SCIENCE CHINA
Information Sciences

• LETTER •

September 2022, Vol. 65 199102:1–199102:3
https://doi.org/10.1007/s11432-019-2776-4

A benchmark for visual analysis of insider threat detection

Ying ZHAO[1], Kui YANG[1], Siming CHEN[2*], Zhuo ZHANG[3], Xin HUANG[3], Qiusheng LI[3], Qi MA[3], Xinyue LUAN[3] & Xiaoping FAN[4]

[1]*School of Computer Science and Engineering, Central South University, Changsha* 410075*, China;*
[2]*School of Data Science, Fudan University, Shanghai* 200433*, China;*
[3]*Qi An Xin Group, Beijing* 100015*, China;*
[4]*School of Information Technology and Management, Hunan University of Finance and Economics, Changsha* 410205*, China*

Received 12 August 2019/Revised 4 November 2019/Accepted 20 January 2020/Published online 7 September 2021

Citation Zhao Y, Yang K, Chen S M, et al. A benchmark for visual analysis of insider threat detection. Sci China Inf Sci, 2022, 65(9): 199102, https://doi.org/10.1007/s11432-019-2776-4

Dear editor,

Malicious insiders are trusted entities that are given the power to violate rules in a given security policy, and insider threats occur when trusted entities abuse such a power. In modern networked work environments, malicious insiders and their malicious activities pose serious threats to organizations and businesses. Trusted employees travel the intranet of an organization with a high degree of freedom. They are familiar with internal systems and can access various resources. Their network behaviors can hardly be controlled. This situation may lead to high-security risks, thereby making data theft, information technology sabotage or fraud likely to occur.

Numerous automatic and interactive analysis methods have been proposed to predict and stop insider threats. However, a significant impediment for insider threat research is the lack of suitable benchmarks. The major reason for this deficiency is that confidentiality and privacy concerns make real-world data inconvenient to disclose. Another reason is that the data for insider threat detection should contain a complete and detailed explanation of human behaviors. To date, only a few data sets, such as the CERT insider threat [1], VAST Challenge 2009 [2], RUU [3], SEA [4], and WUIL [5] data sets, are available (see Appendix B). This state forms a sharp contrast to the blossoming of benchmarks in the fields of computer vision and machine learning. Accordingly, this state is extremely detrimental to the development of insider threat detection technology.

Visual analytics involve humans in the loop to deal with complex scenarios and analysis tasks. The proposed visual analytics are commonly evaluated in non-unified means of user cases or user studies because of the lack of standard data sets, criteria, and processes to verify. Benchmarking is an important method to enable visual analytics researchers and practitioners to test and verify their methods. The IEEE VAST Challenge endeavors to propose benchmarks with immense effort and success. However, compared with the rapid growth of application demands of visual analytics, the number of benchmarks is considerably limited. Different from benchmarks in computer vision and machine learning, which focus on accurate ground truth, constructing a benchmark data set for visual analytics requires carefully-defined analytical scenarios, diverse behavior models, story-telling ground truth, and standard evaluation criteria.

In this study, we introduce an open-source benchmark data set called ITD-2018, which is specifically designed for insider threat detection domain. The scenario of the ITD-2018 data set is set in a virtual Internet company called HighTech. The time of the scenario is set on the eve of the release of a new flagship product. To protect the core interest of the company and ensure the successful release of the new product, the executive decides to form an insider threat intelligence analysis group. The task of the group is to analyze the potential security threats on the basis of the gathered data within the company.

The ITD-2018 data set has a one-month time span and consists of five types of data, namely, work punching in and out log, web browsing log, email log, server logging-in log, and TCP affic log. The data set is saved by days in 'csv' format. The total data size is 126 MB before compression. Each type of data is specified as follows.

• Punching log. This log records the work starting and ending time. The data fields include employee ID, date, and punching in and out times. If the punching in and out times of a record are 0, then the employee was absent on that day.

• Email log. This log records the email servers' activities. The data fields include sending/receiving time, protocol, source IP and port, destination IP and port, a person who sends the email, person(s) who receives the email, and email subject.

