Proactive Insider Threat Detection Through Social Media: The YouTube Case

Miltiadis Kandias, Vasilis Stavrou, Nick Bozovic, Dimitris Gritzalis
Information Security & Critical Infrastructure Protection Research Laboratory
Dept. of Informatics, Athens University of Economics & Business (AUEB)
76 Patission Ave., GR-10434, Athens, Greece
{kandiasm, stavrouv, nbozovic, dgrit} @aueb.gr

ABSTRACT
Insider threat is a major issue in cyber and corporate security. In this paper we study the psychosocial perspective of the insider via social media, Open Source Intelligence, and user generated content classification. Inductively, we propose a prediction method by evaluating the predisposition towards law enforcement and authorities, a personal psychosocial trait closely connected to the manifestation of malevolent insiders. We propose a methodology to detect users holding a negative attitude towards authorities. For doing so we facilitate the use of machine learning techniques and of a dictionary-based approach, so as to detect comments expressing negative attitude. Thus, we can draw conclusions over a user behavior and beliefs via the content the user generated within the limits a social medium. We also use an assumption free flat data representation technique in order to decide over the user's attitude. Furthermore, we compare the results of each method and highlight the common behavior manifested by the users. The demonstration is applied on a crawled community of users on YouTube.

Categories and Subject Descriptors
K.4.1 [**Computers and Society**]: Public Policy Issues – *Abuse and crime involving computers, privacy*.

General Terms
Security, Human Factors.

Keywords
Insider Threat, Social Media, YouTube, Security Officer, Privacy, Behavior Prediction.

1. INTRODUCTION
Threats that an information system may encounter derive from either external or internal environments. In order to mitigate such threats, information security officers and researchers are often asked to identify the optimised analogy between security and functionality. One of the most demanding problems in cyber and corporate security is the insider threat [1]. The malevolent insider manifests when a trusted user of the information system behaves in a way that the security policy defines as unacceptable [2].

WPES'13, November 4, 2013, Berlin, Germany.

Regardless of the numerous technical countermeasures, techniques and methods proposed, insider computer abuse incidents keep occurring. As a result, research suggested [3] that technical and social solutions should be implemented so as to reduce the impact of this threat. Social learning theory [4] assumes that a person commits a crime because she has come to associate with delinquent peers. Similar approaches study the criminal computer behavior and examine the deviant computer-related behavior [5]. Furthermore, behavior analysis leads to studying employees under the prism of predisposition towards malevolent behavior, by examining personal traits that have been proved to abut to this kind of behavior. Shaw's et al. [6] research examined the trait of social and personal frustrations. The most important observation is the "revenge syndrome" such persons develop and the anger they feel towards authority figures. Thus, an employee negatively predisposed towards law enforcement and authorities, is considered to be more possible to manifest delinquent behavior against the organization. Furthermore, research has proved that individuals tend to transfer their offline behavior online [7], thus making it possible to perform psychometric evaluations by utilizing the content a user has made publicly available.

In this paper we utilize the information that users share in social media, in order to propose prediction and deterrence measures against the insider threat. Our goal is to extract conclusions over the users, regarding the personality trait of predisposition against law enforcement and authorities, which is a common characteristic among insiders. Furthermore, we propose a methodology that can extract a user's attitude towards that trait, along with two complementary approaches, which are used as a means of comparison of the results and the development of our core method.

To this extend, we use data crawled by YouTube to apply our methodology on a realistic environment. We formed a community of YouTube users (consisting of Greek users) and classified them via (a) comment classification using two approaches; namely, machine learning techniques and dictionary-based classification on user comments, and (b) comment classification via flat data in an assumption free basis, regarding the data mining process. In addition, users are divided into two categories; those who are predisposed negatively towards law enforcement and authorities (Category P), and those who are not (Category N). The user attitude is extracted by aggregating the individual results from the attitude expressed in comments, uploads, favorites and playlists, in the comment classification approach and by examining each user as a flat file, in the flat data classification.

