Validating an insider threat detection system: A real scenario perspective

Ioannis Agrafiotis, Arnau Erola, Jassim Happa, Michael Goldsmith, Sadie Creese
Department of Computer Science, University of Oxford, UK

{*firstname.lastname*}@*cs.ox.ac.uk*

***Abstract*—There exists unequivocal evidence denoting the dire consequences which organisations and governmental institutions face from insider threats. While the in-depth knowledge of the modus operandi that insiders possess provides ground for more sophisticated attacks, organisations are ill-equipped to detect and prevent these from happening. The research community has provided various models and detection systems to address the problem, but the lack of real data due to privacy and ethical issues remains a significant obstacle for validating and designing effective and scalable systems. In this paper, we present the results and our experiences from applying our detection system into a multinational organisation, the approach followed to abide with the ethical and privacy considerations and the lessons learnt on how the validation process refined the system in terms of effectiveness and scalability.**

***Index Terms*—Insider threat; anomaly detection; real world case study; Machine learning**

I. Introduction

The diffusion of information systems into governmental and private organisations has rapidly changed the landscape of threats these institutions face. In the modern digitalised environment, the threat that insiders may pose has grown significantly and several industry surveys denote the dire consequences of such attacks [1, 2]. As the survey conducted by ISACA reports, almost 60% of the attacks which organisations experienced in 2014 originated from a malicious or accidental insider [3]. Thus, the security controls and policies which organisations adopt to prevent external attacks have been proven to be inadequate to deter insider attacks.

Another important characteristic of insider threats is the impact on the businesses, which according to Ponemon's Cyber Crime report in 2014, the mean annualised cost exceeded seven million [2]. To make matters worse, companies are ill-equipped to respond to insider attacks; according to US State Cybercrime survey, only 50% of those who participated in the study had respond strategies for mitigating the impact from such attacks.

In an attempt to address the challenge of insider threats, research community has proposed various systems and models. These span from defining what an insider threat is and conceptualising the problem [4], to understanding the human element and the psychological factors involved [5], and to implementing information systems that detect anomalies indicative of insider threat behaviour [6–8].

While research in the area of insider threat has advanced, the absence of real world data from organisations creates significant difficulties in validating and refining the proposed models. Insider threat incidents, if detected, are under-reported by organisations, while access to data which is crucial for detection systems is scarce due to ethical and privacy issues. As a result, systems are calibrated and tested on synthetic datasets, with the biases it implies.

For the needs of the Corporate Insider Threat Detection (CITD) project, we developed a system to detect anomalies indicative of insider threat behaviour and validated it on synthetic datasets developed in-house [9]. In this paper, we extend this work and investigate the scalability and performance of our system on an anonymised version of real data, in which a known insider attack has been identified by the organisation's head of security. In particular, we focus on the scalability issues we faced and on the enhancement of performance when we attempted to apply our system to data from a multinational technology company. We report the lessons learnt and the challenges on validating our system while abiding to ethical and privacy principals.

In what follows, Section II reflects on proposed models and detection systems for insider threats, Section III summarises the research undertaken as part of the CITD project that led to the development of the detection system, Section IV provides the results of our validation, while Section V reflects on the outcome and provides insights on how to refine the system further. Finally, Section VI elaborates on the lessons learnt from applying the CITD system to real world data, Section VII presents our future plans and Section VIII concludes the paper.

II. Reflecting on Insider Threat research

A. *Conceptualising the insider-threat problem*

The overgrowing implications of the insider-threat problem has attracted the interest of the research community over the last fifteen years. The in-depth knowledge which insiders possess of the security controls, the monitoring practices, and the modus operandi of organisations allows for targeted and sophisticated attacks with dire consequences. Detecting and preventing insider attacks raises significant challenges and insightful research has been focusing on understanding the human element, identifying patterns of attacks and creating conceptual models of insider behaviour [5, 10, 11].

Pioneers in conceptualising the problem of insider threat is the CERT research programme conducted by Carnegie Mellon University [12–15]. Using System Dynamics as a methodological framework, the CERT project examined various real cases of insider attacks which were classified in four broad categories namely Information Technology(IT) Sabotage, Intellectual Property (IP) theft, Data and Financial Fraud, and Espionage. Observations from the case studies were used to identify 'critical paths' which insiders follow and reveal patterns. Specific emphasis was given on understanding the human element and qualitative characteristics, such as disgruntlement or dissatisfaction, which were central to understanding and modelling insider behaviour. This comprehensive work provided valuable insight into understanding the nature of insider threat and led in a series of MERIT (Management and Education of the Risk of Insider Threat) studies.

Complementary works followed different perspectives for distilling insider attacks. Sarkar [16] built on the CERT models and considered factors who may act as precursors of an attack (i.e. psychology of an attacker, reduced loyalty, historical behaviour). In addition, insiders are categorised in those who act maliciously and those who unintentionally facilitate an attack, either due to negligence, or due to violation of policies and procedures to facilitate their daily tasks. In a similar vein, Pfleeger *et al.* created a taxonomy for capturing the actions of the insiders by eliciting elements from the organisation, the environment and the system which insiders exploited [17]. Finally, the U.S. Department of Homeland Security discussed how behavioural aspects and personality traits may provide indications of insider threats and reflect on prevention measurements [18].

B. Detecting insider threats

A different strand of literature has focused on designing and implementing systems to detect and prevent insider attacks. Despite of the in-depth knowledge of an organisation, insiders may leave digital fingerprints to be captured when an attack is executed [19]. In addition, behavioural characteristics and precursor elements may be mapped to changes in the digital profile of an employee. It is possible then to raise alerts indicative of an insider threat when these changes are detected.

One of the most common methods used to capture changes in employees' digital footprint is anomaly detection. Digital activities of employees can form sequences of actions, which overtime may build a profile used as a baseline for detection. Observed sequences of actions which deviate from the normal profile may be regarded as potentially anomalous behaviour. Parveen and Thuraisingham [20] presented such an approach and introduced the notion of *concept-drift* to indicate employee's behavioural change over time. Their approach uses unsupervised learning techniques and processes vast volumes of streaming data using stream and graph mining [21].

