Gheyas and Abdallah *Big Data Analytics* (2016) 1:6
DOI 10.1186/s41044-016-0006-0

Big Data Analytics

REVIEW Open Access

Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis

Iffat A. Gheyas* and Ali E. Abdallah

* Correspondence:
Iffat.Gheyas@bcu.ac.uk
School of Computing and Digital Technology, Birmingham City University, City Centre Campus, Millennium Point, Birmingham B4 7XG, United Kingdom

Abstract

Cyber security is vital to the success of today's digital economy. The major security threats are coming from within, as opposed to outside forces. Insider threat detection and prediction are important mitigation techniques. This study addresses the following research questions: 1) what are the research trends in insider threat detection and prediction nowadays? 2) What are the challenges associated with insider threat detection and prediction? 3) What are the best-to-date insider threat detection and prediction algorithms? We conduct a systematic review of 37 articles published in peer-reviewed journals, conference proceedings and edited books for the period of 1950–2015 to address the first two questions. Our survey suggests that game theoretic approach (GTA) is a popular source of insider threat data; the insiders' online activities are the most widely used features in insider threat detection and prediction; most of the papers use single point estimates of threat likelihood; and graph algorithms are the most widely used tools for detecting and predicting insider threats. The key challenges facing the insider threat detection and prediction system include unbounded patterns, uneven time lags between activities, data nonstationarity, individuality, collusion attacks, high false alarm rates, class imbalance problem, undetected insider attacks, uncertainty, and the large number of free parameters in the model. To identify the best-to-date insider threat detection and prediction algorithms, our meta-analysis study excludes theoretical papers proposing conceptual algorithms from the 37 selected papers resulting in the selection of 13 papers. We rank the insider threat detection and prediction algorithms presented in the 13 selected papers based on the theoretical merits and the transparency of information. To determine the significance of rank sums, we perform "the Friedman two-way analysis of variance by ranks" test and "multiple comparisons between groups or conditions" tests.

Keywords: Insider threat prediction, Anomaly detection, Machine learning, Cyber security, Individual attacks, Collusion attacks

Background

We live in the digital age and like anything, this new reality has its upsides and downsides. Its major downside is the security risk. As more and more of our sensitive information is moving to the digital world, confidentiality breaches are becoming more common and significant. "HIV patient tells of fears of disclosure after details leak" [1], "Barclays bank leaks thousands of customer records" [2], "Pepsi alerted Coca-Cola to stolen-coke-secrets offer" [3], "PlayStation Network users fear identity theft after major

data leak" [4] —such news headlines are all too familiar. Most confidentiality breaches occur from within the company [5].

Data integrity is another important security concern. Damage to data integrity can often cause more serious problems than confidentiality breaches. For example, our lives may hang in the balance if the attacker alters and manipulates our medical records like blood type and drug allergies. The most damaging data integrity attacks would be conducted at the country's critical infrastructure systems (such as its water supply or electrical system) through the deliberate injection of incorrect data to the SCADA system. Hence real-time attack detection and prediction are dominant topics for IT security. These are the first layer of defence. Attack detection means being able to detect the presence of an attack as early as possible. Prediction means deriving the likelihood of future attacks from current data.

Security threats can come from inside or outside of an organization. The attacks from insiders, be they from employees, suppliers, or other companies legitimately connected to a company's computer system, pose a more pernicious threat than external attacks. These insiders have knowledge of the internal workings of the organization, and full possession of all the rights and privileges required to mount an attack that outsiders lack. Consequently, insiders can make their attacks look like normal operations.

Companies intend to spend more against insider attacks over the coming years [6]. However, all efforts to protect against insider attacks may go to waste if it is not accessible when and where it is needed. The challenge here is to develop automated threat detection systems that do not generate too many false alarms. A false security alarm may result in short-term or prolonged loss of availability; a loss of availability could result in employees not being able to have access to the system and do their jobs effectively during a time-sensitive emergency, when every moment counts. A loss of system availability can paralyse a company. This can lead to higher costs, lost revenue and reputational damage.

Availability, confidentiality and integrity are fundamental aspects of the protection of systems and information. Loss of any one of these items constitutes a security breach. In order to optimize these conflicting requirements we need to develop an insider threat detection and prediction algorithm (IDPA) that minimizes both false negatives and false positives. Different forms of IDPAs have been developed in recent years. Since many algorithms have been proposed for this task, it is natural to ask which the top performers are. To our knowledge, no such screening has been performed before in the literature. We did however find three review papers published recently focusing on insider threat research: (i) Hong, Kim, & Cho [7] provide an overview of IDPAs; (ii) Ophoff et al. [8] categorize the existing insider threat research into five categories; and (iii) Azaria et al. [9] critically review several IDPAs. However, none of these studies performed a systematic review and meta-analysis of published studies to compare the IDPAs. The aim of our study is to systematically review the available literature and perform a meta-analysis to compare and rank the existing IDPAs. This study addresses the following research questions:

I. What is the trend of study in insider threat detection and prediction (IDP)?
II. What are the challenges in IDP?
III. Which algorithms are the best to use in practice?

The remainder of this paper is organized as follows. Section 2 presents the methods used in this study. Section 3 presents the results of the systematic review and meta-analysis. We conclude and discuss our key findings in section 4. Finally, 'best practice' recommendations for the development, implementation and evaluation of IDPAs are presented in section 5.

Methods

PRISMA [10] guidelines were implemented to standardise the features of this systematic literature review.

Literature search strategy and selection criteria

The literature was searched through Web of Science (WoS) by topic from 1950 to 2015 with the key words of "insider threat", "insider threat detection", and "insider threat prediction". We also searched Google Scholar (GS) database using "insider threat detection and prediction" as search terms. We screened the first 100 search records within GS. Additionally, we searched the reference lists of all potentially relevant articles and book chapters. We removed duplicate papers. The abstracts of the identified articles were scrutinised for relevance based on the following questions—if the answers to all of these questions are 'Yes', then we shortlisted the article for further consideration.

I. Is the article concerned about cyber threats from insiders?
II. Is the focus of the paper on the threat detection and/or prediction?
III. Is the article focused on malicious insider actions, as opposed to unintentional or careless insider actions?
IV. Is it a journal article/conference article/ book chapter?
V. Is the article written in English?

After the abstract screening process, shortlisted papers were read in full to ensure that the main focus of the article was on the data-driven approach and also to ensure that earlier versions of the same article were removed.

Meta-analysis: data extraction, interpretation and bias management

This study performs the meta-analysis to compare and rank the existing IDPAs. It is not possible to compare the algorithms against the reported performance in the literature because the majority of the time an algorithm is evaluated only once. Different algorithms are evaluated using different evaluation criteria and different kinds of data. Hence we compare IDPAs based on their theoretical features. Using a systematic literature review we identify the challenges associated with IDP. We then rank IDPAs based on how well their theoretical constructs address those challenges. Publication bias is a common problem, as most papers with poor results either aren't submitted for publication or are rejected at peer review. Hence it is important to take into account publication bias. The best judge of publication bias will be how transparent the authors are about their proposed algorithms. Transparency is a measure of how easily other researchers would be able to follow and evaluate the proposed algorithms. To examine

the publication bias of a study, we consider a set of questions about the public availability of the experimental datasets and the pseudo codes of the proposed algorithms.

Statistical analysis

IDPA's are ranked based on two conditions: (i) the score for the theoretical soundness of the algorithm, and (ii) the score for publication bias. We apply "the Friedman two-way analysis of variance by ranks" ([11], pp. 172–180) to the rankings of IDPAs according to the above two conditions in order to test the null hypothesis that there are no statistically significant differences in performance of the IDPA's. After applying the Friedman two-way analysis of variance by ranks test and noting that it is significant, we use one-tailed "comparisons of groups or conditions with a control" ([11], pp.181–183) tests to perform pairwise comparisons between IDPAs in order to test the null hypothesis that there is no significant difference between the two algorithms. We choose the significance level 0.05 for hypothesis tests.

Results

We organize this section as follows. The results of the study selection process are reported in section 3.1. Section 3.2 highlights the trends for the research in IDP. Section 3.3 discusses the challenges of IDP. Section 3.4 presents the meta-analysis results: a comparative analysis of IDPAs.

Study selection

A simple GS article search yielded 20,500 references. We selected the first 100 GS records for consideration. WoS search yielded 438 papers. One additional paper was identified from the references of other published papers. These 539 references were checked for duplicates. We screened 177 abstracts, evaluated 50 full-text articles, and included 37 articles—a total of 22 studies (59 %) propose novel IDPAs [9, 12–32]. The other 15 papers either propose new features for IDP or discusses challenges associated with IDP [33–47]. Figure 1 presents the flow chart of the study selection process.

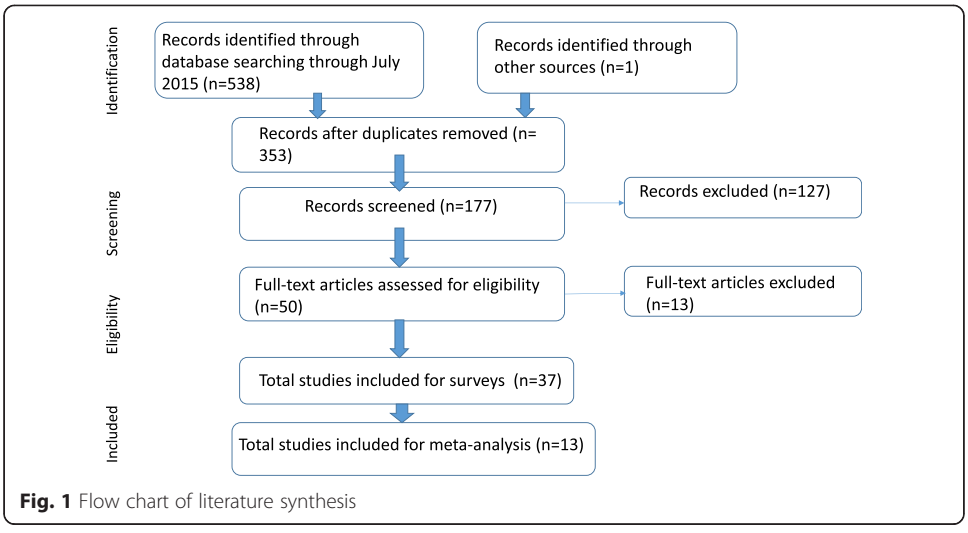

Fig. 1 Flow chart of literature synthesis

In 13 papers (out of these 22 papers presenting novel algorithms), the authors have implemented and evaluated the proposed algorithms. We included these 13 studies in the meta-analysis to find the best IDPAs. We examined all 37 selected articles to address the first two research questions mentioned in section one: (i) the trend of research in IDP, and (ii) the challenges in IDP.