2. RELATED WORK
Dealing with insider threat incidents is one of the most important challenges faced by today's organizational, industrial, and other forms of information infrastructures. To this end, researchers have proposed numerous countermeasures tackling this threat. Such countermeasures include, among others, the development of an in-

formation security common body of knowledge, in order to develop an information security curriculum [8]. Furthermore, the field of risk assessment/management and critical infrastructure protection has contributed towards an elevated understanding over the issue [9-10]. Alternative approaches have, also, been proposed to this extend [11-12].

The area of insider threat prediction involves various methods and techniques [13]. Magklaras et al. [13] introduced a threat evaluation system based on certain profiles of user behavior. Furthermore, he presented a process of constructing a language tailored to describing insider threat incidents, so as to mitigate threats derived from legitimate users in an IT infrastructure [14].

Kandias et al. [15] proposed a combination of technical and psychological approaches to deal with insider threats, while Greitzer et al. and Brdiczka et al. also take into consideration the psychosocial perspective of an insider. Greitzer et al. [16] developed a psychosocial model to assess employees' behavior, while Brdiczka et al. [17] proposed an approach that combines Structural Anomaly Detection from social and information networks and Psychological Profiling of individuals so as to identify threats.

Personal factors that may increase the likelihood someone to develop malevolent behavior are presented by the FBI, too [18]. Personal traits such as anger, revenge or greed along with certain circumstances presented in an organization could lead to the manifestation of an insider. An approach of studying personality traits, described by Shaw, has been introduced by studying the trait of narcissism using Graph Theoretic Analysis [19].

3. TESTING ENVIROMENT

In our research, we focus on a Greek community of YouTube and on information gathered from our previous work [20]. In comparison with our previous work, this paper builds upon our previous research and poses a significant extension.

At first, the subject of this paper is the detection of users who are negatively predisposed towards law enforcement and authorities. Secondarily, in this paper we utilize the dataset and an extension of the methodology proposed in our previous work, so as to perform usage profiling and insider threat prediction. The previous work focused on user awareness and the possibility of political beliefs extraction. Finally, in this paper we have extended the use of machine learning techniques by using flat data classification.

3.1 Data Crawling

We utilized the dataset we crawled during our previous work. Dataset collection was performed using YouTube's REST-based API (https://developers.google.com/youtube/v3/).The dataset includes: (a) 12.964 users, (b) 207.377 videos, and (c) 2.043.362 comments. The time span of the collected data covers a period of 7 years (Nov. 2005 - Oct. 2012). In addition, data was classified into three categories: (a) user-related information, e.g., profile, uploaded videos, subscriptions, favorite videos, playlists, (b) video-related information, e.g., video's license, number of likes and dislikes received, category and tags, and (c) comment-related information, e.g., the content of the comment and the number of likes and dislikes it received.

4. PROPOSED APPROACH

In this section we propose a method that enables the identification of a user's attitude towards law enforcement and authority. Social media offer us the ability to monitor users' online behaviour, record their online life imprint, and identify this attitude, which is expressed via their videos, comments and likes. Video is YouTube basic module. As we cannot process the video itself, we draw a conclusion for each video via its comments.

We detect the attitude towards law enforcement expressed in a comment by performing text classification into two categories: (a) category P, which contains expressions holding negative attitude towards law enforcement, and (b) category N, which contains all the comments that hold a neutral attitude towards law enforcement and authority, or have no such content. Category N may also contain expressions holding positive attitude towards law enforcement. However, in this paper we are interested only in the prevalence or absence of negative attitude towards law enforcement.

Text classification uses machine learning techniques to classify a comment into the appropriate category. A comment is assigned to one of the two categories, so as to indicate the respective attitude towards law enforcement that the specific category depicts. We further examined the efficiency of another technique. Thus, we formed a dictionary that included a series of words and phrases that indicate negative attitude towards law enforcement. These words and phrases belong to various Greek jargons that refer to authorities using derogatory vocabulary and expressions. We scan each comment to detect specific terms and assign it to the appropriate category. According to our findings, machine learning leads to a more reliable result than a simple word check and also, text classification performs better than scanning lists of words in a dictionary. Comment classification using machine learning techniques performed 35% faster, than using a term dictionary.