Other research articles have focused on how to extract features indicative of insider threat based on data which organisations collect and provided different algorithms for detecting suspicious activity [22–24]. A noteworthy example is presented in [25] where the authors obtained access to organisations' databases and explored data from employees' laptops. Their validation approach consider as normal the data obtained, and inserted logs which corresponded to malicious activity providing them the ground truth. They then tested the effectiveness of seventeen different detection algorithms. A visual language to illustrate features was provided as well. Over 100 features were captured, however, there is little discussion on how these were identified.

Several cases have been reported were insiders colluded to execute an attack. Chen and Malin [26] to identify suspicious collaborative behaviour by creating a *user-relationship network* using unsupervised learning techniques. Their model is tested on access logs from an electronic health record system in a medical centre. Other proposed detection systems endeavoured to infer psychological and behavioural factors from users' activity and incorporate these features into the anomaly detection algorithms [27, 28]. Brdiczka *et al.* [7], in particular, refined their initial detection system with the addition of psychological profiling to reduce the number of false-positive alerts. Another interesting approach is using Bayesian networks to infer the behavioural attributes of users based on sentiment analysis on text and social network analysis [29].

Finally, Bishop *et al.* [30] investigated how process modelling may shed light into detecting insider activity. By formalising processes, they explored which tasks are managed by which agents and analysed the various ways which processes may be compromised, suggesting counter measures to increase the resilience of processes to insider attacks.

The design of the proposed systems is driven by the different types of data available to the research community. Most of the published work is validated on synthetic data and on the scarce occasions where real data from organisations is available, malicious activity is inserted. Especially datasets which are based on real data with the insertion of malicious activity, such as DARPA ADAMS, provide valuable insight and a great opportunity to the research community to test and improve detection systems. In our work, we were fortunate to test our system on real world data without the need to insert synthetic malicious activity because the organisation had a well-known case of insider. Real data exhibits patterns in people's behaviour that is difficult to simulate and since it is a rare opportunity to experiment with only real data, this paper endeavours to fill this gap.

III. The Corporate Insider Threat Detection project

The CITD project is sponsored by the UK Centre for the Protection of National Infrastructure (CPNI). The project brings together two groups from the University of Oxford, the University of Leicester and Cardiff University. The aim of the project is to develop an investigative tool to detect potential insider threats while adhering to ethical and privacy considerations.

Our core approach and proposed framework for modelling the insider-threat problem goes beyond traditional technological observations and incorporates a more complete view of insider threats, common precursors, and human actions and behaviours [31]. The conceptual model proposed for insider threats provided a reasoning structure to analysts that can make or draw hypotheses regarding a potential insider threat based on measurements from real-world observations.

Due to the scarce data available on insider attacks, the University of Leicester conducted several interviews with insiders who were convicted for their malicious behaviour and security experts from organisations which suffered from insider attacks. More than 120 cases were identified and examined. The analysis of these case studies provided a framework to fully characterise insider attacks. This included a clear definition of why employees are motivated to attack, who is more prone to attack, the human factors that lead to malicious and accidental threats and how individual's background may affect the likelihood of an attack being realised. We also identified what the common attack vectors and steps within an attack are, and what assets and vulnerabilities are typically targeted [32, 33].

We further deconstructed all 120 cases to identify unique attack steps, which describe atomic activities that took place during these attacks. We then created a chain of attack steps that culminate in the end-goal of the attacker (e.g., committing fraud), designed attack-pattern graphs comprising the attack steps, and highlighted the most prevalent paths for every attack type [34].

The culmination of our efforts was the design of an automated detection system described in detail in [9]. The architecture of the system is illustrated in Figure 1. The system parses various types of log activity (file logs, web logs, email logs, login/log-off logs etc) and builds tree profiles for every employee and every role. Based on these profiles, features indicative of insider threat are extracted. These features are processed based on a semi-supervised approach. Principal Component Analysis (PCA) is used to reduce the dimensionality and cluster similar behaviour. We use as a baseline a certain amount of days and the detection of anomalous behaviour is performed on a daily basis. The values of features are plotted in the reduced dimensional space and if considered abnormal alerts are raised.

We created a three-tier alert system. The first level of alerts consists of policy violations and tripwires based on well-known attacks. The second level of alerts is threshold-based and assesses the Euclidean distances of the plotted features. The third level are deviation-based anomalies running statistical assessments on the values of the Euclidean distances. Human analysts can engage to provide an active learning feedback loop. By adopting an accept-or-reject scheme, the analyst is able to refine the underlying detection model to reduce the false-positive rate.

In [9] we validated the system on synthetic scenarios developed in-house. Here, we reflect on the results from this process, refine the system and further apply it on real dataset.

IV. Applying the system to data from a multinational organisation

Validating the automated part of the CITD detection system does not only provide the means to assess the effectiveness of the tool, but also the opportunity to gain insight into how it could be further optimised to identify anomalies indicative of insider threats. We applied the CITD detection system to data obtained from a multinational corporation. The organisation was aware of an insider acting during a specific period of time and we assumed that all other activities by employees were not malicious.

The most important challenges we faced when applying our system to real data were the ethical and privacy implications. In addition to the data protection act, the organisation had strict policies in place on how employee's data may be processed. From the university side, there are ethics rules that every researcher has to abide by. In order to comply with these policies, we decided to focus only on anonymised data and more importantly never to store the data on a machine outside the organisation's firewall. Access to the anonymised data was only available to the head of security of the organisation.

The sensitivity of the data required us to adopt a bastion approach to deploy our detection system. We provided the organisation with a pre-configured machine with the tool installed, which was based at their headquarters. We also configured a different machine to which we did have access to and used this to upload our code for debugging purposes. The organisation would then synchronise the updated versions of the system to the machine placed at headquarters, try to run the system and report back with results. More specifically, an SSH public key was sent to a gateway machine on the organisation's DMZ. The code never ran from that particular machine and the raw data logs were never accessible to this machine. Figure 2 shows the CITD system configuration in the multinational organisation.

Regarding the way the organisation reported results, we did not obtain a detailed list of the alerts that the system generated. The organisation provided us only statistical information regarding the number and severity of alerts (low or high deviation), the number of employees we generated alerts for and the type of alerts (i.e. suspicious file activity) along with the dates for when these alerts were generated for the malicious insider only.