Trends of the research in IDP

We selected 37 papers for review [9, 12–47]. Each of these papers deals with one or more dimensions of the data-driven modelling framework of insider threat research. Figure 2 depicts the distribution of the papers over time.

The figure shows a clear upward trend from the year 2000 in the number of IDP publications, reflecting increased risk for insider threats. The number of publications peaks in 2009 (which coincides with the year that an army psychiatrist—after receiving a poor performance evaluation— opened fire at Fort Hood, Texas killing 12 people and wounding 31 others), and in 2013 (which coincides with the year that Edward Snowden, a former contractor for the CIA, leaked the Snowden Surveillance Archive to journalists).

To understand the trends and issues in IDP research, we carried out surveys on the following topics: (i) sources of research data (results are presented in section 3.1.1); (ii) Inputs (features) of IDPAs (details are presented in section 3.1.2), (iii) Outputs of IDPAs (details are presented in section 3.1.3), and (iv) Overview of IDPAs (details are presented in section 3.1.4). These surveys include different numbers of studies since many of the selected 37 studies focus on only a subset of issues associated with IDP research.

Sources of insider attack data for research

Only a small percentage of studies use original real-world data – in our survey 4 out of 16 studies (Fig. 3). The major sources of real data are the log files of the online systems, and social media websites. Obtaining security data for research purposes is difficult, if not impossible, due to financial, business and national security concerns. Furthermore, even if one acquires real data, data on those conducting insider attacks are not publicly available for privacy reasons [43]. Often, insider attacks are not even caught as attacker leaves no traces. Consequently, research with real-world data proves challenging. For

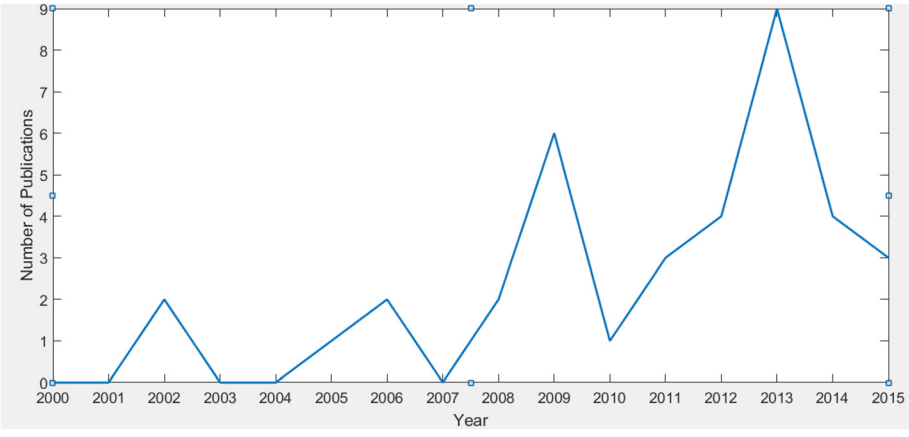

Fig. 2 The trend of security research for the insider cyber threat. The x-axis represents 'Year' and the y-axis represents the number of publications published in any given year

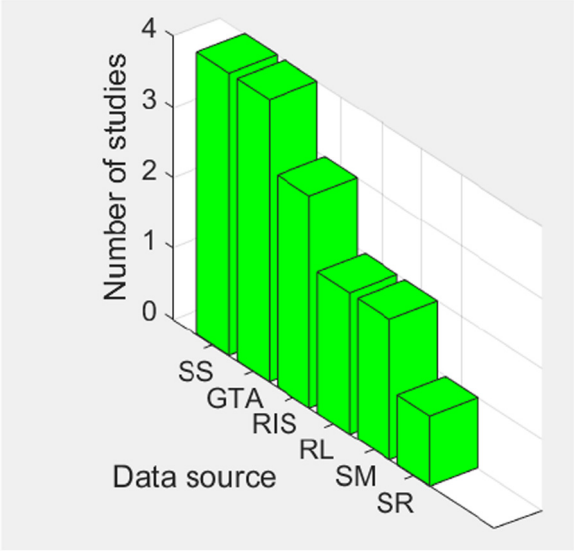

Fig. 3 Usage of various data sources in the selected studies. Real-world system log data is represented by 'RL', social media data by 'SM', real data injected with synthetic anomalies by 'RIS', simulated data drawn from stochastic models which are developed from real data by 'SR', simulated data drawn from stochastic models by 'SS' and game data by 'GTA'. The x-axis represents different data sources and the y-axis represents the number of the studies use the data source

example, Kandias et al. [40] have conducted a content analysis of user comments on YouTube videos looking for any negative comments on law enforcement. Theoretically, these negative comments posted by employees are likely to reflect their intent to commit malicious acts. However, there is no ground-truth whether the individual was involved in any malicious activity.

Our survey reveals that insider threat studies use the following sources of data:

- ➢ Real-world system log data
- ➢ Social media data
- ➢ Simulated data drawn from stochastic models
- ➢ Real data injected with synthetic anomalies.
- ➢ Simulated data drawn from stochastic models which are developed from real data
- ➢ Game-theoretic approach (GTA)

In our survey, the 16 studies use data for assessing the effectiveness of a feature or algorithm [12–15, 18, 21–27, 29, 34, 38, 40]. Figure 3 illustrates the percentage of data sources used in the sample of articles.

The figure shows that the primary sources of research data are the simulated data drawn from the arbitrarily defined models and the data generated through the game (GTA). In GTA, a digital landscape is very carefully created to simulate the real world as closely as possible. The game is based on a set of rules created by experts. Ideally, these rules should simulate the real situation. The game is usually played by multiple players. A central assumption of the game is that the players are rational. A rational player is one who always chooses an action which gives the outcome he most prefers, given what he expects his opponents to do. An insider game is a sequential, stochastic

game with imperfect information. The game is usually played by two teams, one team being called the insiders and the other the defendants. The goal of insiders is to exploit potential attack paths across IT layers that may lead to a security compromise. The goal of defenders is to prevent insiders from achieving their goal. Two teams play in turn. Each team has a set of actions with anticipated costs and expected future benefits of courses of action. On their turn, a team will randomly choose its action based on some probability distribution. On any turn (except for their first turn) a team may opt to play none. Each team's available set of strategy choices are conditional on all previous actions of the opponent. Both teams continue to play until the Nash equilibrium is reached where both teams' strategy is optimized giving what the other player is doing. As players play the game, the data is generated automatically by different players. GTA is not only a good source of research data, it is also a promising source of operational data [16]. Data generated from GTA can be used along with the real data in the case where the real data is not sufficient.

Another popular source of research data is the real data injected with synthetic anomalies. The majority of real world data does not carry information about malicious events. Hence, researchers collect real-wold security data that contains a great deal of legitimate traffic. They simulate insider attacks based on high-profile cases of insider attacks. Then they inject simulated attacks into the real data.

Other sources of research data include the simulated data drawn from stochastic models which are developed from real data. A few studies generate models from real world data and then generate data from estimated models to protect the anonymity and privacy of the users' data and also to control the size of the data stream.

Exploration of the feature space for monitoring insider threats

It is critical that the initial feature space contains all the important features since the success or failure of a pattern recognition system is heavily dependent on the choice of good features. Research suggests that three elements must be present at the same time for an attack to occur: motive, ability and opportunity [39]. Although all of these features are key to the success of insider threat detection and prediction, existing studies concentrate on only a subset of these. Figure 4 depicts the number of studies use each of these feature types. 'Opportunity' is definitively the most widely used feature type for insider threat detection. A total of 36 studies were included in this survey [12–47].

a) Motive The motive refers to the reason or cause why an insider or group of insiders will perpetrate a crime. Previous studies have grouped the features associated with motives into four broad categories [14, 17, 25, 34, 38–40]:

i. Predisposition to malicious behaviour—Honeypots are made to capture the predisposition of the insider towards malicious activities. A honeypot is a trap set to identify malicious insiders. Any interaction with the honeypot is considered abnormal. The usage of honeypots in network measures the probability that an insider will attack against the system if an opportunity arises.
ii. Mental disorders (e.g., paranoia, depression, hearing voices etc.),
iii. Personality factors (e.g., narcissism, neuroticism, sociopathy etc.) and
iv. Current emotional state (e.g., hostility, stress, anger etc.)

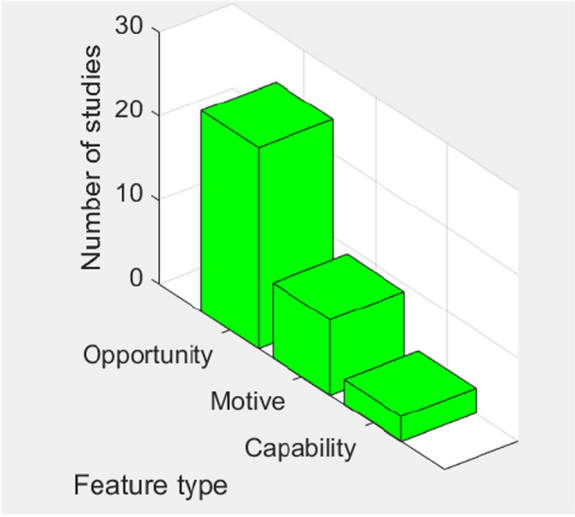

Fig. 4 An overview of the various types of features used in the selected studies. The x-axis represents different feature types and y-axis represents the number of studies use the feature type

Table 1 displays the features that belong to these four categories and the sources from which the values of these features are obtained.

b) Opportunity Crime is all about opportunities. Most studies focus on the following two broad categorizations of features to determine the level of opportunity available to an insider in an organization to perform an attack against the system: insider's role in the system, and activity-based features [39].

Insider's role in the system Each insider has a system role associated his/her user account, such as being a system administrator, advanced user or novice user. A system role defines the kind of access to the system an insider has. A system role also determines if an insider has the privileges required to do a function or operation in the system.

Activity-based features For insider threat detection and prediction, key sources of data are log files. These log files provide real-time data on the usage of resources, users, roles and activities [39]. A sensor can be set to any application or device to log data autonomously. Table 2 lists activity-based features, their major groups and sources of information on these features. The following seven categories of activity-based patterns are being explored widely: (i) file domain, (ii) database domain, (iii) Email domain, (iv) HTTP domain, (v) mobile calls, (vi)print domain, (vii) TCP/IP network flows, and (viii) Other applications such as MS Word, MS Power Point, MS Excel, JPG, and TXT. Figure 5 shows the number of selected papers used these feature vectors. A total of 35 studies were included in this survey [12–24, 26–47]. Table 2 presents the features that belong to these domains and the sources of information for each feature domain.