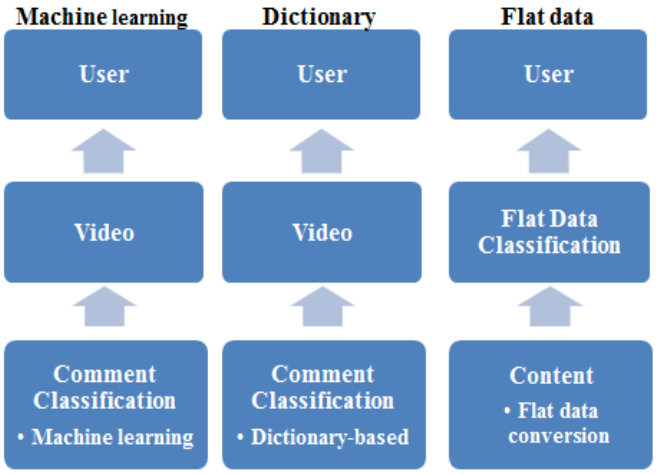

Figure 1. Classification approaches

Comment classification enables the extraction of conclusions over the attitude expressed in a video's comments. This approach facilitates classification of videos and further categorization into one of the predefined categories (P for those with negative attitude towards law enforcement and authorities and N for the others). So, assignment of a comment into a corresponding category implies the attitude of the user who uploaded it, commented it, or liked it, towards law enforcement. The same applies to a list of videos, e.g., favorite videos and playlists. By acquiring the category a video falls into, a conclusion can be drawn for the attitude expressed in the list. Being able to classify a user's content, we may extract conclusions for user's comments, uploaded videos, favorite videos or playlists. By aggregating these results we can elicit adequate information to draw a conclusion over the user's overall attitude.

4.1 Comment Classification

In order to further analyze the collected dataset, we have chosen to store it in a relational database. We classify comments into the

predefined categories using two techniques, i.e. (a) machine learning, and (b) dictionary-based detection.

4.1.1 Machine Learning Approach

In order to classify comments into one of the predefined categories of attitude towards law enforcement, training of an appropriate classifier is required. Comment classification is performed as text classification [21] that uses machine learning techniques to train the system and decide the category which a text falls into. The machine is trained using input text examples, along with the corresponding categories to which they belong. The machine's output is a prediction of the label that is assigned to a text seen for the first time. Machine learning algorithms "learn" from the text examples they receive and construct underlying models that are able to determine the label of any text given as input. Label assignment requires the assistance of an expert, who can distinguish and justify the categories each text belongs to.

Forming the training sets requires two processes: The first is comment selection from the database and the second is proper label assignment to comments according to the category they belong. The later categorization was supported by a domain expert (i.e., Sociologist), who could assign and justify the chosen labels on the training sets. Thus, we developed a reliable classification mechanism. We chose 430 comments from category P and 470 from category N of the training set for each language. The expert contributed by assigning a category label to each comment.

Apart from the training set, we also created a test set, which is required to evaluate the efficiency of the resulting classifier. The test set consists of pre-labeled data that are fed to the machine to check if the initial assigned label of each comment is equal to the one predicted by the classification algorithms. The test set labels were also assigned by the domain expert.

We performed comment classification using: (a) Naïve Bayes Multinomial (NBM), (b) Support Vector Machines (SVM), and (c) Logistic Regression (LR), so as to compare the results and pick the most efficient classifier. We compared each classifier's efficiency based on the metrics of precision, recall, f-score, and accuracy [22]. Accuracy metric measures the number of correct classifications performed by the classifier. Precision measures the classifier's exactness. Higher or lower precision score implies few or many false positive classifications respectively. Recall measures the classifier's completeness. Higher or lower recall score means few or many false negative classifications, respectively. F-score is the harmonic mean of both metrics. Table 1 presents each classifier's efficiency, based on accuracy, precision, recall, and f-score. LR and SVM achieve the highest accuracy. Precision and recall are proper metrics to evaluate each classifier [22].