It is important to note that the purpose of our testing was not to identify whether other employees acted maliciously but to test whether the system was able to detect a known insider in historical data; at no point was any other employee under consideration of being an insider nor under scrutiny from the organisation; more importantly because the alerts raised by our system are indicative of anomalous behaviour but not necessarily of suspicious behaviour only. Anomaly detection is merely one element of a multi-pronged approach towards detecting malicious behaviour with complementary elements from data on employee behaviour and off-line activities such as travelling and performance reviews providing additional insight. The person who acted maliciously during the period of data we processed, had admitted their actions and had been prosecuted by the organisation. No personal details about this person nor the nature of their attack were revealed to us.

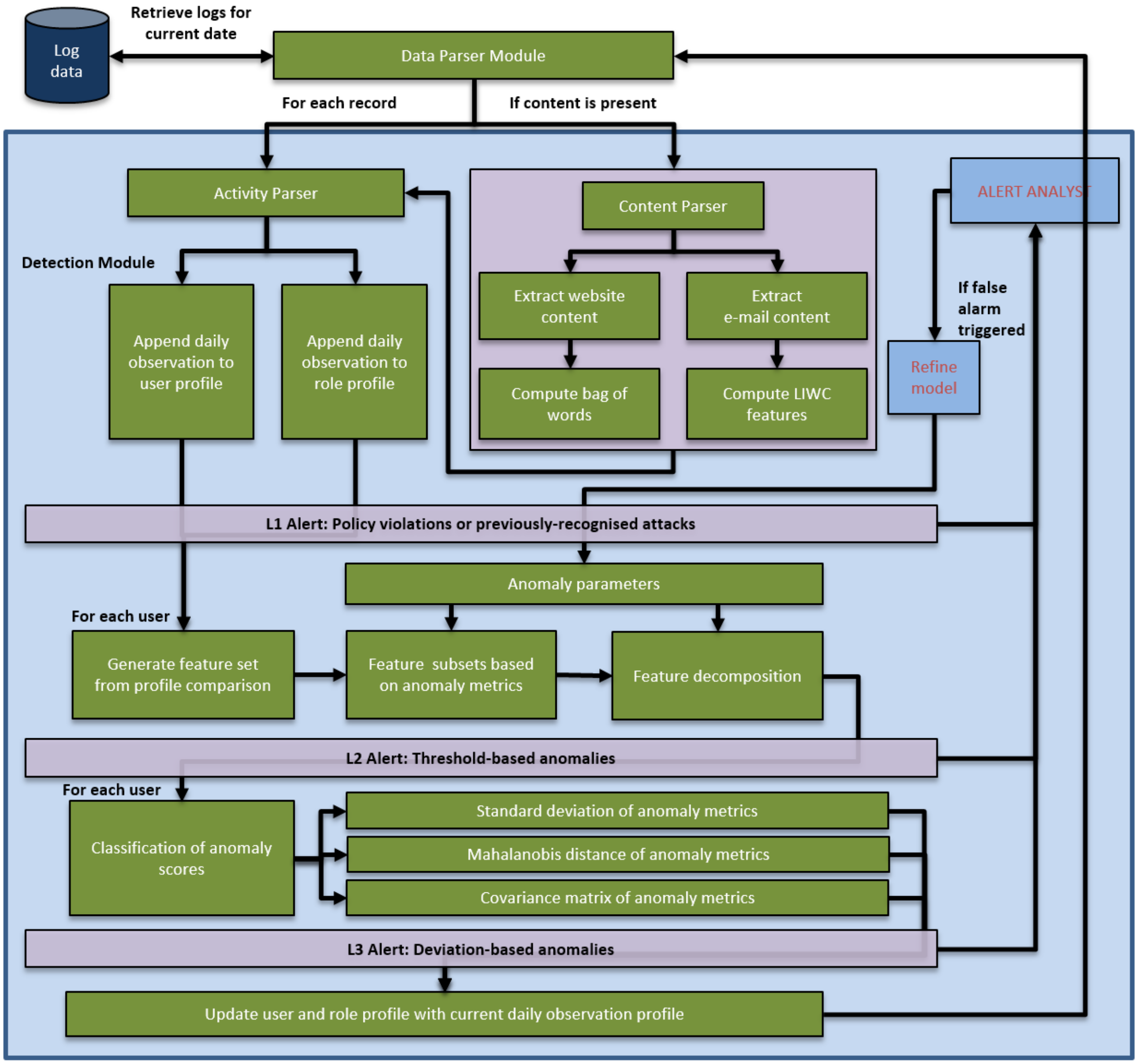
Log data
Retrieve logs for current date
Data Parser Module
For each record
If content is present
Activity Parser
Content Parser
ALERT ANALYST
Detection Module
Append daily observation to user profile
Append daily observation to role profile
Extract website content
Extract e-mail content
Compute bag of words
Compute LIWC features
If false alarm triggered
Refine model
L1 Alert: Policy violations or previously-recognised attacks
Anomaly parameters
For each user
Generate feature set from profile comparison
Feature subsets based on anomaly metrics
Feature decomposition
L2 Alert: Threshold-based anomalies
For each user
Classification of anomaly scores
Standard deviation of anomaly metrics
Mahalanobis distance of anomaly metrics
Covariance matrix of anomaly metrics
L3 Alert: Deviation-based anomalies
Update user and role profile with current daily observation profile

Fig. 1. Architecture of the insider threat detection system described in detail in [9]. The system is constructed by a number of parsers and components that interpret log data records to create user and role tree profiles. There exist three levels of alerts: policy violations and tripwires based on well known attacks, threshold-based anomalies and deviation-based anomalies. Alerts are visualised to analysts who may provide feedback on the validity of the alert and reconfigure the sensitivity of the system.

A. *Deciding on relevant datasets*

The organisation records different types of log activities. The format of the logs differs significantly, requiring several parsers to process the data. There was also not a straightforward way to link every activity to a user. Some logs did not include a username as a field and linking this type of activity to a specific individual required further analysis and examination of other type of logs.