Our survey reveals that the most widely used feature domains are the file domain and the database domain, and the least used feature domain is the device domain (Fig. 5). Eldardiry et al. [22] compared the independent predictive powers of file, device,

Table 1 Key features and sources of information regarding the motivation

Feature domain	Features	Sources of information
Predisposition to Malicious Behaviour	# Predisposition towards Law Enforcement & authorities # Demonstation of delinquent Behaviour in the Past	# Social Networking Sites # Interaction with Honey Files # Employees' criminal record history
Mental disorders	# Paranoia # Depression, # Schizophrenia # Bipolarity	# Clinical Diagnoses # Social Networking Sites
Personality factors	# Narcissism # Psychopathy # Sociopathy # Neuroticism, # Conscientiousness # Openness, # Self-focuse # Social Distancing # Moral disengagement # Agreeableness # Excitement seeking, a facet of Extraversion # Feelings of Closeness # Sense of Collective Responsibility, # Blaming or Devaluation of a Victim, # Perception of Punishmensy and Balance of Punishment & Rewards	# Language Style Matching (LSM) # Email Communication Patterns # Website Visite Patterns # Temporal Pattern Analysis of Online and PC Usage Behaviour # Sentiment Analysis of an Insider's Communications # Visits of Socila Media Websites
Negative emotion	# Hostility # Anger # Stress	# Dynamic Environmental Stressors including personal life stressors & job stressors # Patterns of Communication and Social Interaction

logon, http and email domains. According to this study, file and device domains are the most important feature domains, and logon and http domains are the least important feature domains.

c) Capability Capability is used to determine the extent to which an insider can compromise a system with his/her current level of information technology (IT) sophistication. "Capability' represents the demonstrated skill level of an insider monitored by the IT system. The feature 'Capability' is broken down into three sub-features [33]:

i. Unique applications run by the insider in different sessions: Typically, the higher the number of unique applications, the higher the insider's sophistication.
ii. Multiple applications run simultaneously by the insider in different sessions: The higher the number of applications run simultaneously per session, the more sophisticated the insider.
iii. The consumption of CPU and RAM by the user in different sessions: The higher the CPU and RAM usage, the higher the sophistication level of insiders.

Output variable

The output of IDPAs is whether or not a pattern is anomalous [12–47]. The magnitude of anomaly is an indication of future attacks. The higher the abnormality,

Table 2 Sources and features of the insider's online activities

Domain name	Source	Features
Logon	Daily System logs	# Auxiliary Information such as user ID, PC ID, activity code, time of day # after hour logons, # logons on user's PC, # logons on other PC(s), # login duration time, # login frequency.
File	Daily System logs, Access logs	# Auxiliary Information such as user ID, PC ID, activity code, time of day, # accessed directory(s), # file created, copied, moved, written, renamed or deleted,
Database	Database audit logs	# Auxiliary Information such as user ID, PC ID, activity code, time of day # Which data items were accessed? # Were any modifications made? .
HTTP	Web server logs	# Auxiliary Information such as user ID, PC ID, time of day # URL and domain information activity codes (upload or download) # URLs visited # Whether the website is encrypted # browser information (internet explorer, Firefox, or Chrome)
Removable device (e.g., USB drives)	Event logs	# Auxiliary Information such as user ID, PC ID, activity code, time of day # Device name and type are logged with usage code
Email	Email transaction logs	# Auxiliary Information such as user ID, PC ID, activity code, time of day # Source and destination of email traffic # Communication patterns # Attachment names
Mobile calls	Call logs	# Source and destination of mobile calls # Duration, date and time of calls # Communication patterns
Print	Printer activity logs	# Auxiliary Information such as user ID, PC ID, activity code, time of day # Name of document printed # Number of copies
TCP/IP network flows	TCP/IP network flow logs	# The source and destination of IP packets on a TCP/IP network # The size of traffic sent over the connections # t The average duration of connections # Positive and failed events from different IP addresses # Time difference between IP events
Other Applications (e.g., MS Word, MS Power Point, PDF, MS Excel, JPG, TXT)	Event logs, Error logs	

the higher the risk of an insider attack. Any sudden changes in individual and group behaviours are considered to be malicious. Typical output is a point estimate of the likelihood for an attack. A sole point estimation cannot account for the uncertainties associated with the estimation. Only two out of the 37 selected studies provide interval estimates of the risk of an insider attack at time step t—they estimate the probabilities of low, medium and high risks of the insider threat at time step t.

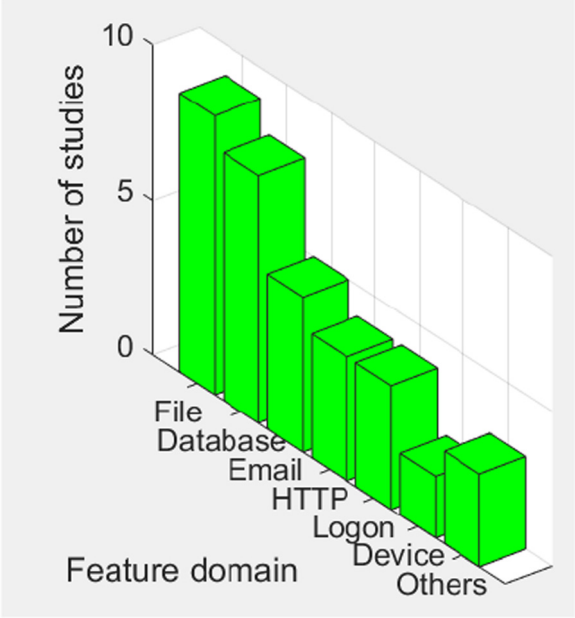

Fig. 5 Application of various feature domains in the selected studies. The x-axis represents different feature domains and the y-axis represents the number of studies use the feature domain

Overview of Insider Threat Detection and Prediction Algorithms (IDPAs)

IDPAs can be classified into different categories based on any of the following considerations: (i) features being used, (ii) the detection strategy being used, and (iii) the main underlying machine learning algorithm of the IDPA. Classifications of IDPAs are briefly discussed below.

Classification of IDPAS based on the features being used

Based on the features used to model the threat, the existing IDPAs fall into six model categories: i) intention models (IM), ii) individual behaviour models (IBM), iii) community behaviour models (CBM), iv) resource usage models (RUM), v) capability model (CM), and vi) mixed system (MS). These models are briefly discussed below. The observed percentages of insider threat detection and prediction algorithms with various types of models are shown in Fig. 6. A total of 31 studies were included in this survey

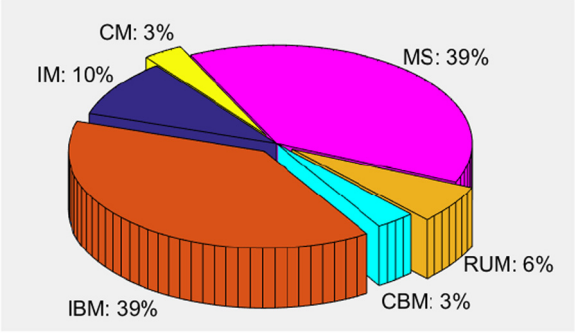

Fig. 6 Application of various models in the existing IDPAs. 'IDPAs' represents insider threat detection and prediction algorithms. Individual behaviour model is represented by 'IBM', community behaviour model by 'CBM', resource usage model by 'RUM', mixed system by 'MS', capability model by 'CM', and intention model by 'IM'

[12–32, 34, 35, 37–42, 44, 45]. Our survey reveals that nearly 80 % of the existing IDPAs are either IBM or MS.

Intention models

Assess potential threats based on the psychological profiles of insiders. Only about 10 % of IDPAs are intention models (Fig. 6).

Individual behaviour models

Assess threats by examining the insider's normal course of activities. These models learn regular patterns of activity for each insider from the insider's past activities. Activities can be anything that involves information systems, including more CPU and RAM usage, accessing a restricted file, altering the content of the file, and copy the file to their Google Drives. This is one of the most popular models used by any IDPA (Fig. 6). Nearly 40 % of IDPAs are based on individual behaviour models.

Community behaviour models

Assess potential threats from an insider based on the normative behaviour of the community within which the insider belongs. This approach is based on two assumptions: (i) an insider's activity pattern changes continually due to a change in role or situation and (ii) users/insiders tend to be team and goal oriented. It works on the principle that all members who access similar resources will also exhibit a similar behaviour. Therefore all these members together form a dynamic community centred on a single resource where the membership changes by adding or losing individuals during the course of its existence. These algorithms learn regular patterns of activity for each insider based on the community members' contemporary activity patterns. Only 3 % of IDPAs are community behaviour models (Fig. 6).

Resource usage models

Assess the security threat posed to the system by examining the cumulative effect of collective activities of all insiders on a particular resource(s) (e.g., file server, database server). Only 6 % of IDPAs can detect collusion attacks by a group of malicious insiders (Fig. 6).

Capability models

Assess insider threat based on insiders' level of IT sophistication. It monitors insiders IT resource usage activity. For example which IT applications an employee usually uses, how the employee executes the application programs (from command-line or dialog interfaces), how many applications an insider runs simultaneously, and so on. The capability model can be replaced by the individual behaviour model, since the individual behaviour model should be able to automatically recognize the IT capability of the insider from the insider's activity streams. Hence only 3 % of IDPAs are capability models (Fig. 6).

Mixed systems

Employ more than one of the above mentioned models at a time. Nearly 40 % of IDPAs are mixed systems (Fig. 6).

Classification of IDPAs based on detection strategies

Based on security threat detection strategies, the existing insider threat detection and prediction algorithms are clustered into two groups: i) anomaly detection (AD), and ii) signature-based intrusion detection (SID). The SID approach discovers the patterns of the known attacks and then examines ongoing traffic, activity, transactions, or behaviour for matches with known patterns of attacks. However, it is well known that SIDs are only able to detect known attacks. On the other hand, the AD approach can detect previously unknown attacks. It learns normal behaviour patterns and classifies patterns that deviate from established normative patterns as anomalies. Figure 7 shows how many of the existing IDPAs use AD and how many use SID. A total of 29 studies were included in this survey [12–16, 18–31, 34, 35, 37–42, 44, 45]. Our survey reveals that the AD approach is more popular than the SID approach. The AD approach is more popular because it can detect previously unknown attacks.