Table 1 Metrics comparison of classification algorithms

	Metrics					
Classifier	**NBM**		**SVM**		**LR**	
Classes	**P**	**N**	**P**	**N**	**P**	**N**
Precision	71	70	83	77	86	76
Recall	72	68	75	82	74	88
F-Score	71	69	79	79.5	80	81
Accuracy	70		80		81	

Logistic Regression achieves better precision value for category P and recall value for category N. SVM achieves slightly better precision value for category N and recall value for category N. Logistic Regression achieves a slightly better f-score assessment.

SVM and LR achieve similar results regarding all metrics, so we chose Logistic Regression because of the better f-score value achieved for each category. By choosing an algorithm based solely on the recall or precision metrics, we would favor one metric at the expense of another. Consequently we had to choose between fewer false positive and fewer false negative classifications. A better f-score assessment indicates a balanced combination of false positive/negative classifications.

Another reason why we chose Logistic Regression is that in Support Vector Machines the result of the classification is purely dichotomous and no probability of category membership is given [23], whereas Logistic Regression algorithm outputs the probability of each category membership. We utilized the probability of Logistic Regression to determine a threshold above which a comment is classified into Category P. We chose to apply a threshold on the comments characterized by negative attitude towards law enforcement, in order to reduce false positive classifications and avoid classifying a comment into category P without strong confidence. We did not apply any threshold to Category N, as our target of interest is category P. The threshold we determined is 72%, i.e., if a comment belongs to category P and its probability of category membership is lower than 72%, it will be classified as non-negative. We tested several threshold values, while observing whether the false positive rate of category P would improve, based on our test set. After utilizing all information provided by the Social Scientist involved in the above mentioned procedure, we verified that the set of comments with probability of category membership lesser than 72% included enough false positive classifications.

4.1.2 Dictionary-Based Approach

An alternative approach we implemented was to classify comments using a dictionary. Comments classification was conducted with the help of a field expert (Sociologist), so as to form a dictionary containing terms and phrases, part of which belong to various Greek jargons, expressing negative attitude towards law enforcement and authorities.

The dictionary contains approximately 65 terms, referring to issues related to law enforcement, authorities, etc. Thus, if a comment contains a term of the dictionary, it is assigned to the P-category. This means that the comment holds a negative attitude towards law enforcement. We only check if the comment belongs to the category P, otherwise it is categorized as non-negative. A disadvantage found in such approaches is the low resulting recall value. On the other hand, they manage to achieve a high precision value.

4.2 Video Classification

For determining the attitude expressed in a video, we analyze its comments, thus we are able to extract its viewer's attitude instead of the content of the video per se. If the video contains comments with negative attitude towards law enforcement the video falls into category P. If the video does not contain any comment expressing negative attitude towards law enforcement, it is classified as non-negative.

Users' understanding of the video is more important in this case than the content of the video, itself. This is so because we aim at classifying users instead of videos. During video classification we tried to eliminate comments that had no likes and more than one dislikes. We chose this approach so as to avoid classifying a video into category P due to a comment with no acceptance from other

users. If a comment receives only dislikes and no likes, then it is possible that the content of the video has nothing to do with negative attitude towards law enforcement and should not be calculated in the video classification process.

4.3 List Classification

The procedure followed to extract a conclusion about a list of videos is similar to the formerly mentioned video conclusion extraction method. The difference is that we utilize videos instead of comments. So, if the list contains videos that belong to the category P, the list is also classified as possessing negative attitude towards law enforcement. The list classification method applies to the user's uploaded videos, favourite videos and her playlists. We followed a similar approach to the one we used in video classification. We eliminated any videos from the list that had only dislikes and no likes (as explained in the comment classification case).