We had several meetings with the head of security to understand the format of the data collected, as well as which insider attacks are prevalent in such environment. The most common and serious threat for this organisation is IP theft. The organisation stores the most sensitive documents in an internal database and only authenticated users are allowed to access these files. To the best of our knowledge, they do not have an automated detection system to identify suspicious downloads of files by employees nor a team to manually examine which employees exhibit suspicious behaviour. We aimed to test the system on multiple data sources, but in the end we limited our testing to only data produced by file-access logs, which were of most interest to the organisation. We designed the system, however, to be able to parse and process all the different data sources.

We decided to consider data regarding:

- File-access logs
- Patent DB interactions

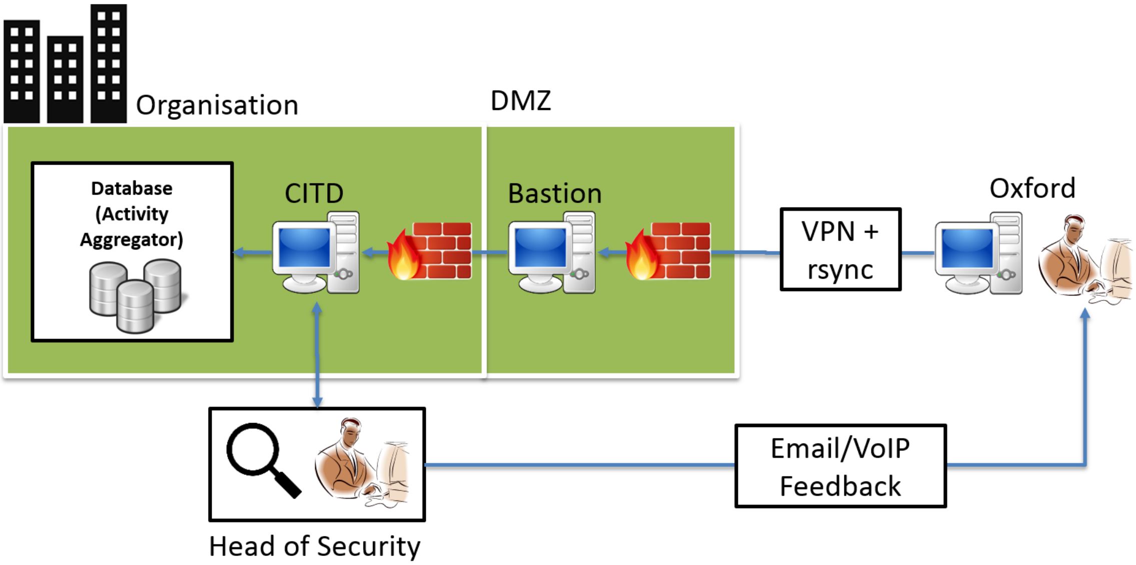

Fig. 2. The organisation has sensors placed collecting network-based and host-based activities (meta-data). These datasets were aggregated to a single point. Prior to our visit to the organisation we obtained data type descriptions and wrote relevant parsers to convert their data types to a CSV format. All CSV files relate to activities only. How the data is aggregated and monitored is determined by the organisation.

- Directory DB interactions

B. Scalability issues

In [9], the synthetic scenario which contained the biggest number of users and comprised the most data, included 20,000 data entries per day and 300 employees who were spread across five different roles. In this case however, the file access logs provided more than 750,000 data entries per day. These logs were not chronologically ordered and did not contain information about the role of each employee. Additional processing was required to link employees to roles. It is worth noting that the roles were anonymised (R1, R2 etc) even though we never got access to this file.

The overwhelming amount of data and the additional processing to link employees to activities and roles, resulted in hardware limitations, which could have been surpassed by deploying the system in a more powerful machine. We decided, however, to focus on data which derived from authenticated users. Unauthenticated file logs resulted only in denial of sensitive files. According to the head of security authenticated logs provide the most interesting information regarding insider threats. Parsing only authenticated data entries resulted in a significant reduction in the amount of data from 750,000 to 44,000 entries per day. Another advantage of authenticated logs lies in the straightforward link between a specific action and a user, meaning that no further processing of the logs was required.

Authenticated file-access logs provided information about the user, the date and time when the activity was conducted, the type of activity, the size of the file and the standard response code from the server. The information acquired from the logs differed significantly to the information which the file system logs of the synthetic scenarios in [9] provided. These differences rendered some of the anomaly metrics used for the synthetic scenarios irrelevant and new anomalies pertinent to the organisation were defined.

The file access logs comprise five different types of activities namely X1, X2, X3, X4, X5 [1]. We decided to focus on these activities and monitor deviations in employees' requests. Attributes attached to these activities were the size of the file and the standard server response code. Therefore, we defined five anomaly metrics for the file system logs which share the same attributes (size of the file and standard http responses). These are:

- X1_anomaly
- X2_anomaly
- X3_anomaly
- X4_anomaly
- X5_anomaly

Alerts on X1_anomaly, X2_anomaly, X3_anomaly and X4_anomaly would suggest that the user downloaded an unusual number of files, whereas an X5_anomaly alert would denote that a user created or modified many files (resulting in a possible integrity issue, or sabotage efforts).

C. Results

We applied the CITD detection system to file access logs over two different periods (1 September to 31 October and 1 December to 31 December). The overall number of employees was 16,000 and although we were able to modify the code to cater for the roles which employees possess, due to the sensitivity of data we never got access to the file, thus the

[1] Due to non-disclosure agreements the format of the logs, and the names of the anomalies which were indicative of the type of the log are anonymised in this paper.

number of roles is unknown. The organisation had already identified suspicious activity and the insider had acted mainly on late September and October. Unsurprisingly, in the runs of the system that considered data from December 2014 period we did not generate an alert for this insider. We did, however, generate several alerts on our last run which focused on data from September 2014 and October 2014 as predicted.

In more detail, the first run of the system was on data from December 2014 and consisted of a training period of seven days. Determining what activity should constitute the normal profile is a challenging task. In the case of our data, there may well be malicious activity in these seven days. Our validation piece was focused only on detecting a well-known insider. Alternatively, there should be a more sanitised version of data to provide the grounds for the development of normal profile.