Classification of IDPAs based on the underlying machine learning algorithms

Choosing the right machine learning algorithms is crucial for accurate estimation of insider threats. We performed a survey to detect any trend in the use of machine learning algorithms. A total of 30 out of the 37 selected studies were included in this survey [12–31, 34, 35, 37–42, 44, 45]. Our survey shows a clear preference towards graph-based learning algorithms (i.e., Bayesian Networks—BN) (Fig. 8). Insider threat detection and prediction are multivariate time series problems. By analogy, a time series is a directed graph. Perhaps this is the reason why dynamic BNs have been studied so extensively. The other underlying learning algorithms include: SVMs (Support Vector Machines), FIS (Fuzzy Inference Systems), Gaussian Mixture Models (GMM), K-nearest neighbour algorithm (KNN), game theoretic approach (GTA), and regression-based

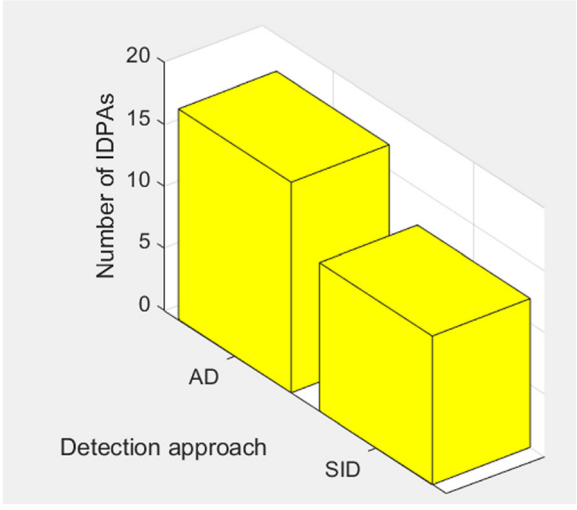

Fig. 7 The application of detection methods in the selected IDPAs. 'IDPAs represent insider threat detection and prediction algorithms. 'AD' represents anomaly detection and 'SID' represents signature-based intrusion detection. The x-axis represents different detection approaches and the y-axis represents the number of IDPAs employ the detection approach

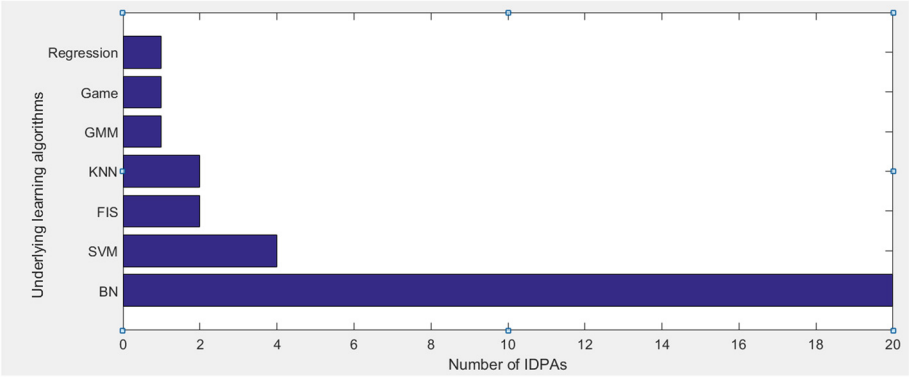

Fig. 8 The main machine learning algorithms used in the existing insider threat detection algorithms. The y-axis represents different learning algorithms and the x-axis represents the number of insider threat detection and prediction algorithms (IDPAs) employ the learning algorithm

approach. Figure 8 shows the number of different learning algorithms used in the insider threat detection and prediction algorithms.

Challenges associated with IDP

We reviewed the selected 37 studies carefully to identify the challenges in the IDP [9, 12–47]. The results are briefly discussed below.

Challenge 1: unbounded patterns vs. time lags between activities

Insider threat detection is a temporal phenomenon. Therefore, features can be in frequency domain or in time domain. However, for this particular problem, any sudden change in behaviour is monitored, since sudden changes in behaviour and actions can be a signal that the insider is involved in malicious attacks. Hence, a short-term analysis is performed on signals from insider behaviour. That means the signals must be in time phase. In the time domain, the correct order as well as the time span between activities is important. Most features are online commands. So these features are nominal variables. For example. Saving a file to a removable disk or network drive is a nominal variable. These variables can have only two possible values—1 (if the command is invoked) or 0 (if the command is not given). There is no in-between state. Now, the question is, how we can capture both the sequence of and the time lag between commands. One way is to consider the last k commands to determine if the insider activity is malicious or benign. This strategy will fail to capture the time lag effect. Another option is to divide the time into equal time steps and then consider all of the commands during the last k-time steps as predictors. This strategy will generate the space of unbounded patterns. The only algorithm that can handle the unbounded patterns is graph algorithms (BN). Dynamic Bayesian Networks (DBNs) is one of the best solutions available for incorporating both the sequence and the varying time difference between activities.

Challenge 2: data non-stationarity vs. individuality

Insider threat data is highly irregular non-stationary multivariate time series. Having permanent global technology shock causes non-stationarity in the signal, where a sudden technological breakthrough can make current IT infrastructure and processes

obsolete in a very short space of time. Moreover, employees' roles, business processes, systems and organization structures change all the time.

The problem with non-stationary data is that the multivariate joint probability distribution of variables is not stationary in time, the underlying situation (probability distribution) keeps changing. One possible solution is to use a Community behaviour model (CBM) since they do not use temporal patterns to identify malicious insiders. CBM learns regular patterns of activity for each insider based on the community members' contemporary activity patterns. However this model cannot incorporate the individuality of insiders' unique working styles. On the other hand, Individual Behaviour models (IBMs) are designed to capture the individuality. IBM model uses the temporal changes of insider's behaviour to anticipate insider attacks. The problem is that IBM cannot deal with non-stationary data. Hence several studies propose a mixed system where these two models (IBM and CBM) are used together.

Challenge 3: high dimensionality vs interacting features

In insider threat detection, most of the features are nominal variables. If a nominal variable has *k* categories, *k-1* dummy variables are needed to index the *k* categories. Moreover, IDP problems are multivariate time-series problems. In a time series process, each time step is treated as a separate feature. As a result, the space is extremely high-dimensional. One way to handle the high-dimensionality is using the ensemble of classifiers approach where each classifier is trained with a small subset of features. However, this approach cannot handle the interacting features. A single classifier with all input features can handle interaction effects but suffers from high-dimensionality. The only algorithm that captures both the high-dimensionality and the interaction effects is the Bayesian networks since it has an inherently hierarchical structure.

Challenge 4: detecting collusion attacks

Two or more insiders may collude together to commit a malicious act without creating any suspicion. For example, a large group of insiders can all make a few small changes to a restricted file to corrupt the data. Small changes on the part of an insider can't be distinguished by the individual behaviour model or the community behaviour model. Detecting collusion attacks requires an aggregate view of the insider activity on a resource. Resource Usage Models assess the security threat posed to the system by examining the collective usage patterns of a particular resource.

Challenge 5: reducing false alarms

Reducing false alarms without compromising catch detection is a big challenge for insider threat detection. Conventionally, an AD approach is used for an insider threat detection. In AD approach, any deviation from the normal behaviour pattern is considered as an abnormality where an abnormality is considered to be malicious. However, most anomalies are not malicious. This approach is bound to create many false alarms. Several studies suggest using an intention model along with other models. Intention Models assess potential threats based on the psychological profiles of insiders. Whenever a conventional model detects an anomaly in the insider's behaviour, the intention model is consulted to assess whether there is any reason to believe that the insider has any motive to be malicious. If the insider possesses a significant degree of interest and motivation, only then an alarm is raised.

Challenge 6: class Imbalance problems and undetected insider attacks

Class imbalance is prevalent in insider threat detection and prediction. The class imbalance problem occurs when the number of instances of one class (positive class) is much lower than the instances of the other class (negative class). The problem is that in such case, classifiers always predict the majority class. To make the matter worse the majority of insider attacks are undetected. Traditional approaches of dealing with the class imbalance problems include increasing the base rate (i.e., the prior probability) of an insider attack, oversampling the minority class and over-penalizing false negatives. However, these approaches increase false positives. AD is the best available strategy to address class imbalance and undetected insider attacks. Two standard anomaly detection strategies are: i) supervised one-class classification, and ii) unsupervised classification. None of these two approaches use actual class labels: malicious or benign. Supervised one-class classifiers consider all instances to be normal. It defines the probability of normality of a particular pattern based on the relative frequency of its occurrence which forms the new output variable. Then the classifier is trained in usual way using the dataset with the new output values. When a new input pattern is presented to the classifier, it predicts the probability that the pattern is normal. In contrast, unsupervised classification groups patterns with similar features into distinct clusters. When an unknown pattern is presented, the algorithm estimates the probability that the pattern is benign based on the distance of the new pattern from its closest cluster centre. Both of these AD approaches (on-class classification and unsupervised classification) can effectively address the problems of class imbalance and undirected insider attacks.

Challenge 7: uncertainty

Insider threat detection and prediction algorithms predict the probability of malicious attacks. Most of the algorithms developed so far predict point estimates (i.e., single probability values) since it is simple to predict point estimates from a nonparametric model. However, a single probability value fails to capture the uncertainty associated with the estimate. Hence interval estimates are preferred over point estimates. A few existing studies use the FIS as an interval estimation mechanism. The FIS provide estimates for the probabilities that a certain pattern pertains to (i) “high risk pattern”, (ii)“medium risk pattern” and (iii) “low risk pattern”. The pattern is classified as belonging to the risk category, which assigns the highest membership probability to the pattern. The given pattern can belong to all three risk categories. However, if the pattern belongs to more than one risk category with high membership probabilities, then the uncertainty is high.

Challenge 8: the number of free parameters in the model

The higher the number of free parameters in the model that must be optimised simultaneously, the higher the computational burden and the risk of overfitting effects.