4.4 User Classification

For drawing a conclusion over the user's overall attitude towards law enforcement, we examined the user according to an aggregation of the partial conclusions over her comments, uploads, favourites and playlists classifications utilizing a voting system. Based on these observations, one may look for indications of the attitude towards law enforcement within the user generated content. A user is classified to category P, if there are comments or videos detected to her content that contain negative attitude towards law enforcement. A user may not express any attitude towards law enforcement via her comments or uploaded videos; however, the user may have added videos with similar content to playlists. It is likely to detect users who have added only a limited amount of content expressing negative attitude towards law enforcement. Similar content may also be detected in the favourite videos or playlists. In these cases, it is possible that the user is not predisposed negatively towards law enforcement and the authorities. However, it may be an indication that further examination is needed in order to decide whether the user shares this psychosocial characteristic or not. Inductively, the proposed method could be useful as a median in the hands of an experienced field expert who is able to weight the results and extrapolate the conclusions of our method within the limits of a critical infrastructure.

5. FLAT DATA CLASSIFICATION

The method was designed to address the same problem from a slightly different perspective. The aim was to design a method that was assumption free and could scale well. For that we chose a flat representation of the data, something that can ease their partitioning and distributed processing. We also chose a method in which we make no assumption for the users, the videos or their comments. The only input of the system is a list of negatively and non-negatively predisposed users that are used to train our system. This enables us to use the system irrespective of language or other barriers as long as we provide the system with a good initial pool of identified users, as they were indicated by the field expert. To evaluate our system's performance we conducted extensive experiments using the same dataset as in the evaluation of the above mentioned method (Section 4.1.1), and compared the results.

5.1 Data Transformation

The data collected and used for the experiments were stored in a relational database, so testing the alternative method required decoupling the data in a flat structure. The data where represented in one relation consisting of user information bundled with each individual comment. This flat approach was chosen because it is significantly different from our previously proposed methods thus enabling us to prove or disprove the correctness of our method. This method can also facilitate the partitioning and distributed processing of the data. Also, since this method is capable of evaluating a single post/user tuple at a time, it is a better starting point for a real-time warning system.

The focus of the method is on the comment content, as that is augmented with user data as collected from the crawler. The flat relation consists of the following attributes: username, comment's text content, unique video identifier that the comment refers to user's country, age, genre, and number of subscribers and video views. Lastly, there is an additional field filled with text information that the user has chosen to publicly share on her profile.

5.2 Methodology

After formatting the data, we separated the comments into two classes based on the writer's class as evaluated by a human expert. This is a method that uses no assumptions on our part, except of the knowledge and skill of a human expert. So, by using this 44.000 strong, two class, training set we trained a simple Bayesian classifier. The resulting predictions for each user/comment tuple were combined trough a simple mechanism, to provide insight into the behavior of the individual user. More specifically, since the classification produces results for user/comment tuples, it was important to find a way to use that information for marking the interesting users as such.

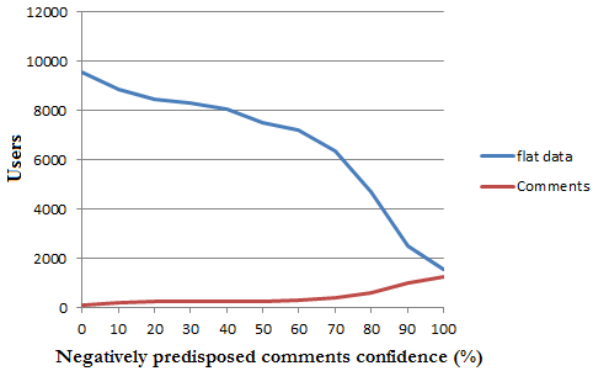

Figure 2 Relation between detected users and detected comments percentage

In contrast to the method used before and in order to avoid marking a user over a single post (that may, or may not, be a true positive), we used a simple metric to overcome this case. We opted to use the division of the number of *P* comments a user has made by his total number of comments. This method can help us differentiate the users in an efficient and resource cheap way. This quotient acts as a threshold for identifying users that warrant closer inspection. In our case, the number of users flagged (blue line) in relation to the chosen threshold is summarized in Figure 2, which also depicts the number of predisposed users (red line) detected by the above mentioned machine learning-based approach (Section 4.1.1). For each threshold depicted in Figure 2, we present the number of users detected by both methods. As the confidence increases, the fractions of users detected by both methods begin to equalize. After manual survey of the detected users, we found out that both sets (flat data and comment classification) contain the same users. Especially in the case of 100% negatively predisposed comments percentage, the sets were almost identical.