• Web browsing log. This log records all website visiting behaviors. The data fields include time, source IP and port, destination IP and port, and visiting host name. If a visit

* Corresponding author (email: simingchen3@gmail.com)

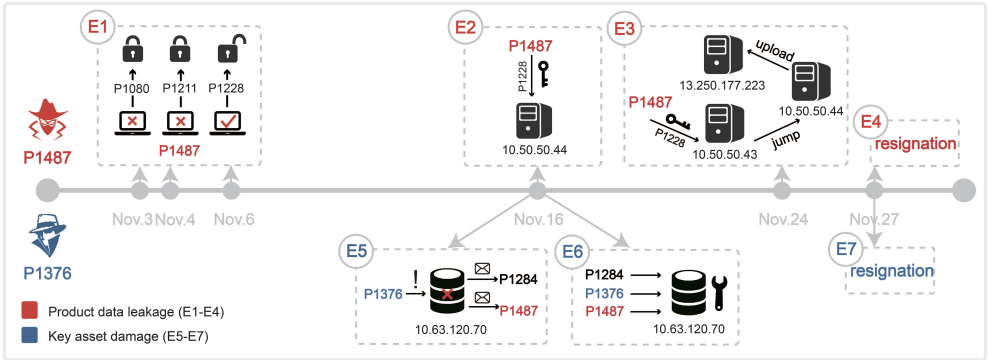

Figure 1 (Color online) Overview of the two main plots. (E1) The SPY (P1487) cracked a leader's account to gain the high data acquisition right. (E2) When solving a sudden database failure, he used the cracked account to locate the target server where the confidential product data was stored. (E3) A few days later, P1487 used the cracked account to log into a server, and used the server as a jumping server to log into the target server. Lastly, P1487 uploaded the confidential data to an external server. (E4) After completing his mission, P1487 filed for resignation at the end of the month. (E5) The DB deleter (P1376) accidentally carried out an incorrect operation and caused a database failure on a critical server. Two other employees received database alarm emails. (E6) These three people simultaneously maintained the database that night. (E7) P1376 filed for resignation at the end of the month because of the serious effect of his misconduct.

is directly through IP, the DNS process can be omitted and the head of HTTP records the host name as null.

• Server logging in log. An employee can use their own workstation or jump servers to log into servers or databases. The data fields include logging time, user name, protocol, destination IP and port, source IP and port, and login result.

• TCP traffic log. This log records TCP connections occur within the company. The data fields include starting and ending times of connection, protocol, destination IP and port, source IP and port, and unlink and downlink total byte numbers. An email, web browsing, or server logging behavior can generate one or multiple TCP records.

HighTech consists of 5 departments and 299 employees (i.e., 1 CEO, 24 in the finance department, 18 in the HR department, 88 in the development 1, 62 in the development 2, and 106 in the development 3), 299 workstations and 34 servers (i.e., 1 for OA, 1 for email, 1 for git, 1 for jira, 2 for lib, 20 for dev, and 8 for backing up). The organization structure of the company is an easy-to-understand four-layer tree. The root node of the tree stands for the CEO. Five sub-trees are under the root node and correspond to the five departments respectively. For the finance and HR sub-trees, the top nodes are the department managers and the leaf nodes are the ordinary employees. For the three development sub-trees, the top, middle, and leaf nodes are the department managers, team leaders, and ordinary employees, respectively. Generally, each employee has a dedicated workstation by default, OA and email servers serve all employees, development servers are only used by the employees of the development departments, and backup servers are connected with other servers for backup operations. The employees perform a series of representative behaviors during working time on the basis of their jobs and pre-defined characteristics. The normal behavior patterns are detailed in Appendix E.