Comparative analysis of various IDPAs

Not all of the selected papers propose novel algorithms. Moreover, many of the proposed algorithms have never been implemented. In only 13 papers out of the 37 selected papers, authors have actually implemented and evaluated their proposed algorithms. Hence, we selected these 13 papers for the meta-analysis [12–15, 18, 21–27, 29]. Figure 9 depicts the

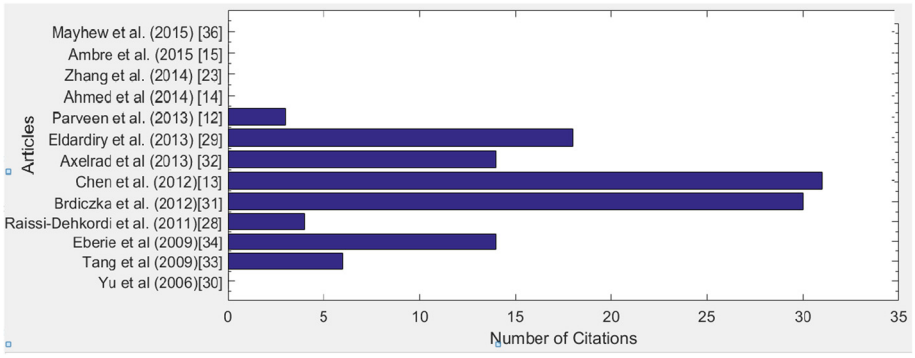

Fig. 9 Number of citations for each selected publication based on the Google Scholar. Only those publications are displayed where the authors implemented their proposed algorithms and the evaluation are presented. The horizontal x-axis represents the number of citations for publications for the author on the vertical y-axis

total number of citations received by all these 13 papers based on Google Scholar's citation count. We also looked at the papers that cited these 13 selected research papers. However, none of these citation papers use the IDPAs presented in these research papers. The most cited papers are: Chen et al. [13] and Brdiczka et al. [24].

Each of the 13 selected papers proposed one IDPA except for one paper (Parveen et al. [12]). In this paper, Parveen et al. [12] present two IDPAs. These algorithms are—GBAD and OCSVM. In the rest of the paper, we refer GBAD by 'Parveen et al. [12]: GBAD' and OCSVM by 'Parveen et al. [12]: OCSVM'. An overview of all these 14 algorithms are provided in Additional file 1: Table S1. The authors of these 14 algorithms have evaluated their proposed algorithms on different datasets. The majority of the studies surveyed are evaluated on large datasets. Since they mainly use synthetic data or real data with only benign activities, the size of the data is not a problem in this case. The noise level is the prime factor to determine the optimal data size; and noise levels vary from data to data. Therefore, no comparisons can be made among different algorithms and studies in terms of the size of the datasets used. In addition these algorithms are evaluated on the predictive performance using different metrics. For example, Axelrad et al. [25] measure performance in terms of error rate, logarithmic loss and quadratic loss. Zhang et al. [18] and Ambre & Shekokar [15] do not offer a predicting performance assessment. Tang, Zhou and Wang [26] measure performance in terms of inference and learning time. Our goal is to compare all the algorithms using the true positive rate (detection rate) and the false positive rate (false alarm rate). Detection rates (DRs) and false alarm rates (FARs) have been derived from the reported performance values that are found in the individual papers. We are able to convert the performance values published in nine studies [12–14, 21–24, 27, 29] (out of the 13 studies) where levels of information are sufficient to calculate DR and FAR. The estimated DRs and FARs of ten IDPAs (from nine studies) are shown in Additional file 1: Table S1 and Fig. 10. Figure 11 displays the Euclidean distance of the performance of each of the ten algorithms from the desired performance metrics (DR: 100 % and FAR: 0 %).

Since the IDPAs are evaluated on different datasets, we can't compare their empirical performance. However these empirical results do provide two interesting observations:

First, OCSVM and GBAD proposed by Parveen et al. [12] were evaluated on the same dataset and GBAD's performance was much worse than that of OCSVM.

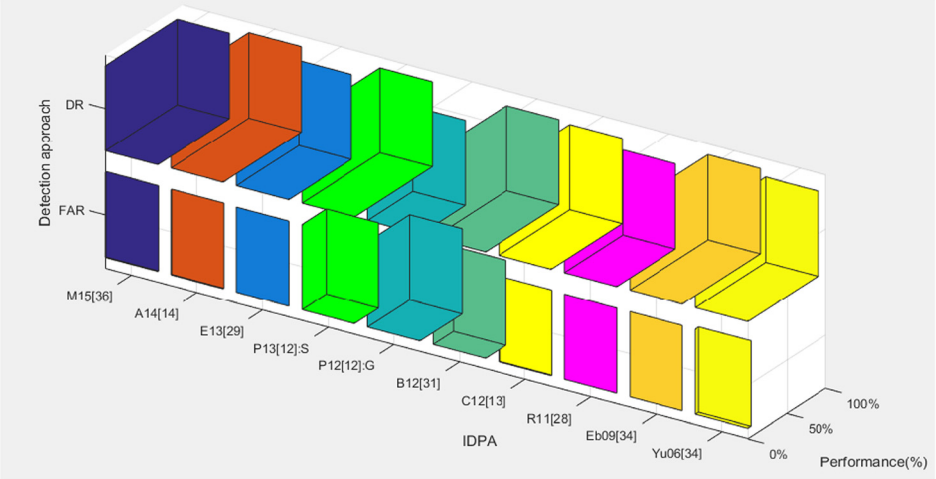

Fig. 10 Detection rates and false alarm rates of the IDPAs (taken or estimated from the original paper by the authors of the IDPA). 'IDPA' represent insider threat detection and prediction algorithm'. 'DR' represents 'Detection Rates' and 'FAR' represents 'False alarm rates'. 'M15 [29]' represents the IDPA proposed by 'Mayhew et al. [29]', 'A14 [14]' represents 'Ahmed et al. [14]', 'E13 [22]' represents 'Eldardiry et al. [22]'. 'P13 [12]:S' represents Parveen et al. [12]: OCSVM', and 'P13 [12]:G' represents 'Parveen et al. [12]: GBAD'—Parveen et al. [12] present two algorithms—OCSVM & GBAD. 'B12 [24]' represents 'Brdiczka et al. [24]', 'C12 [24]' represents 'Chen et al. [24]'. 'R11 [21]' represents 'Raissi-Dehkordi et al. [21]'. 'Eb09 [27]' represents 'Eberie et al. [27]'. 'Yu06 [23]' represents 'Yu et al. [23]'

Second, Eberie et al. [27] propose a similar algorithm (they also call their algorithm 'GBAD') and their algorithm performed extremely well on the real-world datasets. Their difference is that Parveeen's GBAD is an ensemble comprising k number of BNs compared to Eberie's GBAD which is a single BN model. Parveen's GBAD is a complex algorithm. Learning an optimal BN is a NP-hard problem and Parveen's GBAD consists of multiple BNs. In fact it is possible that the authors would not find the optimal model solution. Also, the way the outputs of the ensemble members are combined in Parveen's GBAD is questionable as they perform simple majority voting or weighted voting to reach consensus. These linear approaches cannot account for the interaction effects.

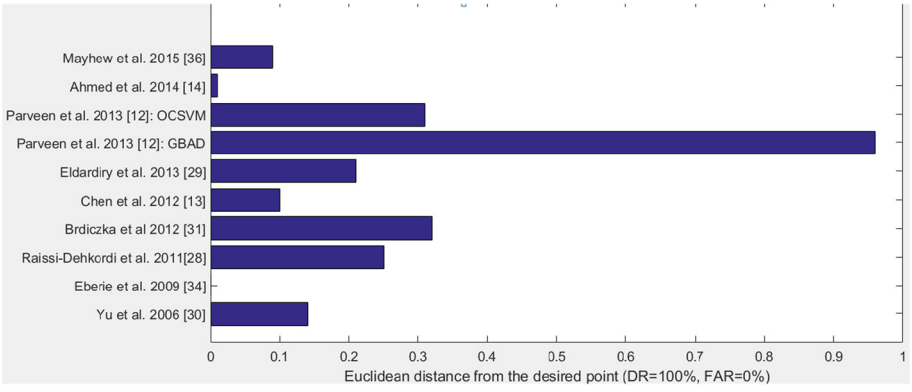

Fig. 11 Euclidean distance of the performance vector of each of the IDPAs from the desired performance vector. The performance vector includes detection rates (DR) and false alarm rates (FAR) that are extracted from the papers by the authors of the algorithms. The desired performance vector is [DR = 100 %, FAR = 0 %]. The x-axis represents Euclidean distance and y-axis represents the different IDPAs (insider threat detection and prediction algorithms) proposed by different authors. The higher the bar, the lower the performance

Additional file 1: Table S1 does not report the learning time of the algorithms, since among 14 algorithms only one algorithm was evaluated in terms of the learning time. However, the table shows the number of free parameters for the chosen 14 IDPAs. The learning time is proportional to the number of free parameters in the model. The higher the number of free parameters, the longer the learning time.

A comparison of the IDPAs based on theoretical merits

We compare the IDPAs according to their theoretical abilities to address the challenges (the challenges along with the best available ideas for achieving solutions are described in section 3.3) associated with IDP. If an IDPA can address a challenge concerning IDP, it scores 1 point. Table 3 shows the scoring of each of the 14 algorithms for each criterion.

Figure 12 compares the total scores of all of 14 algorithms over all criteria. The higher the score the better the algorithm. The top three algorithms based on their theoretical merits are Mayhew et al. [29], Brdiczka et al. [24] and Eberie et al. [27].

Figure 13 shows how many algorithms are capable of addressing each of the 10 challenges. The majority of algorithms cannot cope with the collusion attack, uncertainty, and data non-stantionarity; while the majority of algorithms can address individuality, interaction effects, class imbalance and undetected insider attacks.

Publication quality

Our study reveals that none of the selected IDPAs are assessed rigorously by neutral experts. This observation suggests that there may be publication bias. Publication bias is a potential threat in all areas of research. Hence it is important to evaluate publication bias in studies. Bias articles provide vague information regarding design, implementation and evaluation of the proposed algorithms so that no one else can try to replicate the study. We measure the publication quality based on ten questions on a 5-point scale about how transparent a publication is. The goal of these questions is to determine if the findings of the study are replicable. Five studies have a low risk of bias – Chen et al. [13], Eberie et al. [34], Parveen et al. [12], Raissi-Dehkordi et al. [21], and Yu et al. [23] and the remaining eight studies have a high risk of bias (Table 4, Fig. 14).

Meta-Analysis: publication-bias adjusted ranking of IDPAs

The most promising algorithms are those that are both theoretically strong and easily implementable. Hence, we adjust the rankings of all algorithms for the publication bias. To do this, we first rank the algorithms in terms of their total scores for theoretical merit and publication bias. We give the rank of 1 to the lowest score, 2 to the next lowest, and so on with N being assigned to the most promising algorithm with the highest score (Table 5).

