Tebri, H., Boughanem, M., and Chrisment, C. (2005). Incremental profile learning based on a reinforcement method. In *Proceedings of the 2005 ACM symposium on Applied computing* (pp. 1096–1101). ACM.

[63] Thatte, G., Mitra, U., and Heidemann, J. (2011). Parametric methods for anomaly detection in aggregate traffic. *IEEE/ACM Transactions on Networking (TON),* 19(2), 512–525.

[64] Wang, K., and Stolfo, S. J. (2004). Anomalous payload-based network intrusion detection. In *International Workshop on Recent Advances in Intrusion Detection* (pp. 203–222). Springer, Berlin, Heidelberg.

[65] Yeung, D. Y., and Ding, Y. (2002). User profiling for intrusion detection using dynamic and static behavioral models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 494–505). Springer, Berlin, Heidelberg.

[66] Yeung, C. M. A., Gibbins, N., and Shadbolt, N. (2009). Multiple interests of users in collaborative tagging systems. In *Weaving Services and People on the World Wide Web* (pp. 255–274). Springer, Berlin, Heidelberg.

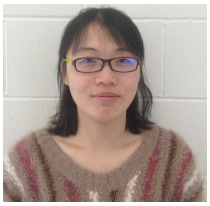
[67] Yu, J., Liu, F. F., and Zhao, H. H. (2012). Building user profile based on concept and relation for web personalized services. In *International Conference on Innovation and Information Management*.

[68] Zhuowei, L., Das, A., and Nandi, S. (2003). Utilizing statistical characteristics of N-grams for intrusion detection. In *Proceedings International Conference on Cyberworlds,* (pp. 486–493). IEEE.

[69] Zwietasch, T. (2014). *Detecting anomalies in system log files using machine learning techniques* (Bachelor's thesis).
[70] "Federal Agency Security Breaches Caused by Lack of User". (2016). Available at: *http://www.businesswire.com* (accessed October 2016).
[71] "Monitoring privileged user actions". (2016). Available at: *https://www.sans.org* (accessed October 2016).

Biographies

Arash Habibi Lashkari is an assistant professor at the Faculty of Computer Science, University of New Brunswick (UNB) and research manager of the Canadian Institute for Cybersecurity (CIC). He has more than 22 years of academic and industry experience developing technology that detects and protects against cyberattacks, malware and the dark web. Dr. Lashkari has been awarded 3 gold medals as well as 12 silver and bronze medals in international computer security competitions around the world. In 2017, he has been selected as the top 150 researchers who will shape the future of Canada. Also, he won the Runner up Cybersecurity Academic Award of the year at ICSIC conference in Canada. He is the author of 10 books in English and Persian on topics including cryptography, network security, and mobile communication as well as over 80 journals and conference papers concerning various aspects of computer security. His research focuses on cybersecurity, big data security analysis, Internet traffic analysis and the detection of malware and cyber-attacks as well as generating cybersecurity datasets.



Min Chen is a postdoctoral fellow at Canadian Institute for Cybersecurity (CIC) on the Faculty of Computer Science, University of New Brunswick. She has extensive academic experience in the areas of machine learning, service computing and cybersecurity. She has several conference and journal publications in the research area of machine learning and service computing. Currently, she is interested in studying user profiling in the respective of cybersecurity with machine learning technology. Her research focused on modeling user behavior as a prevention technique for planned attacks.

Ali A. Ghorbani has held a variety of positions in academia for the past 35 years and is currently the Canada Research Chair (Tier 1) in Cybersecurity, the Dean of the Faculty of Computer Science (since 2008), and the Director of the Canadian Institute for Cybersecurity. He is the co-inventor on 3 awarded patents in the area of Network Security and Web Intelligence and has published over 200 peer-reviewed articles during his career. He has supervised over 160 research associates, postdoctoral fellows, graduate and undergraduate students during his career. His book, Intrusion Detection and Prevention Systems: Concepts and Techniques, was published by Springer in October 2010. In 2007, Dr. Ghorbani received the University of New Brunswick's Research Scholar Award. Since 2010, he has obtained more than $10M to fund 6 large multi-project research initiatives. Dr. Ghorbani has developed a number of

technologies that have been adopted by high-tech companies. He co-founded two startups, Sentrant and EyesOver in 2013 and 2015. Dr. Ghorbani is the co-Editor-In-Chief of Computational Intelligence Journal. He was twice one of the three finalists for the Special Recognition Award at the 2013 and 2016 New Brunswick KIRA award for the knowledge industry.