SCIENCE CHINA
Information Sciences

• **Supplementary File** •

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 Engineer, 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] *Hunan University of Finance and Economics, Changsha* 410205*, China*

Appendix A Introduction

We introduce an open-source benchmark data set tailored to visual analytics community, called ITD-2018, which is specifically designed for insider threat detection domain. The background of the ITD-2018 data set is set in a virtual Internet company with hundreds of employees. The data set consists of five types of data sources, namely, work punching in and out, web browsing, email, server logging-in, and TCP traffic logs. These data sources contain a wide range of information categories, such as multi-dimensional data, multi-entity graph, text information, and time-varying information. The data set is embedded with numerous coherent events. These events form multiple intricate storylines and involve numerous major players, relevant assets, and human behaviors. Some events pose great threats to the core interest of the company.

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.

We provide detailed ground truth and quantitative effectiveness scoring criteria 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, 2], 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 scheme, raise the reflection and implication for visual analysis, and discuss the limitation of our approach and future directions.

Appendix B Previous Data

In the insider threat analysis domain, the lack of public real-world data sets is a long-standing problem because of confidentiality and privacy issues. To address this problem, researchers have created and released some synthetic data sets to promote the research in this domain.

The CERT data set [3] is the most well-known public data set for insider threat detection. This data set has been updated several times and the current version is R6.2. The advantages of this data set are that it contains five types of insider behavior data sources (i.e., log-on/log-off, http traffic, email, file operation, and external storage device usage), and these data sources are embedded into five plot lines. The drawbacks of the data set are the huge size and lack of clear scenario definition, sufficient ground truth, and systematic evaluation standards.

The IEEE VAST Challenge [4–8] is the most famous annual contest in the visual analytics community. The VAST Challenge provides synthetic data sets to help researchers test whether their visual analytics solutions would be beneficial for particular analytical tasks. For the past decade, many IEEE VAST Challenge data sets related to security and intelligence analysis scenarios have been conducted [9,10]. The IEEE VAST Challenge 2009 [11] focused on insider threat analysis, which particularly focused on a serious information exfiltration incident at an embassy. The contestants were required to find spies and identify their suspicious behaviors by using three data sources, namely, network traffic data, social network data, and video surveillance. The data set features attractive scenario design and multi-faceted data sources, including video data,

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

Table B1 Basic information about the ITD-2018 data set and comparison among the previous public insider threat data sets

Data Name	Data Types	Threat Types	Plot Complexity	Ground Truth	Evaluation Scheme	Data Generation Methods
CERT insider threat dataset [3]	logon/logoff, usb device usage, http traffic, file access, and email	information technology sabotage, property theft, and data fraud	complex	simple	–	programme-driven
IEEE VAST Challenge 2009 [11]	network traffic data, social network data, and video surveillance	data leakage	complex	detailed	detailed	programme-driven
SEA 1998 [12, 13]	unix command	masquerade attack	simple	simple	–	terminal-driven
RUU 2011 [14]	registry-based activity data, process action data, and file accesses data	masquerade attack	simple	simple	–	terminal-driven
WUIL 2014 [15]	file operation data	masquerade attack	simple	simple	–	terminal-driven
ITD-2018 (Current Work)	work punching in and out log, web browsing log, email log, server logging-in log, and TCP traffic log	identity theft, data leakage, and key asset damage	complex	detailed	detailed	programme-driven

detailed ground truth, and reference evaluation standard. These advantages have constantly been the traditional virtues of a series of VAST Challenge data sets in the past decade. Learning from the VAST Challenge, we organized four years' ChinaVis Data Challenge. The supplementary material provides a brief introduction of the ChinaVis Data Challenge.

The SEA [12, 13], RUU [14], and WUIL [15] are three public data sets for masquerade attack detection. Masquerade attack [14] is a situation that involves a user illegally impersonating another legitimate user. In practice, the majority of masquerade attacks occur within an organization. The SEA data set records the command operations of 70 users under a UNIX system, whilst the commands of 20 users of them are injected with artificial masquerade attacks. The masquerade attacks in the RUU and WUIL data sets are collected from the system operations of real-world users. However, the SEA, RUU, and WUIL data sets only concern one single type of insider threat, and lack clear scenario definition and systematic evaluation standards.

Apart from the aforementioned data sets, quite a few public data sets exist in other domain and relatively contain insider threats. For example, KDD CUP 99 [16] is a well-known data set in data mining and network security domain. The data set contains numerous network intrusion behaviors. Similar data sets include the NSL-KDD and DARPA 2000.

Our ITD-2018 data set is a new insider threat benchmark data set. Compared with the previous data sets, it has multiple state-of-the-art features. Table B1 summarizes the differences between our data set and the previous ones. (1) Diverse data sources. ITD-2018 data set contains five heterogeneous data sources. (2) Vivid story plots. Two main plots and seven extension plots are embedded in ITD-2018 data set, which forms multiple coherent and intricate storylines. (3) Detailed ground truth. ITD-2018 data set 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 criteria 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.