We then estimate the rank sum of each algorithm (column 3 of Table 5). We perform *the Friedman two-way analysis of variance by ranks* ([11], pp.172–180) to test the null hypothesis that there is no significant difference among 14 algorithms. To do this test, we calculate the value of the statistic which we denote as F_r

$$F_r = \left[\frac{12}{Nk(k+1)}\sum_{j=1}^{k} R_j^2\right] - 3N(k+1)$$

Where N = Number of algorithms = 14

Table 3 Comparisons of the proposed Insider threat detection algorithms based on theoretical strengths

Criteria	Ambre et al. 2015 [15]	Mayhew et al. 2015 [29]	Ahmed et al. 2014 [14]	Zhang et al. 2014 [18]	Axelrad et al. 2013 [25]	Eldardiry et al. 2013 [22]	OCSVM: Parveen et al. 2013 [12]	GBAD: Parveen et al. 2013 [12]	Brdiczka et al. 2012 [24]	Chen et al. 2012 [13]	Raissi-Dehkordi et al. 2011 [21]	Eberie et al. 2009 [27]	Tang et al. 2009 [26]	Yu et al. 2006 [23]
Unbounded patterns and time lags between activities[a]	1	0	0	1	1	0.5[k]	0	1	1	0	0.5[l]	1	1	0
Data Non-Stationarity[b]	0	1	0	1	0	1	0	0	0	1	0	0	0	0
Individuality[c]	1	1	1	1	1	1	1	1	1	0	1	1	1	1
High Dimensionality[d]	1	0	0	1	1	1	0	1	1	0	0	1	1	0
Interaction effects[e]	1	1	1	0.5[m]	1	0.5[n]	0[o]	0[o]	1	1	1	1	1	1
Collusion Attacks[f]	0	1	0	0	0	0	0	0	0	0	1	0	0	0
False alarms[g]	0	1	0	0	0	0	0	0	1	0	0	1	0	0
Class imbalance & undetected insider attacks[h]	0	1	1	1	0	1	1	1	1	1	1	1	0	1

Table 3 Comparisons of the proposed Insider threat detection algorithms based on theoretical strengths *(Continued)*

Uncertainty[i]	0	0	1	0	0	0	0	0	0	0	0	0	0	1
Number of free parameters[j]	0	0.14	0.14	0	0	0	0.25	0	0	1	0.25	0	0	0.14
Total	4	6.14	4.14	5.5	4	5	2.25	4	6	4	4.75	6	4	4.14

[a]Criterion 1: Unbounded patterns vs. time lags between activities: If the proposed algorithm is based on a hierarchical Bayesian networks (HBNs), then the algorithm scores 1 point; otherwise it scores zero
[b]Criterion 2: Data non-stationarity: If the algorithm employs a CBM(s), then it scores 1 point; otherwise it scores zero
[c]Criterion 3: Individuality: If the algorithm employs an IBM(s), then it scores 1 point; otherwise it scores zero
[d]Criterion 4: High dimensionality: If the algorithm is based on a HBN(s), then it scores 1 point; otherwise it scores zero
[e]Criterion 5: Interaction effects: If the algorithm is based on a HBN(s) or a machine learning algorithm that uses all features simultaneously to predict the output, then it scores 1 point; otherwise it scores zero
[f]Criterion 6: Collusion attacks: If the proposed algorithm employs a RUM, then it scores 1 point; otherwise it scores zero
[g]Criterion 7: False alarms: If the insider threat detection algorithm employs an IM(s), then it scores 1 point; otherwise it scores zero
[h]Criterion 8: Class imbalance and undetected insider threats: If the insider threat detection algorithm adopts the AD-based approach, then it scores 1 point, otherwise it scores zero
[i]Criterion 9: Uncertainty: If the insider threat detection algorithm is based on a FIS, then it scores 1 point; otherwise it scores 0
[j]Criterion 10: The number of free parameters in the model: Each algorithm receives a score that is the inverse of its number of free parameters. We assign score zero to the HBNs since the HBN-learning problem is known to be NP-hard
[k]The proposed algorithm is a hybrid of the Markov networks and GMMs. Markov networks can handle unbounded patterns but GMMs cannot handle unbounded patterns. So the score is 0.5
[l]The proposed alhorithm consists of two types of models: individual behaviour models and resource usage models. All of the variables in the resource usage model is in frequency domain. Hence resource usage model is not affected by the challenge. However, the variables of individual behaviour model is in time domain and is susceptible to the challenge
[m]This algorithm is a hybrid of Naïve Bayes (NB) and HBNs. HBNs can capture feature interactions, but NB cannot capture feature interactions. So the score is 0.5
[n]This algorithm is a hybrid of of GMMs and Markove networks. In this algorithm, GMM deals with a small subset of features at a time. Therefore the GMM cannot address the interaction of features. However, the Markov network can address the interactions of features. Hence the score is 0.5
[o]GBAD & OCSVM are the ensembles of classifiers. The outputs of these classifiers are combined to produce the final output. Hence these algorithms cannot address the interaction effects

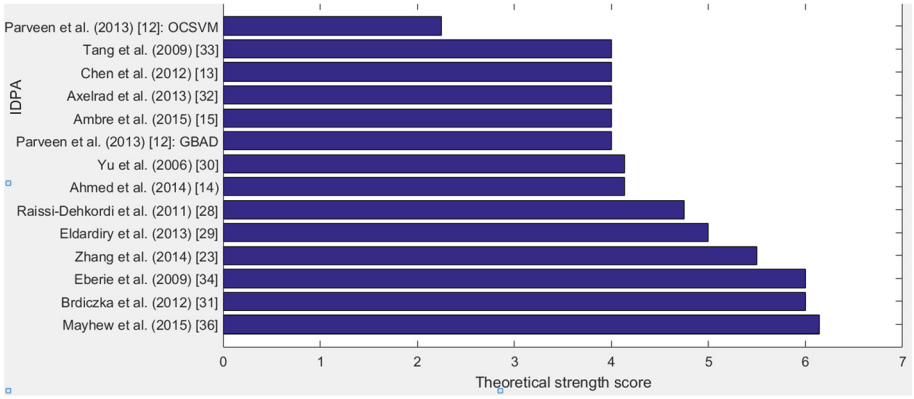

Fig. 12 Comparison of IDPAs based on the theoretical merits. The horizontal x-axis represents the theoretical strength score for IDPAs proposed by the authors on the vertical y-axis. The higher the score, the better

k = number of variables or conditions =2 [variables are theoretical merits and publication bias]

R_j = sum of ranks for the jth variable

$\sum_{j=1}^{k}$ directs one to sum the squares of the sums of ranks over all variables

$$\sum_{j=1}^{2} R_j^2 = (54)^2 + (84)^2 = 9972$$

$$F_r = \left[\frac{12}{14 \times 2 \times (2+1)} \times 9972\right] - 3 \times 14 \times (2+1) = 6172$$

Since the test statistic (6172) is larger than the tabled value of F_r (6.00) at the 0.05 significance level, we reject the null hypothesis. This indicates that at least some of these 14 algorithms are significantly different. In order to find which algorithm are significantly different, we need to perform pairwise tests. So we perform 'Multiple Comparisons between Groups' ([11], pp.181-183) to test the null hypothesis that if there is

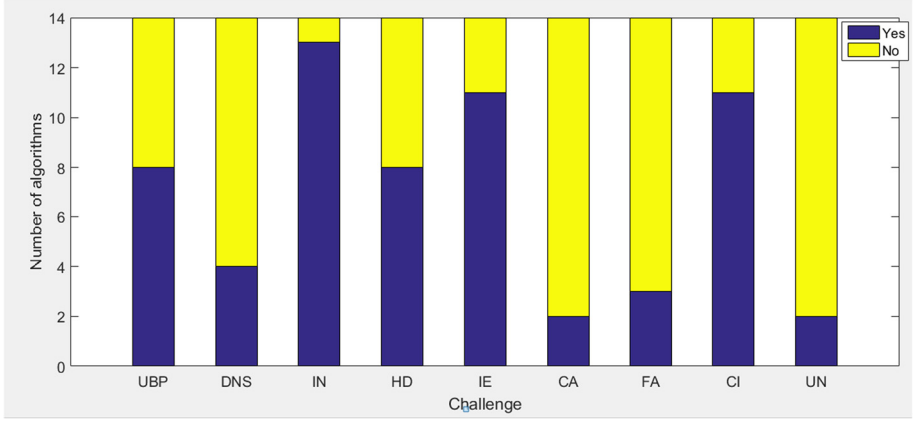

Fig. 13 The number of algorithms capable of addressing diverse challenges. 'Yes' represents the number of algorithms (IDPAs) that can address a particular challenge and 'No' represents the number of algorithms that cannot address the challenge. 'UBP' stands for 'Challenge 1: unbounded patterns vs. time lags between activities'. 'DNS' stands for 'Data non-stationarity'. 'IN' stands for 'individuality'. 'HD' stands for 'high dimensionality'. 'IE' stands for 'interaction effects'. 'CA' stands for 'Collusion Attacks'. 'FA' stands for 'False Alarms'. 'CI' stands for 'Class Imbalance problems and undetected insider attacks' and 'UN' stands for 'Uncertainty'

Table 4 Risk of bias assessment. Assessment of sources of risk of bias within publications

Criteria	Ambre et al. (2015) [ref] [15]	Mayhew et al. 2015 [ref] [29]	Ahmed et al. 2014 [14]	Zhang et al. 2014 [18]	Axelrad et al. 2013 [25]	Eldardiry et al. 2013 [22]	Parveen et al. 2013 [12]	Brdiczka et al. 2012 [24]	Chen et al. 2012 [13]	Raissi-Dehkordi et al. 2011 [21]	Eberie et al. 2009 [27]	Tang et al. 2009 [26]	Yu et al. 2006 [23]
Explain and justify the design and choice of components used in the proposed algorithm[a]	2	3	2	3	1	3	3	3	2	3	3	4	4
A clear description of the exact features used to train the proposed algorithm is given[a]	3	1	1	4	1	3	4	3	4	4	4	2	4
Feature selection method is clear[a]	4	3	5	3	4	3	3	1	4	4	4	2	4
Model parameter optimization method is cearly described[a]	3	3	5	1	3	2	3	1	4	4	5	2	4
A pseudocode of the proposed algorithm is presented[a]	1	1	1	3	1	3	4	3	4	4	4	4	4
The proposed algorithm is compared with the benchmark algorithms[b]	0	0	0	0	0	0	5	0	5	5	0	5	5
The benchmark algorithms are chosen carefully[a]	0	0	0	0	0	0	2	0	2	3	0	3	4
Detailed evaluation results are provided[a]	1	5	4	1	2	3	4	3	4	3	5	1	5
The key characteristcs of the experimental dataset are clearly described[a]	2	1	1	3	1	2	3	3	3	3	4	1	5
The experimental data are made available to other researchers[b]	0	0	0	0	0	0	5	0	0	0	5	0	0
Sum	16	17	19	18	13	19	36	17	32	33	34	24	39