The results of the evaluation of the alternate system, that used the specified training set, are encouraging. A ten-fold cross validation showed a precision of 73% and a recall of 93% on the user/com-

ment relation. Detailed results and an overview of the system's performance with different classifiers are presented in Table 2.

Table 2 Flat data classification algorithm metrics comparison

	Naïve Bayes Metrics			
Classes	**Precision**	**Recall**	**F-score**	**Accuracy**
P	72	92	81	81
N	93	73	82	

6. RESULTS PRESENTATION

According to the above mentioned methods, we are able to extract a series of results regarding the crawled YouTube users. The core decision of these processes is, however, the predisposition of the user towards law enforcement and authorities.

6.1 Methods Comparison

Here we present the results obtained for each classification approach utilized (i.e. Logistic Regression as LR, Dictionary approach as Terms). Table 3 depicts the number of users per attribute who have been classified as expressing negative attitude towards law enforcement and authorities. In line with the flat data thresholds, and depending on the scenario applied, the security officer may decide which of the above mentioned attributes could be processed, so as to reveal a user attitude.

Table 3. Machine Learning vs. Dictionary approach

	Category P Users/Classification Attribute			
Method	**Comments**	**Uploads**	**Favourite**	**Playlists**
LR	1400	838	2300	2200
Terms	1070	516	1900	1500

Figure 2 depicts the common users identified by the flat data and the machine learning-based approach. For each threshold one may notice the amount of users detected by flat data classification, as well as the number of common users who have been classified as negatively predisposed based solely on their comments. Table 4 depicts the number of comments, videos, and users that have been classified as predisposed negatively towards law enforcement.

Table 4. Machine Learning vs. Dictionary Classifications

	Category P Classifications		
Method	**Comments**	**Videos**	**Users**
LR	11986	7195	4850
Terms	11389	4784	3845

The deviation that appears between the results of logistic regression and the dictionary-based approach indicate that logistic regression has detected correlations from the training set that simple word-checking could not. The detected deviation could also be a result of false positive or false negative classifications. However, the false positive/negative classifications are rather low for both categories according to metrics appearing inTable 1.

The 1400 users, who were detected having posted negatively predisposed comments, are verified by the flat data approach. The threshold of 72%, that we set in the machine learning-based methodology, above which a comment is classified as predisposed negatively, can be enhanced by the Figure 2. As one may notice, the threshold above which both lines start to converge is above 70%, with the blue line dropping increasingly. At the threshold of 100%, the fraction of users detected by the machine learning-based approach is 10% lower than the flat data one. This variation is explained by the fact that the two approaches are based on different algorithms (the core method is based on Logistic Regression and flat data classification is based on Naïve Bayes).

7. CONCLUSION

In this paper we deal with the insider threat prediction issue. Malevolent insiders and predisposition towards computer crime has been closely linked to the psychosocial trait of negative attitude towards law enforcement and authorities. In specific, we proposed a detection method for users holding this trait towards law enforcement, based on comment classification via machine learning in YouTube social medium.

The desirable results are achieved by classifying each video's comments into the predefined categories (i.e. Category P and N for negative and non-negative attitude, respectively). Thus, we were able to extract a conclusion over the user's attitude towards law enforcement and authorities as expressed in the content by aggregating partial classification attributes (comments, uploads, favorites, playlists). We also presented two additional methods for user detection, along with the one described above; i.e. a dictionary-based comment classification and an assumption free flat data comment/user classification. The former approach is similar to the first method, albeit comment classification is performed via a dictionary containing terms and expressions holding negative attitude towards law enforcement, while the rest of the procedure remains the same. The latter utilizes a sum of nine user parameters aggregated in a single tuple for each user.