Appendix C Scenario definition and modelling

Scenario is a story that takes place over a finite time period and contains a series of characters, places, and plot lines. Scenario definition and modelling are the primary steps in data creation, and determine the design of all contents in a scenario. This section illustrates the detailed process of our scenario definition and modelling.

Appendix C.1 Scenario definition

Our scenario is set in a virtual and medium-sized Internet company called HighTech. The virtual company is based on a real-world company. We set HighTech with a simple but general organization structure. HighTech has three types of departments (i.e., finance, human resource (HR), and development), and four job levels (i.e., executive, department manager, team leader, and ordinary employee). The HighTech staff members perform a series of representative activities during working time, such as emailing and web browsing. The characteristics of every employee are carefully defined to

Table C1 Attributes of the three types of objects

Object Types	Attribute Types	Attributes Description
Employee	Personal attributes	ID, age, and gender
	Work attributes	Department, job level, working age in HighTech, superior employee, email address, and dedicated workstation
	Behavior attributes	Preferred times of punching in and out, probabilities of late, leave early, absenteeism and overtime work, favorite themes of websites/webpages, probabilities of handling different email subcategories, commonly used email topics, and probabilities of logging into different servers
Department	–	Daily work schedule (see Table C2)
Asset	–	IP address and asset type: workstation or server (OA, email, git, jira, lib, dev, or backup server)

Table C2 Daily work schedules of defferent departments

Department	Time Periods				
	Punching In	Morning Work	Lunch Break	Afternoon Work	Punching Out
Finance	8:00	8:00 – 12:30	12:30 – 13:30	13:30 – 17:00	17:00
HR	9:00	9:00 – 12:30	12:30 – 13:30	13:30 – 18:00	18:00
Development 1	9:00	9:00 – 12:30	12:30 – 13:30	13:30 – 18:00	18:00
Development 2	9:00	9:00 – 12:30	12:30 – 13:30	13:30 – 18:00	18:00
Development 3	10:00	10:00 – 12:30	12:30 – 13:30	13:30 – 19:00	19:00

ensure that employees can act diversely. Moreover, we set the time of the scenario on the eve of the release of a new flagship product. This makes the scenario includes not only common events (e.g., overtime work and resignation) but also some special events related to the product release (e.g., product data leakage and key asset damage). These events can constitute multiple plot lines with different levels of importance and difficulty to discover. Based on these settings, the story is expected to be pleasing and sufficiently challenging, thereby inspiring data users and researchers for problem solving.

Appendix C.2 Scenario modelling

Scenario modelling creates objects, relationships, and behaviors that exist in the scenario and formulates them with diverse models. To achieve a high-quality scenario modelling, we conducted in-depth discussions with the manager and two employees and reviewed the relevant anonymized data sources of the referred company. Considering the common structure of the referred company, we tried to model the scenario with most representative objects and relationships. After several rounds of discussions, we determined that we needed to model three types of objects (i.e., employees, departments, and non-human assets) and two important relationships (i.e., personnel and asset relationships). Behaviors of objects should be diverse and faithful to the features of high-tech company. We eventually modeled four types of behaviors (i.e., punching in and out, web browsing, emailing, and logging into servers). Moreover, the threat events with different levels of concealment were designed in scenario modelling.

Object modelling abstracts the characteristics of the three types of objects and transforms them into descriptive attributes. Table C1 shows that the employee object is constructed with three types of attributes, namely, personal, work, and behavior attributes. The personal attributes include an employee's ID, age, and gender. The work attributes, such as the department and job level, describe an employee's work and duties. The behavior attributes are related to the personal and work attributes. We use probability and preference to model the chance that an employee exhibits a specific type of behavior. For example, male employees have a high probability of browsing sports websites, and HR staff members often handle recruitment emails. The department object has one attribute, namely, the time requirement for the daily work. HighTech is set with five departments: one finance, one HR, and three development departments. The time requirements of the departments are slightly different from one another, thereby reflecting diverse job duties (see Table C2). The non-human asset objects refer to workstation computers and servers. Asset attributes are the IP address and asset type. The servers in the scenario can be divided into OA, email, and development servers (e.g., git, dev, jira, lib, and backup servers) on the basis of their functionalities.