[a]Rating scale: Very good=5, Good= 4, Average=3, Poor=2, Very poor=1, Absolutely no information =0
[b]Rating scale: Yes=5, No=0
Total score less than 25= High risk of bias

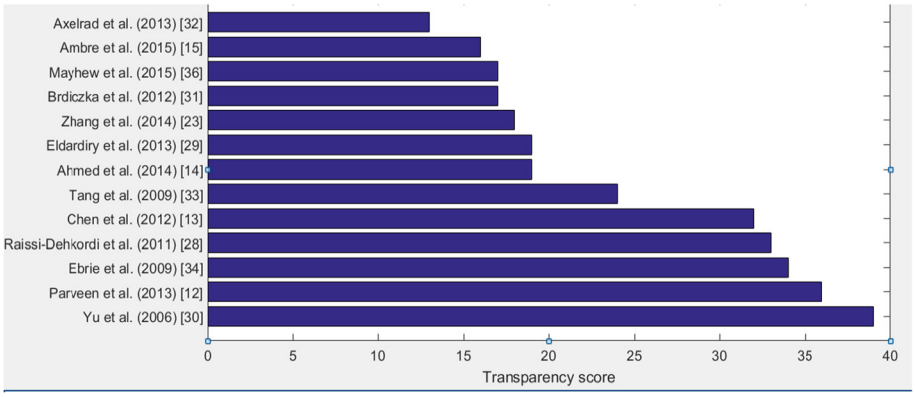

Fig. 14 Comparison of articles based on the transparency score. The horizontal x-axis represents the transparency score for publications for the authors on the vertical y-axis. The higher the transparency score, the less bias the article

any significant difference in performance between pairs of algorithms at the significance level of 0.05 (α = 0.05) by using the following inequality

$$|R_u - R_v| \geq q(\alpha, \#c)\sqrt{\frac{Nk(k+1)}{6}} \; where \;\; R_u > R_v$$

Where R_u = Rank sum of algorithm 'u' over theoretical merits and publication bias
R_v = Rank sum of algorithm 'v' over theoretical merits and publication bias
α = 0.05 and # c = k−1 = 2−1 = 1
Tabled value of $q(0.05, \# c)$ for one-tailed test is 1.645. Therefore

Table 5 Rankings and rank sums of IDPAs

Algorithm	Ranking based on theoretical merits	Ranking based on publication bias	Rank sum
Amber et al. 2015 [15]	2	2	4
Mayhew et al. 2015 [29]	8	3	11
Ahmed et al. 2014 [14]	3	5	8
Zhang et al. 2014 [18]	6	4	10
Axelrad et al. 2013 [25]	2	1	3
Eldardiry et al. 2013 [22]	5	5	10
Parveen et al. 2013 [12]: OCSVM	1	10	11
Parveen et al. 2013 [12]: GBAD	2	10	12
Brdiczka et al. 2012 [24]	7	3	10
Chen et al. 2012 [13]	2	7	9
Raissi-Dehkordi et al. 2011 [21]	4	8	12
Eberie et al. 2009 [27]	7	9	16
Tang et al. 2009 [26]	2	6	8
Yu et al. 2006 [23]	3	11	14
Sum	54	84	138

$$|R_u - R_v| \geq q(0.05, 1)\sqrt{\frac{Nk(k+1)}{6}}$$
$$\geq 1.645 \times \sqrt{\frac{14 \times 2 \times (2+1)}{6}}$$
$$\geq 6$$

Any difference which is equal to or exceeds 6 will indicate a significant difference between that pair of algorithms.

We then rank the algorithms based on the number of algorithms has been outperformed by the algorithms. The result is presented in Table 6. Currently, the algorithms proposed by Ebrie et al. [27], Yu et al. [23], Raissi-Dehkordi et al. [21] and Parveen et al. [12]: GBAD are the optimal and best-to-date algorithms based on the theoretical merits as well as the publication biases and limitations.

Conclusions

This study addresses three primary research questions in IDP:

Question 1. What are the trends and developments shaping the IDP problem?

Key findings include:

- The number of publications per year is increasing on IDP.
- GTA is used as a major source of generating data on insider attacks. Overall, GTA can be a reliable source of data for both research and commercial purposes if the games are designed correctly.

Table 6 Overall Rankings of the IDPAs based on their theoretical merits taking into account publication bias

Rank	Algorithms belong to the rank	Algorithm(s) that are outperformed by the algorithm(s) that belong to the rank group
1	Eberie et al. 2009 [27]	(1) Parveen et al. (2013) [12]: OCSVM, (2) Mayhew et al. (2015) [29], (3) Zhang et al. (2014) [18], (4) Eldardiry et al. (2013) [22], (5) Brdiczka et al. (2012) [24], (6) Chen et al. (2012) [13], (7) Ahmed et al. (2014) [14], (8) Tang et al. (2009) [26], (9) Amber et al. (2015) [15], (10) Axelrad et al. (2013) [25]
2	Yu et al. (2006) [23]	(1) Chen et al. (2012) [13], (2) Ahmed et al. (2014) [14], (3) Tang et al. (2009) [26], (4) Amber et al. (2015) [15], (5) Axelrad et al. (2013) [25]
3	# Parveen et al. (2013): GBAD [12], # Raissi-Dehkordi et al. (2011) [21]	(1) Amber et al. (2015) [15], (2) Axelrad et al. (2013) [25]
4	# Mayhew et al. (2015) [29], # Parveen et al. (2013) [12]: OCSVM, # Zhang et al. (2014) [18], # Eldardiry et al. (2013) [22], # Brdiczka et al. (2012) [24], # Chen et al. (2012) [13], # Ahmed et al. (2014) [14], # Tang et al. (2009) [26], # Amber et al. (2015) [15], # Axelrad et al. (2013) [25]	N/A

- The key features of the insider attack include the psychological symptoms and the online activities of the insiders.
- Typically, the output of an IDPA is the point estimate of the likelihood of an insider attack.
- The online activities are the most widely used features for IDP. Existing IDPAs assess threats based on the individual and group perspectives of online activity data of insiders.
- Most of the current IDPAs rely on AD approach to detect attacks ahead of time. AD uses unlabelled data for training and can discover unseen attacks.
- Graphical models/ Bayesian Networks are by far the most common tools for IDP.

Question 2. What are the challenges of IDP?

Challenges associated with IDP include unbounded patterns, uneven time lags between activities, nonstationarity, individuality, high dimensionality, interacting features, collusion attacks, high false alarm rates, acute class imbalance, undetected insider attacks, uncertainty, and the large number of free parameters in the model.

Question 3. What are the best to date IDPAs?

The algorithms proposed by Eberie et al. [27] and Yu et al. [23] are the best to date algorithms for IDP based on the theoretical merits and the transparency of these approaches.

Best practice recommendations

This review study offers important insights into promising IDPAs for the future

- A good algorithm should adopt a holistic approach for IDP comprising individual behaviour models, community behaviour models, resource usage models and intention models to make the IDPA robust against both individual and collusion attacks, the high rate of false alarms, and data nonstationarity. The algorithm proposed by Mayhew et al. [29] is an example of a holistic system.
- A good IDPA should be a BN like the algorithm proposed by Eberie et al. [27]. A BN will make the IDPA robust against unbounded patterns, uneven time lags between activities, high dimensionality and interacting features.
- The BN should employ a Mamdani FIS –like the algorithm proposed by Yu et al. [23] –for making predictions at every level of the hierarchy. This strategy will make the algorithm robust against the uncertainty.
- A good IDPA should use the AD approach so as to address the problems associated with class imbalance and undetected insider attacks.
- It will be a complex model—there is no magic solution to this problem. Good search algorithms need to be employed. This problem can be reduced by running multiple global and local search algorithms together to optimize the model parameters [48].

Authors should take the following measures to minimize the bias and to improve the transparency of publications.

- The computer code for the proposed algorithm should be made public so that other researchers can easily evaluate the algorithm.
- Detailed pseudo-codes of the proposed algorithm should be provided for the readers with limited or no knowledge about the specific programming language.
- A reference and link to research data should be made in any publications arising from it. If data cannot be made available for reuse on ethical grounds, authors can build models from data and then simulate data from the model with added noise. This data can be made public.
- The best to date algorithms should be chosen for the benchmark.

Additional file

Additional file 1: Table S1. Overview of the proposed insider threat detection algorithms. (XLSX 25.6 kb)

Abbreviations

AD, anomaly detection; BN, Bayesian networks; CBM, community behaviour model; CM, capability model; DR, detection rate; FAR, false alarm rate; FIS, fuzzy inference system; GBAD, graph-based anomaly detection; GMM, Gaussian mixture model; GS, Google Scholar; GTA, game-theoretic approach; IBM, individual behaviour model; IDP, insider threat detection and prediction; IDPA, insider threat detection and prediction algorithm; IDPAs, insider threat detection and prediction algorithms; IM, intention model; IT, information technology; KNN, K-nearest neighbour algorithm; MS, mixed system; OCSVM, one-class SVM; RUM, resource usage model; SID, signature-based intrusion detection; SVM, support vector machine; WoS, Web of Science

Acknowledgement

We would like to thank the anonymous reviewers for their valuable comments on our manuscript.

Authors' contributions

IAG (first and corresponding author) jointly conceived the study with AEA. IAG performed literature review and manuscript write-up. AEA performed critical revision of the article. Both authors read and approved the final manuscript.

Authors' information

Dr Iffat A. Gheyas: is a Lecturer in Big Data Analytics at Birmingham City University. She received her PhD in Computing Science from Stirling University in 2010 specializing in Artificial Intelligence and Machine Learning for Data Mining and Knowledge Discovery. From 2011 to 2013, she served as a research fellow on the Cloud Stewardship Economics project (funded by Technology Strategy Board) at the University of Aberdeen. Within this project, she developed a cyber-security monitoring tool. It analyses network traffic behaviour to detect security attacks. From 2013 to 2014, she served as a research fellow on the European Commission FP7 project "Security Economics", initially at the University of Aberdeen, and then at Durham University. As part of this project, she developed risk mitigation algorithms for addressing the dynamic threats to the National Grid System through resilience. She joined Birmingham City University in August 2014 as a lecturer.
Professor Ali E. Abdallah: Ali E. Abdallah BSc, MSc, DPhil(Oxon): Ali E. Abdallah is a professor of information security at Birmingham City University. Prior to current appointment, he was a professor of information assurance at London South Bank University, a lecturer in computer science at the University of Reading and a research officer at Oxford University Computing Laboratory. He lectures in information security, information risk management, software security, distributed systems and formal methods. His research interests are closely linked to his teaching and strongly emphasize the relevant theoretical underpinnings. He leads "information security" research at BCU focusing on topics ranging from identity management, access control and privacy to securing shared information in virtual organizations and the development of high assurance secure and resilient software.