Then, we trained a simple Bayesian classifier who proved to have similar results to the core machine learning technique, described in this paper. Thus, using the flat data classification results, we compared the users our machine learning-based approach indicated as negatively predisposed via their comments. Comparison indicated similar results with small deviation. Though, the results of these methods demonstrate the robustness and adequate performance of the proposed.

Taking into consideration the above mentioned observations, one could argue over the applicability of the proposed method. Thus, the dialectic over the issue should follow the road of selecting the appropriate field of application for these profiling techniques, similarly to other proactive traditional detection techniques, even of differentiated scope [25-31]. Under a legal and an ethical point of view [24], it should be clear that such checks could be only implemented under strict terms and conditions. For example, they could be used (a) in a pre-final recruitment stage (for shortlisted candidates, with a compromised possibility of subject identifications), (b) when users have given their explicit consent, or (c) with the use of a trusted third party as an arbiter who can order the background check of selected candidates. Therefore, these checks could be used to assist recruitment of experts by critical infrastructures, especially on key roles (e.g., security officers, etc.).

For future work we plan on improving our classification methods and extract further demographics and statistics over the results. In addition, we plan on applying meta-training techniques in our methodology, so as to detect common behavioral patterns among the users and determine a threshold correlation between negatively predisposed users and their negative comments. Also, we plan on further examining the importance of each decision making attribute used in the machine learning-based methodology using the assumption free flat data classification results.

Acknowledgment

This work has been co-funded by the European Union (European Social Fund, ESF) and Greek national funds, through the Operational Program Education and Lifelong Learning of the National Strategic Reference Framework (Program Heraclitus II: Investing in Knowledge Society through the ESF).

8. REFERENCES

[1] Randazzo, R., Keeney, M., Kowalski, E., and Cappelli, D., Moore, A. 2005. *Insider threat study: Illicit cyber activity in the banking and finance sector*.CMU/SEI-2004-TR-021.

[2] Theoharidou, M., Kokolakis, S., Karyda, and M., Kiountouzis, E. 2005. The insider threat to information systems and the effectiveness of ISO17799. *Computers & Security*. 24, 6, 472-484.

[3] Lee, J., and Lee, Y. 2002. A holistic model of computer abuse within organizations. *Information Management & Computer Security.* 10, 2, 57-63.

[4] Akers, R., Krohn, M. D., Lanza, L., and Radosevich, M. 1979. Social learning and deviant behavior: A specific test of a general theory. *American Sociological Review*, 636-655.

[5] Rogers, M., Smoak, N., and Liu, J. 2006. Self-reported deviant computer behavior: A big-5, moral choice, and manipulative exploitive behavior analysis. *Deviant Behavior*. 27, 3, 245-268.

[6] Shaw, E., Ruby, K., and Post, J. 1998. The insider threat to information systems: The psychology of the dangerous insider. *Security Awareness Bulletin*. 2, 98, 1-10.

[7] Amichai-Hamburger, and Y., Vinitzky, G. 2010. Social network use and personality. *Computers in Human Behavior*, 26, 6, 1289-1295.

[8] Gritzalis, D., Theoharidou, M., and Kalimeri, E. 2005. Towards an interdisciplinary InfoSec education model. *Proc. of the 4thIFIP World Conference on Information Security Education*, Moscow Engineering Physics Institute, 22-35.

[9] Theoharidou, M., Kotzanikolaou, P., and Gritzalis, D. 2011. Risk assessment methodology for interdependent critical infrastructures. *International Journal of Risk Assessment and Management*. 15, 2, 128-148.

[10] Coles-Kemp, L., and Theoharidou, M. 2010. Insider threat and information security management. *Insider Threats in Cyber Security*. Springer, 45-71.