Two main plots and one extension plot are included. The main plots focus on product data leakage and key asset damage, including seven events (see Figure 1). The extension plot contains seven independent events (E8–E14). (E8) Four employees, namely, P1183, P1273, P1169, and P1151, uploaded data to the external server 13.250.177.223 through jump servers. (E9) P1281 encountered a family-related incident, prompting him to file for resignation. (E10) Four employees, namely, P1149, P1352, P1383, and P1389, planned to travel together. These employees frequently browsed travel websites, and sent their leave mails to their own leaders. (E11) Every Thursday morning at 9:30, the HR department would send emails to invite all employees to participate in group sports exercises. Employees who wished to participate would reply and depart between 19:00 and 19:20. (E12) Most employees in the finance department worked overtime due to the busy financial work at the end of November 2017 in the company. (E13) Eight employees, namely, P1147, P1283, P1284, P1328, P1334, P1376, P1487, and P1494, used VPN to remotely connect to the company's intranet to work overtime during the weekend. (E14) A bug in the TCP log system caused the SMTP network protocol of some email records to be marked as HTTP (see Appendix E).

A programme-driven method is proposed to generate the ITD-2018 data set. Firstly, we carefully define the data set scenario, and use diverse models to formulate the relationships and behaviors of the employees and non-human assets in the scenario. Secondly, we design and implement a data generator. This generator adopts a single-person-single-day strategy to generate the background data and a script-driven method to generate the threat data. Lastly, the background and threat data are merged with a contradiction elimination process to form the final data set. The data generation processing is introduced in Appendixes C and D.

A detailed ground truth and quantitative effectiveness scoring criterion is provided to help data users evaluate their technologies and systems by using the ITD-2018 data set. Moreover, the ITD-2018 data set was applied in the ChinaVis Data Challenge 2018[1)], and 77 entries submitted by 342 participants were received. We introduce the evaluation scheme used in the data challenge and share our evaluation experiences. Based on the evaluation results of the 77 entries and feedback from the participants, we rethink our data design and evaluation (see Appendix F).

In summary, our ITD-2018 data set is a new insider threat benchmark data set. Compared with the previous data

1) ChinaVis Data Challenge 2018. http://www.chinavis.org/2018/challenge.html.

sets, it has multiple state-of-the-art features. (1) Diverse data sources. ITD-2018 contains five heterogeneous data sources. (2) Vivid story plots. Two main plots and seven extension plots are embedded in ITD-2018, which forms multiple coherent and intricate storylines. (3) Detailed ground truth. ITD-2018 provides a very detailed ground truth, including elaborated scenarios, storyline narratives, and major players, assets and time information of each plot. (4) Complete evaluation scheme and experience. We provide a quantitative effectiveness scoring criterion to enable data users to evaluate their analysis results by using ITD-2018 data set. We share our experience gained from evaluating 77 entries of the ChinaVis Data Challenge 2018 by using the scoring criteria. We also summarize the difficulty of detecting each event based on the analysis results of the 77 entries (see Appendix G).

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant No. 2018YFB1700403), National Natural Science Foundation of China (Grant Nos. 61872388, 62072470), and Natural Science Foundation of Hunan Province (Grant No. 2020JJ4758). ITD-2018 project at Github: https://github.com/csuvis/InsiderThreatData. Thanks all the organizers, reviewers, and participants of ChinaVis Data Challenge 2018.

Supporting information Appendixes A–G. The supporting materials include two parts: (1) ITD-2018 data set; (2) reviewer guide of ChinaVis Data Challenge 2018. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Glasser J, Lindauer B. Bridging the gap: a pragmatic approach to generating insider threat data. In: Proceedings of IEEE Security and Privacy Workshops, 2013. 98–104
2 Grinstein G, Scholtz J, Whiting M, et al. VAST 2009 challenge: an insider threat. In: Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST), 2009. 243–244
3 Salem M B, Stolfo S J. Modeling user search behavior for masquerade detection. In: Proceedings of International Workshop on Recent Advances in Intrusion Detection, Berlin, 2011. 181–200
4 DuMouchel W, Ju W H, Karr A F, et al. Computer intrusion: detecting masquerades. Statist Sci, 2001, 16: 58–74
5 Camiña J B, Hernández-Gracidas C, Monroy R, et al. The windows-users and -intruder simulations logs dataset (WUIL): an experimental framework for masquerade detection mechanisms. Expert Syst Appl, 2014, 41: 919–930