Competing interests

The authors declare that they have no competing interests.

Received: 13 October 2015 Accepted: 4 August 2016
Published online: 30 August 2016

References

1. BBC News. HIV patient tells of fears of disclosure after details leaks. 2015. http://www.bbc.co.uk/news/uk-34135866. Accessed 10 Sept 2015.
2. Ashford W. Barclays bank leaks thousands of customer records. 2014. http://www.computerweekly.com/news/2240214060/Barclays-under-scrutiny-after-leak-of-27000-customer-records. Accessed 10 Sept 2015.
3. Fox News. Pepsi alerted Coca-Cola to stolen-coke-secrets offer. 2006. http://www.foxnews.com/story/2006/07/06/pepsi-alerted-coca-cola-to-stolen-coke-secrets-offer.html. Accessed 10 Sept 2015.
4. Theguardian. PlayStation Network users fear identity theft after major data leak. 2011. http://www.theguardian.com/technology/2011/apr/27/playstation-users-identity-theft-data-leak. Accessed 10 Sept 2015.
5. OPSWAT. The top six corporate email risks. 2015. http://www.policypatrol.com/the-top-six-corporate-email-risks/. Accessed 10 Sept 2015.

6. Vormetric Data Security. 2015 vormetric insider threat report. 2015. http://enterprise-encryption.vormetric.com/rs/vormetric/images/CW_GlobalReport_2015_Insider_threat_Vormetric_Single_Pages_010915.pdf. Accessed 10 Sept 2015.
7. Hong J, Kim J, Cho J. The trend of the security research for the insider cyber threat. Int J Future Gener Commun Netw. 2010;3(2):31–40.
8. Ophoff J, Jensen A, Sanderson-Smith J, Porter M, Johnston K. A descriptive literature review and classification of insider threat research. In: Proceedings of Informing Science & IT Education Conference (InSITE). 2014. p. 211–23.
9. Azaria A, Richardson A, Kraus S, Subrahmanian VS. Behavioral analysis of insider threat: a survey and bootstrapped prediction in imbalanced data. IEEE Trans Comput Soc Syst. 2014;1(2):135–55.
10. Moher D, Liberati A, Tetziaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6:7.
11. Siegel S, Castellan Jr NJ. Nonparametric Statistics for the behavioural sciences. 2nd ed. New York: McGraw Hill; 1988.
12. Parveen P, Mcdaniel N, Weger Z, Evans J, Thuraisingham B, Hamlen K, Khan L. Evolving insider threat detection stream mining perspective. Int J Artif Intell Tools. 2013;22(5):1–24.
13. Chen Y, Nyemba S, Malin B. Detecting anomalous insiders in collaborative information systems. IEEE Trans Dependable Secure Comput. 2012;9(3):332–44.
14. Ahmed MB, Akram A, Asif M, Ur-Rehman S. Using genetic algorithm to minimize false alarms in insider threat detection of information misuse in windows environment. Math Probl Eng. 2014;2014(2014):1–11.
15. Ambre A, Shekokar N. Insider threat detection using log analysis and event correlation. Proc Comp Sci. 2015; 45(2015):436–45.
16. Liu D, Wang X, Camp J. Game-theoretic modelling and analysis of insider threats. Int J Crit Infrastruct Prot. 2008; 2008:75–80.
17. Legg PA, Buckley O, Goldsmith M, Creese S. Automated insider threat detection system using user and role-based profile assessment. IEEE Syst J, PP. 2015;99:1–10.
18. Zhang R, Chen X, Shi J, Xu F, Pu Y. Detecting insider threat based on document access behaviour analysis. In: Han W, editor. LNCS, 8710. Switzerland: Springer International Publishing; 2014. p. 376–87.
19. Yaseen Q, Panda B. Enhanced insider threat detection model that increases data availability. In: Natarajan R, Ojo A, editors. LNCS, 6536. Heidelberg: Springer; 2011. p. 267–77.
20. Wang H, Liu S, Zhang X. A prediction model of insider threat based on multi-agent. In: 2006 1st International Symposium on Pervasive Computing and Applications. Urumqi: IEEE; 2006. p. 273–8.
21. Raissi-Dehkordi M, Carr D. A multi-perspective approach to insider threat detection. In: The 2011 Military Communications Conference – Track 3 – Cyber Security and Network Operations. Baltimore: IEEE; 2011. p. 1164–9.
22. Eldardiry H, Bart E, Liu J, Hanley J, Price B, Brdiczka O. Multi-domain information fusion for insider threat detection. In: 2013 IEEE Security and Privacy Workshops. San Francisco: IEEE; 2013. p. 45–51.
23. Yu Y, Graham JH. Anomaly instruction detection of Masqueraders and threat evaluation using fuzzy logic. In: 2006 IEEE International Conference on Systems, Man, and Cybernatics. Taipei: IEEE; 2006. p. 8–11.
24. Brdiczka O, Liu J, Price B, Shen J, Patil A, Chow R, Bart E, Ducheneaut. Proactive insider threat detection through graph learning and psychological context. In: IEEE Symposium on Security and Privacy Workshops (SPW). San Francisco: IEEE; 2012. p. 142–9.
25. Axelrad ET, Sticha PJ, Brdiczka O, Shen J. A Bayesian network model for predicting insider threats. In: 2013 IEEE Security and Privacy Workshops. San Francisco: IEEE; 2013. p. 82–9.
26. Tang K, Zhou M-T, Wang W-Y. Insider cyber threat situational awareness framework using dynamic Bayesian network. In: 4th International Conference on Computer Science & Education. Nanning: IEEE; 2009. p. 1146–50.
27. Eberie W, Holder L. Applying graph-based anomaly detection approaches to the discovery of insider threats. In: IEEE International Conference on Intelligence and Security Informatics, 2009. ISI'09. Dallas: IEEE; 2009. p. 206–8.
28. Yaseen Q, Panda B. Knowledge acquisition and insider threat prediction in relational database systems. In: 2009 International Conference on Computational Science and Engineering. Vancouver: IEEE; 2009. p. 450–5.
29. Mayhew M, Atighetchi M, Adler A, Greenstadt R. Use of machine learning in big data analytics for insider threat detection. In: 2015 IEEE Military Communications Conference, MILCOM 2015. Tampa: IEEE; 2015. p. 915–22.
30. Young WT, Goldberg HG, Memory A, Sartain JF, Senator TE. Use of domain knowledge to detect insider threats in computer activities. In: 2013 IEEE Security and Privacy Workshops. San Francisco: IEEE; 2013. p. 90–7.
31. Wang H, Xu H, Lu B, Shen Z. Research on security architecture for defending insider threat. In: 2009 Fifth International Conference on Information Assurance and Security. Xi'an: IEEE; 2009. p. 30–3.
32. Yaseen Q, Panda B. Tackling insider threat in cloud relational databases. In: 2012 IEEE ACM Fifth International Conference on Utility and Cloud Computing. Chicago: IEEE; 2013. p. 215–8.
33. Magklaras GB, Furnell SM. A preliminary model of end user sophistication for insider threat detection in IT systems. Comput Secur. 2005;24:371–80.
34. Taylor PJ, Dando CJ, Ormerod TC, Ball LJ, Jenkins MC, Sandham A, Menacere T. Detecting insider threats through language change. Law Hum Behav. 2013;37(4):207–75.
35. Martinez-Moyano IJ, Rich E, Conrad S, Andersen DF, Stewart TR. A behavioural theory of insider-threat risks: a system dynamics approach. ACM Trans Modeling Comput Simul. 2008;18(2):2–27.
36. Schultz EE. A framework for understanding and predicting insider attacks. Comput Secur. 2002;21(6):526–31.
37. Magklaras GB, Furnell SM. Insider threat prediction tool: evaluating the probability of IT misuse. Comput Secur. 2002;21(1):62–73.
38. Kandias M, Galbogini K, Mitrou L, Gritzalis D. Insiders trapped in the mirror reveal themselves in social media. In: Lopez J, Huang X, Sandhu R, editors. LNCS, 7833. Heidelberg: Springer; 2013. p. 220–35.
39. Kandias M, Mylonas A, Virvilis N, Theoharidou M, Gritzalis D. An insider threat prediction model. In: Katsikas S, Lopez J, Soriano M, editors. LNCS, 6264. Heidelberg: Springer; 2010. p. 26–37.
40. Kandias M, Stavrou V, Bozovic N, Mitrou L, Gritzalis D. Can we trust this user? Predicting insider's attitude via You Tube usage profiling. In: 2013 IEEE 10th International Conference on Ubiquitous Intelligence & Computing and 2013 IEEE 10th International Conference on Autonomic & Trusted Computing. Vietri Sul Mere: IEEE; 2013. p. 347–54.

41. Myers J, Grimaila MR, Mills RF. Towards insider threat detection using web server logs. In: CSIIRW'09 Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies. New York: ACM; 2009.
42. Eberie W, Holder L. Mining for insider threats in business transactions and processes. In: IEEE Symposium on Computational Intelligence and Data Mining, 2009, CIDM'09. Nashville: IEEE; 2009. p. 163–70.
43. Greitzer FL, Ferryman TA. Methods and metrics for evaluating analytic insider threat tools. In: 2013 IEEE Security and Privacy Workshops. San Francisco: IEEE; 2013. p. 90–7.
44. Nance K, Marty R. Identifying and visualizing the malicious insider threat using Bipartite graphs. In: Proceedings of the 44th Hawaii International conference on System Sciences. IEEE: Kauai; 2011. p. 1–9.
45. Gritzalis D, Stavrou V, Kandias M, Stergopoulos G. Insider threat: enhancing BPM through social media. In: 2014 6th International Conference on New Technologies, Mobility and security (NTMS). Dubai: IEEE; 2014. p. 1–6.
46. Glasser J, Lindauer B. Bridging the gap: a pragmatic approach to generating insider threat data. In: 2013 IEEE Security and Privacy Workshops. San Francisco: IEEE; 2013. p. 98–104.
47. Claycomb WR, Nicoll A. Insider threats to cloud computing: directions for new research challenges. In: 2012 IEEE 36th International Conference on Computer Software and Applications. Washington, DC: IEEE; 2012. p. 387–94.
48. Gheyas IA, Smith LS. Feature subset selection in large dimensionality domains. Pattern Recogn. 2010;43:5–13.