[11] Kandias, M., Virvilis, N., and Gritzalis, D. 2011. The insider threat in Cloud computing. *Proc. of the 6thInternational Conference on Critical Infrastructure Security*. Springer. 95-106.

[12] Kandias, M., Mylonas, A., Theoharidou, M., and Gritzalis, D. 2011. Exploitation of auctions for outsourcing security-critical projects. *Proc. of the 16th IEEE Symposium on Computers and Communications*. IEEE, 646-651.

[13] Magklaras, G., and Furnell, S. 2001. Insider threat prediction tool: Evaluating the probability of IT misuse. *Computers and Security*. 21, 1, 62-73.

[14] Magklaras, G., Furnell, S., and Brooke, P. 2006. Towards an insider threat prediction specification language. *Information Management & Computer Security*. 14, 4, 361-381.

[15] Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., and Gritzalis, D. 2010. An insider threat prediction model. *Proc. of 3rd International Conference on Trust, Privacy and Security in Digital Business*. Springer, 26-37.

[16] Greitzer, F., Kangas, L., Noonan, C., Dalton, A., and Hohimer, R. 2012. Identifying at-risk employees: Modeling psychosocial precursors of potential insider threats. *Proc. of the 45thHawaii International Conference on System Science*. IEEE, 2392-2401.

[17] Brdiczka, O., Liu, J., Price, B., Shen, J., Patil, A., Chow, R., and Ducheneaut, N. 2012. Proactive insider threat detection through graph learning and psychological context. *Proc.* of *IEEE Symposium on Security and Privacy*. IEEE, 142-149.

[18] FBI, 2012. *The Insider Threat: An introduction to detecting and deterring an insider spy*. http://www.fbi.gov/about-us/investigate/counterintelligence/ the-insider-threat

[19] Kandias, M., Galbogini, K., Mitrou, L., and Gritzalis, D. 2013. Insiders trapped in the mirror reveal themselves in social media. *Proc. of the 7thInternational Conference on Network and System Security*. Springer, 220-235.

[20] Kandias, M., Mitrou, L., Stavrou, V., and Gritzalis, D., 2013. Which side are you on? A new Panopticon vs. privacy. *Proc. of the 10th International Conference on Security and Cryptography*, 98-110, Iceland.

[21] Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*. ACM, 34, 1, 1-47.

[22] Manning,C., Raghavan, P., and Schütze, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

[23] Dreiseitl, S., and Ohno-Machado, L. 2002. Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*. 35, 5, 352-359.

[24] Mitrou, L. 2010. The impact of communications data retention on fundamental rights and democracy: The case of the EU Data Retention Directive. *Haggerty/Samatas*.127-147.

[25] Mylonas, A., Meletiadis, V., Tsoumas, B., Mitrou, L. and Gritzalis, D. 2012. Smartphone Forensics: A Proactive Investigation Scheme for Evidence Acquisition. *Proc. of the 27th International Information Security and Privacy Conference.* Springer, 249-260.

[26] Mylonas, A., Meletiadis, V., Mitrou, L., and Gritzalis, D. 2013. Smartphone sensor data as digital evidence. *Computers & Security* (Special Issue: Cybercrime in Digital Economy).

[27] Gritzalis, D. 2004. Embedding privacy in IT applications development. *Information Management and Computer Security*. 12, 1, 8-26.

[28] Gritzalis, D., and Mallios, J. 2008. A SIP-based SPIT management framework. *Computers & Security*. 27, 5-6, 136-153.

[29] Gritzalis, D. 1997. A baseline security policy for distributed healthcare information systems. *Computers & Security*. 16, 8, 709-719.

[30] Gritzalis, D. 1998. Enhancing security and supporting interoperability in healthcare information systems. *Informatics for Health and Social Care*. 23, 4, 309-324.

[31] Spirakis, P., Katsikas, S., Gritzalis, D., Allegre, F., Darzentas, J., Gigante, C., Karagiannis, D., Kess, P., Putkonen, H., and Spyrou, T. 1994. SECURENET: A network-oriented intelligent intrusion prevention and detection system. *Network Security.* 1,1.