We used PCA to reduce the features on two dimensions and the system generated $\sim$4,000 alerts for $\sim$3,000 employees. Of these alerts, $\sim$1000 were of Red severity, indicating deviations greater than four units. In order to reduce the number of alerts, we introduced new features tailored to the needs of the organisation based on the types of threats they believe they find. The new features were the number of unique new files for every activity, the number of unique current files for any activity and the overall size of files a user accessed for every activity. We calculated these features for every role as well.

We added a subset of the new features to the system to ensure that its performance would be acceptable (the complexity of PCA algorithm is $O(n*m^3)$ where n is the number of days we process the data for and m is the number of features. It is clear that introducing new features increases the computation time substantially). Our decision on which features to add was based on the experiences that the head of security had regarding previous incidents (i.e X5 logs were not considered at all and the number of X1 logs was crucial). We tested the system again on the same data and for the same training period, while increasing the dimensions of PCA to three. We managed to reduce the number of alerts from $\sim$4000 to $\sim$2700 (33% reduction). This time however, we generated only 297 alerts of Red severity (70% reduction) for 235 users, which is a significant improvement.

The final run of the system considered data from September and October 2014, period where we knew an insider acted maliciously. We generated 129,797 alerts (of which 42,420 were of Red significance) for 4,129 individuals. Concerning the individual who the organisation suspected of insider activity, we generated several alerts of Orange significance (deviations greater than three but less than four units) pertaining to this person. What is of interest about this individual, however, is not the significance of the alerts but the fact that he/she raised an alert for every anomaly the system is calculating, resulting in twelve alerts in a single day. Bearing in mind that on average we produced for the period of two months $\sim$0.5 alerts per employee per day ($a/e*d = 129,797/4,129*60 \sim 0.5$ where a is the number of alerts the system produced, e is the number of employees the system produced an alert for and d is the number of days we have data for), getting 12 alerts in a single day is a significant indication. There was no other employee who generated so many different type of alerts for such a short period of time and the ratio of alerts per user over the period of 60 days followed a normal distribution with the average alerts per day for user's being 0.5 and the units of standard deviation being 1.5, meaning that 95% of employees for whom we generated an alert for had between 0 and 3.5 alerts per day.

More specifically, for the **29-9-2014** we generated alerts regarding:

- X1_anomaly
- X2_anomaly
- X3_anomaly
- X4_anomaly
- this_anomaly
- any_anomaly
- new_anomaly
- current_anomaly
- hourly_anomaly
- user_anomaly
- role_anomaly
- total_anomaly

The individual had increased activity on accessing files (X1, X2, X3, X4 anomalies). In addition, the individual had significant variations from his normal profile (user_anomaly), accessed files on different hours than usual (hour_anomaly) while he requested an unusual number of new files (new_anomaly), as well as an unusual number of files he had visited in the past (this_anomaly and current_anomaly). Finally, his overall behaviour deviated from the behaviour that people holding his/her role exhibited.

For the **30-9-2014** we generated alerts regarding:

- hourly_anomaly

On this day the individual accessed files on unusual hours.

For the **20-10-2014** we generated alerts regarding:

- X1_anomaly
- hourly_anomaly

The individual requested an unusual number of files (X1_anomaly) and accessed these files on unusual hours.

For the **21-10-2014** we generated alerts regarding:

- hourly_anomaly

On this day the individual accessed files on unusual hours.

For the **21-10-2014** we generated alerts regarding:

- X1_anomaly
- X2_anomaly
- this_anomaly
- hourly_anomaly

The individual requested an unusual number of files (X1_anomaly, X2_anomaly, this_anomaly) and accessed these files on unusual hours.

While the number of employees we generated alerts for concerns 25% of the staff (as a multinational organisation, people's working hours tend to change significantly), further statistical analysis on the generated alerts, reduces significantly the number of false positives.

We are in the process of further refining these results, by focusing on the anomalies which are more indicative of IP theft (i.e. X1_anomaly) and the interpretation of the alerts produced for individuals per day (i.e. how many alerts per day is a significant indication of suspicious behaviour, are there alerts which appear only when insiders act (in this case for example the insider generated always an hour_anomaly). In addition, we are introducing more features to reduce the noise (i.e. number of rejected requests for documents) and the number of false-positive alerts.

V. Reflecting on the results to refine the system

Applying the CITD system to a real-world dataset inevitably generated false-positive alerts (alerts for behaviours that do not signify an insider threat) and false negatives (behaviours that should have generated alerts but remained unidentified). In our case study, however, the ground truth becomes complicated. We can be certain for the minimum number of false negative alerts which the system generated, however, the number of false-positive alerts remains questionable. An alert could either indicate a false-positive or an attack that has not been detected thus far by the organisation.

Our system provides a variety of configuration options to cater for cases where the results are not representative of the ground truth. Our next steps on the validation methodology will focus on decreasing the number of false-positive and false-negative alerts by examining the source of their origin. The generation of an increasing number of false-positive alerts has its origins in the erroneous perception of what the norm is. There may be cases where the behaviour of the employees varies daily and their activities differ significantly (i.e. research and development departments where the working hours, number of emails sent, number of files accessed change radically over a period of time). In such cases a more dynamic representation of the norm is suggested.

We therefore intend to explore the use of dynamic profiles. Instead of aggregating all the benign data to capture the normal profile of a user, we will consider as normal profile the last thirty days of the users' data. For every following day, we would delete the oldest data entry of the normal profile and insert the newest observation provided it did not constitute an anomaly. Furthermore, there might be cases where the norm may appear as suspicious behaviour. For example, the number of emails that a sales person will send and receive is far greater compared to the number of emails a software developer is likely to receive. In these scenarios the system could be configured to be less sensitive in what constitutes an anomaly.

Regarding false negatives, their source of origin is the undetected attacks. The behaviour of the insiders may be very subtle or hidden under legitimate usage of the system. In such scenarios, the system can be reconfigured to become more sensitive to changes in behaviour. In case where subtle attacks were executed in the past, policy rules may be designed to indicate such situations. For example, an insider whose end-goal is to download a sensitive file, may start downloading various files to conceal the real attack, thus the normal profile will not defer significantly from the characteristics of the attack. We are in the process of creating trip-wired rules based on the policies of the organisation to capture these more subtle attacks.

A. Anomaly metric performance assessment

In all applications of the system, we assigned the same value to all the weights of each anomaly, assuming no prior knowledge of different types of attacks the organisation experienced. Once we have a better understanding of which anomaly metric is error-prone we may wish to reduce its influence in the generation of the second-tier and third-tier alerts. Similarly, since we have evidence that for the organisation theft of sensitive files is a prominent type of attack, we will increase the value of the weight for the file_anomaly.

The aforementioned suggestions consider changes after the PCA algorithm is applied. We intend to make changes to render the PCA algorithm more efficient and improve its outcome, which is the eigenvector and the eigenvalues. By default, when applying PCA we include the data of the day under investigation to the normal profile. We will, however, apply PCA only to the normal profile and then project the data of the day under investigation to the new eigenvalues and eigenvectors.

In addition, we will try to run PCA with more dimensions to increase the variability of our data. By default the system considers only the first two dimensions of the eigenvectors which reduce dramatically the number of variables to consider, thus rendering the system faster, and allow for visualisation tools to project the data in a two dimensional space. Reducing the dimensions, however, results in loss of information. Increasing the default number of the final dimensions may provide better results in cases where there are many features to consider.

VI. Lessons learnt

The process of validation provided useful lessons not only for the CITD project but for the broader research community. Below we describe how organisational development and operational lessons could be broadly applied to other projects and inform decisions for the overall system architecture and considerations before validating these tools. We also present lessons regarding algorithmic inputs and alert outputs which will inform the future work for the CITD project.

A. Organisational development

In terms of broader organisational development, a critical task for the successful deployment of detection tools is to establish procedures that will adhere to security policies and privacy rules that organisations have in place. We have demonstrated that it is feasible to validate a system without ever viewing the anonymised version of the input data. It is also important for the organisations to understand that the alerts the system may produce are not evidence of insider activity and may provide indications of deviations which may well be benign. An issue that may occur is that sometimes organisations may have a legal obligation to follow any alerts detection and monitoring systems provide, even during trial periods. In such cases, organisations may be reluctant to provide full access to datasets or report detailed results of the system.

Interviewing system administrators and employees responsible for securing the operations of the organisation is an important part of the deployment and validation process. It allows to shed light into what type of attacks are prevalent, what data types are relevant in identifying these attacks and which metrics tools could rely on to capture suspicious activity. We benefited greatly from our interactions with the head of security and managd to reduce the number of false positives alerts.

B. Operational issues

In terms of operational issues, deploying the CITD detection tool to data from a multinational organisation allowed us to identify which parts of the system are applicable to different contexts without alterations and which parts require modifications. We expect that deployemnt of other detection tools would lead to similar conclusions. It is evident for example that the format of the data to be used as input is a significant factor in system's adaptability. Environments which record logs with similar format would require a minimum number of changes. In real-world datasets, however, the format of the data will be very different, requiring appropriate parsers to interpret it. Due to the different format of the data, the anomaly metrics identified in [9] may not be always applicable and novel anomalies deriving from the logs may need to be defined.

Writing parsers to tailor data for input is a straightforward process. Understanding the data, however, and highlighting interesting information which could give rise to suspicious behaviour is a demanding and time-consuming task. We believe this would be a challenge for every detection system. Other challenges relevant to detection systems which create user profiles include identifying activities which belong to the same user and linking different datasets to construct a normal profile. We also need to acknowledge the limitations in the availability of datasets due to legal restrains and monitoring practices, which limit the efficiency of our system (email content, albeit rich in information, is extremely difficult to get access to due to ethical and legal considerations).

Focusing on the CITD system, once the data parsers are in place, activities and their attributes are well defined, and the anomaly metrics are set, the core functionality of the system does not require any modifications to process this information and output alerts. The tree structure profile is able to process any list of activities and attributes, irrespective of their name, and the anomalies which depend on role and time characteristics are agnostic to the anomaly metrics tailored for specific contexts (i.e. role_anomalies and hourly_anomalies are computed in any context without further modification needed). We encountered, however, scalability issues due to the amount of information stored in these profiles. In order to render the system scalable for the magnitude of data an organisation produces, we are considering filtering the data stored as activities and attributes in the system. This solution may limit the system's ability to extract certain features, however, storing every single file a user has visited may not always provide useful results. Instead, storing sensitive files which organisations strive to protect may be a more efficient way to move forward.

C. Algorithmic lessons

Regarding algorithmic lessons, PCA and the rational behind triggering alerts in all three tiers is applicable in any context and requires only subtle modifications. These modifications pertain to the maximum number of dimensions in which PCA is executed. The number of dimensions depends on the features generated for the anomaly metrics and since these are context-dependent the highest possible dimensions are calculated by $d-1$ where d denotes the number of activities in each context. It may also be useful to create alerts that will pertain to certain organisations, allowing users to define which features should be selected.

D. Output of alerts

Concerning the output of alerts, we believe that the first-tier alerts can be tailored to capture policies for specific contexts, thus identifying subtle attacks. Second-tier alerts capture suspicious behaviour by distinguishing anomalies when a user's behaviour compared to the normal profile differs more than a certain threshold. Weights associated with anomalies allow us to focus on specific behaviours and tune the system if needed. Third-tier alerts are the most effective according to our results. These alerts are generated when a user's behaviour deviates from the normal behaviour more than a specified number of units. Our findings suggest that alerts which indicate deviation more than 2 standard units provide the most decisive indicators of suspicious behaviour and statistical analysis on the alerts per user may reduce the number of false positives and provide useful insights.

VII. Future work

The validation strategy focused on the anomaly detection algorithms and explored the effectiveness of the third tier alert. As insiders obtain better knowledge of the detection systems in place and endeavour to mask their abnormal behaviours, it is getting more difficult to compute effective anomaly scores. Identifying attack patterns can be extremely helpful, as these patterns can provide the mechanism for combining different features and anomaly metrics used by detection systems, and set the context for understanding anomalous behaviour.

We have captured different insider attacks in the form of attack trees which provide a solid framework for identifying subtle attacks, cases where employees collude to orchestrate an attack and solid ground for reasoning regarding the employees' motivations [34]. Incorporating these attack patterns into the system in the form of policy alerts would be one avenue for future research. It would allow the system to cater for the aforementioned attacks and reason about the behavioural aspects of people based on their online activity.

Therefore, we intend to focus on formally describing and designing policies for the first level alerts. These policies would have a two-fold purpose; to capture violations of policies designed by Human Resources (HR); and to describe behaviours that may be indicative of problematic behaviour even if this behaviour does not constitute a violation of a HR policy, based on the attacks trees identified in.

Due to the lack of data indicative of the psychological profile and personality trait of employees, our system does not consider behavioural indicators for detecting concerning behaviour. If such data were available, these would form another feature dimension and would feed the anomaly metrics. Unfortunately, recording accurately data for stress indicators or organisational commitment is an extremely difficult task and further research is needed particular in exploring how the use of language, login times, or variation in keystrokes can be linked to personality traits.

At the moment the results deriving from PCA are processed by standard deviation and Mahalanobis distances to identify significant changes in employees' profiles. More methodologies and machine learning techniques could be adopted. A useful exercise would be to explore the results of these techniques when applied to different contexts and correlate the effectiveness of these techniques with specific insider attacks. In addition, we could focus on designing Bayesian networks based on the attack-pattern trees and link these with anomaly metrics.

Another avenue for future work could be the development of a dashboard visualisation to provide a better understanding of the alerts the system is raising. As described in Section III, the alerts regarding the insider were of average severity but the number of these over a short period of time was significantly different to the distribution of alerts for other users. The number of alerts on a daily basis flagged by detection systems must be manageable by security analysts in organisations. Dealing with an overwhelming number of alerts will be counter-productive in the long-term for the detection of insider threats, especially in the case where most of these are false positive alerts. Visual representation of alerts though provides opportunities to better realise patterns in these alerts and insights into the nature of the alerts. We have started testing visualisation techniques and the results are promising.

VIII. Conclusions

While insider threat is becoming an increasing concern within industry, the difficulties researchers are experiencing in gaining access to real datasets to validate their systems is slowing down the implementation of more effective methods to deter insider attacks. This work reflects on the results, our thoughts and our experiences in applying our CITD system to a real multinational organisation.

Our approach abided to the organisational ethical and privacy considerations. During our experiments we had at no time access to the datasets and the organisation was responsible to run the system. All the detection output was filtered by the head of security and only statistical information about the alerts was provided as well as feedback about our system's computational and detection performance. We used this information to continually update our approach and improve the system's accuracy.

Our tool processed historical file logs of $16,000$ users and extracted features to create normal baselines of behaviour. The organisation had already identified malicious activity from one of the employees and our validation method focused on identifying alerts for this individual. From the explored detection methods, dimensionality reduction using PCA combined with anomaly detection using standard deviation provided very good results. Although the system generated false-positive alerts, which could be related to a non-identified attack or to a real false identification, further statistical analysis of the results provided significant indications of the insider's malicious activity.

The opportunity to apply the system on real datasets revealed scalability limitations concerning the vast amount of data which needs to be processed and identified issues on linking certain activities to users. We provided strategies for refining the system and highlighted the value of having a data curator both for data protection and system performance feedback purposes.

In the future, we intend to deploy our tool in other organisations, incorporate attack patterns as part of our behavioural detection and establish means to develop policy tripwires (signatures), not strictly intended as means to indicate maliciousness of a policy breaking, but as means to indicate where problematic behaviours exists. In turn, this may inform policy-makers where shifts in policies should change due to operational issues.

Validating the CITD system in real data provided the grounds to assess the effectiveness of the system. Our results suggest that the PCA rational and the three tier approach can capture a wide range of alerts with relative success and generate a small number of false-positive alerts. In addition, projecting data in higher dimensions decreases the number of false-positive alerts. We are confident that building the normal profile for users and roles which employees hold, and calculating deviations based on PCA results is the first and decisive step towards designing a system for detecting anomalous behaviour indicative of insider threat.

Acknowledgements

This research was conducted in the context of a collaborative project on Corporate Insider Threat Detection, sponsored by the UK National Cyber Security Programme in conjunction with the Centre for the Protection of National Infrastructure, whose support is gratefully acknowledged. The project brings together three departments of the University of Oxford, the University of Leicester and Cardiff University. We would like to acknowledge Dr Phil Legg for his contribution to the detection system code base in preparation for deployment of the tool in said multinational company. Finally, we would also like to thank the multinational company for their continued collaboration on this project.

References

[1] Ponemon Institute and Attachmate Corporation. (2013) The risk of insider fraud second annual study: Executive summary. [Online]. Available: http://www.attachmate.com/resources/analyst-papers/bridge-ponemon-insider-fraud-survey.htm

[2] H. P. E. U. Kingdom, "Ponemon cyber crime report: It, computer and internet security," 2015. [Online]. Available: http://www8.hp.com/uk/en/software-solutions/ponemon-cyber-security-report/

[3] I. . RSAConference, "State of cybersecurity: Implications for 2015," 2015. [Online]. Available: http://www.isaca.org/cyber/Documents/State-of-Cybersecurity_Res_Eng_0415.pdf

[4] J. Hunker and C. W. Probst, "Insiders and insider threats – an overview of definitions and mitigation techniques," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 2, no. 1, pp. 4–27, 2011.

[5] E. D. Shaw and H. V. Stock, "Behavioral risk indicators of malicious insider theft of intellectual property: Misreading the writing on the wall," Symantec, Tech. Rep., 2011.

[6] F. L. Greitzer and R. E. Hohimer, "Modeling human behavior to anticipate insider attacks," *Journal of Strategic Security*, vol. 4, no. 2, pp. 25–48, 2011.

[7] O. Brdiczka, J. Liu, B. Price, J. Shen, A. Patil, R. Chow, E. Bart, and N. Ducheneaut, "Proactive insider threat detection through graph learning and psychological context," in *Proc. of the IEEE Symposium on Security and Privacy Workshops (SPW'12), San Francisco, California, USA*. IEEE, May 2012, pp. 142–149.

[8] P. A. Legg, N. Moffat, J. R. C. Nurse, J. Happa, I. Agrafiotis, M. Goldsmith, and S. Creese, "Towards a conceptual model and reasoning structure for insider threat detection," *Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications*, vol. 4, no. 4, pp. 20–37, 2013.

[9] P. A. Legg, O. Buckley, M. Goldsmith, and S. Creese, "Automated insider threat detection system using user and role-based profile assessment," 2015.

[10] D. M. Cappelli, A. P. Moore, and R. F. Trzeciak, *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes*, 1st ed. Addison-Wesley Professional, 2012.

[11] C. Colwill, "Human factors in information security: The insider threat who can you trust these days?" *Information Security Technical Report*, vol. 14, no. 4, pp. 186–196, 2009.

[12] F. L. Greitzer, A. P. Moore, D. M. Cappelli, D. H. Andrews, L. A. Carroll, and T. D. Hull, "Combating the insider cyber threat," *Security Privacy, IEEE*, vol. 6, no. 1, pp. 61–64, 2007.

[13] A. P. Moore, D. M. Cappelli, and R. F. Trzeciak, "The "big picture" of insider it sabotage across U.S. critical infrastructures," Tech. Rep., 2008. [Online]. Available: http://www.sei.cmu.edu/library/abstracts/reports/08tr009.cfm

[14] A. P. Moore, D. M. Cappelli, T. C. Caron, E. Shaw, D. Spooner, and R. Trzeciak, "A preliminary model of insider theft of intellectual property," Tech. Rep., 2011. [Online]. Available: http://www.sei.cmu.edu/library/abstracts/reports/11tn013.cfm

[15] D. M. Cappelli, A. P. Moore, and R. F. Trzeciak, *The CERT Guide to Insider Threats: How to Prevent, Detect, and Respond to Information Technology Crimes*, 1st ed. Addison-Wesley Professional, 2012.

[16] K. R. Sarkar, "Assessing insider threats to information security using technical, behavioural and organisational measures," *Information Security Technical report*, vol. 15, no. 3, pp. 112–133, 2010.

[17] S. L. Pfleeger, J. B. Predd, J. Hunker, and C. Bulford, "Insiders behaving badly: Addressing bad actors and their actions," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 1, pp. 169–179, 2010.

[18] N. Cybersecurity and C. I. Center, "Combating the insider threat," 2014. [Online]. Available: https://www.us-cert.gov/sites/default/files/publications/Combating%20the%20Insider%20Threat_0.pdf

[19] M. A. Maloof and G. D. Stephens, "Elicit: A system for detecting insiders who violate need-to-know," in *Recent Advances in Intrusion Detection*, ser. Lecture Notes in Computer Science, C. Kruegel, R. Lippmann, and A. Clark, Eds. Springer Berlin Heidelberg, 2007, vol. 4637, pp. 146–166.

[20] P. Parveen and B. Thuraisingham, "Unsupervised incremental sequence learning for insider threat detection," in *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*, June 2012, pp. 141–143.

[21] P. Parveen, J. Evans, B. Thuraisingham, K. Hamlen, and L. Khan, "Insider threat detection using stream mining and graph mining," in *Privacy, security, risk and trust (PASSAT), 2011 IEEE Third International conference on social computing*, Oct 2011, pp. 1102–1110.

[22] H. Eldardiry, E. Bart, J. Liu, J. Hanley, B. Price, and O. Brdiczka, "Multi-domain information fusion for insider threat detection," in *Security and Privacy Workshops (SPW), 2013 IEEE*, 2013.

[23] N. Nguyen and P. Reiher, "Detecting insider threats by monitoring system call activity," in *Proceedings of the 2003 IEEE Workshop on Information Assurance*, 2003.

[24] J. Myers, M. R. Grimaila, and R. F. Mills, "Towards insider threat detection using web server logs," in *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*, ser. CSIIRW '09. New York, NY, USA: ACM, 2009, pp. 54:1–54:4. [Online]. Available: http://doi.acm.org/10.1145/1558607.1558670

[25] T. E. Senator, H. G. Goldberg, A. Memory, W. T. Young, B. Rees, R. Pierce, D. Huang, M. Reardon, D. A. Bader, E. Chow *et al.*, "Detecting insider threats in a real corporate database of computer usage activity," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1393–1401.

[26] Y. Chen and B. Malin, "Detection of anomalous insiders in collaborative environments via relational analysis of access logs," in *Proceedings of the first ACM conference on Data and application security and privacy*. ACM, 2011, pp. 63–74.

[27] G. B. Magklaras and S. M. Furnell, "Insider threat prediction tool: Evaluating the probability of IT misuse," *Computers and Security*, vol. 21, no. 1, pp. 62–73, 2002.

[28] F. L. Greitzer and R. E. Hohimer, "Modeling human behavior to anticipate insider attacks," *Journal of Strategic Security*, vol. 4, no. 2, p. 3, 2011.

[29] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.

[30] M. Bishop, B. Simidchieva, H. Conboy, H. Phan, L. Osterwell, L. Clarke, G. Avrunin, and S. Peisert, "Insider threat detection by process analysis," in *IEEE Security and Privacy Workshops (SPW)*. IEEE, 2014.

[31] P. A. Legg, N. Moffat, J. R. C. Nurse, J. Happa, I. Agrafiotis, M. Goldsmith, and S. Creese, "Towards a conceptual model and reasoning structure for insider threat detection," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 4, no. 4, pp. 20–37, 2013.

[32] J. R. Nurse, O. Buckley, P. Legg, M. Goldsmith, S. Creese, G. R. Wright, M. Whitty *et al.*, "Understanding insider threat: A framework for characterising attacks," in *Security and Privacy Workshops (SPW), 2014 IEEE*. IEEE, 2014, pp. 214–228.

[33] J. R. Nurse, P. A. Legg, O. Buckley, I. Agrafiotis, G. Wright, M. Whitty, D. Upton, M. Goldsmith, and S. Creese, "A critical reflection on the threat from human insiders–its nature, industry perceptions, and detection approaches," in *Human Aspects of Information Security, Privacy, and Trust*. Springer, 2014, pp. 270–281.

[34] I. Agrafiotis, J. R. Nurse, O. Buckley, P. Legg, S. Creese, and M. Goldsmith, "Identifying attack patterns for insider threat detection," *Computer Fraud & Security*, vol. 2015, no. 7, pp. 9–17, 2015.