Relationship modelling abstracts two important relationships in the scenario. (1) *Personnel relationship* refers to HighTech's organizational structure. Figure C1 shows that we use 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. (2) *Asset relationship* is modelled by a multi-entity graph. In the graph, a vertex can represent an employee, workstation, or server, and an edge indicates a relationship between an employee and a workstation, a workstation and a server, or a server and another server. 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

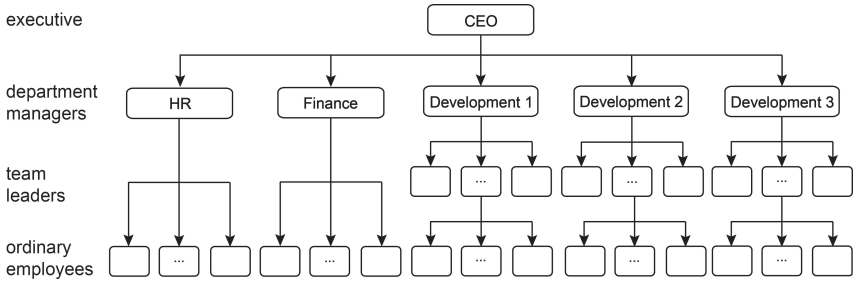

Figure C1 HighTech's organizational structure.

departments, and backup servers are connected with other servers for backup operations.

Behavior modelling formulates the employees' behaviors during working time. Each employee's behavior should be unique but consistent with certain rules and patterns to reflect his/her work duties and personal preferences. The scenario has four types of behaviors, whilst each type is constructed by at least one model.

(1) *Daily work schedule modelling* simulates the distribution of various behaviors over time to formulate the occurrence time of any activity. Table C2 shows that we divide a day into five time periods, namely, punching-in, morning work, lunch break, afternoon work, and punching-out. The punching-in period has a typical punching-in peak within 10–20 min before the morning working period. Thus, we create a left-skewed distribution function to simulate the temporal pattern of punching-in activities. After punching in, the majority of employees often browse webpages for a while, handle emails thereafter, and eventually start other working tasks. We use three trapezoidal distributions to divide the morning working period into three chronological sub-periods, namely, web browsing sub-period lasting for approximately 20 min, email handling sub-period for approximately 30 min, and main working sub-period until 12:30. During the main working sub-period, the majority of employees may handle emails and access the web systems in the OA servers. Meanwhile, employees from the development departments may log into the development servers. During the lunch break period, we simulate a short peak of web browsing because the majority of employees intend to browse news or online entertainment for a short relaxation. For the afternoon working period, the temporal behavior patterns are similar with those of the morning main working sub-period. After work, the majority of employees punch out within 1.5 hours. However, some employees may work overtime, which is simulated by a long-tailed trapezoidal distribution. Additionally, the working hours of the executive and department managers are flexible.

(2) *Web browsing modelling* uses theme and graph models to simulate web browsing behaviors. Firstly, we build a website/webpage and theme library that contains the information of various websites/webpages and their relevant themes. All themes are divided into two categories, namely, intranet-related and extranet-related. The intranet-related themes refer to the web-based systems deployed on the company's intranet servers, such as the OA, email, and git systems. The extranet-related themes are websites or webpages, including 21 theme subcategories (e.g., searching, recruitment, technology, e-commerce, social, videos, sports, news, finance, automotive, lottery, real estate, tourism, etc.). Secondly, we select theme preferences for each employee on the basis of his/her personal and work attributes. For example, male employees prefer to visit sports and game websites, and development employees like technical forums. Lastly, we set that all intranet web browsing behaviors should be restricted to the asset relationship graph. For example, the OA and email systems can be accessed by all employees, whereas the git, jira, and lib systems can only be accessed by development employees.

(3) *Email behavior modelling* uses graph and topic models to simulate emailing behaviors. Firstly, we define two categories and seven subcategories of emails: internal emails (i.e., emails within a team, emails among different levels of employees, emails among different departments of employees, software or hardware alarms, and internal notifications) and external emails (i.e., recruitment emails and spam). We use the company's organization tree to build an email communication graph to regulate email communications. That is, email communications should occur on the links of the graph. In the graph, a vertex represents an email address and an edge indicates the link between two email addresses. Secondly, we set the probability for each email subcategory to control the occurrence frequency. In this scenario, alarm emails and spam are most likely to occur, followed by emails within a team in the finance or development departments and recruitment emails in the HR department. Lastly, we set textual topics for each email subcategory. Each employee is assigned preferred email topics on the basis of his/her work duty. For example, the finance department staff members mainly send and receive emails with financial, accounting, and reimbursement topics. The email topics of HR staff members mainly include performance appraisal, labor contracts, and welfare guarantees.

(4) *Server login modelling*: The asset relationship graph indicates that, only development employees are able to log into development servers. We summarize several common types of server operations, such as database management, data backup, upload, and download. Development employees randomly access each server and perform different types of operations.

Threat event modelling is responsible for detailing the threat events that occurred in HighTech. In contrast with all the aforementioned models designed for constructing normal behaviors, a threat event is a collection of abnormal behaviors that performs malicious activities, thereby affecting the normal operations of the company and even threatening the new product launch. To realistically design threat events, we had in-depth communication with the three employees in the real-world reference company to determine the threat event details. We describe the threat event details in Appendix E.

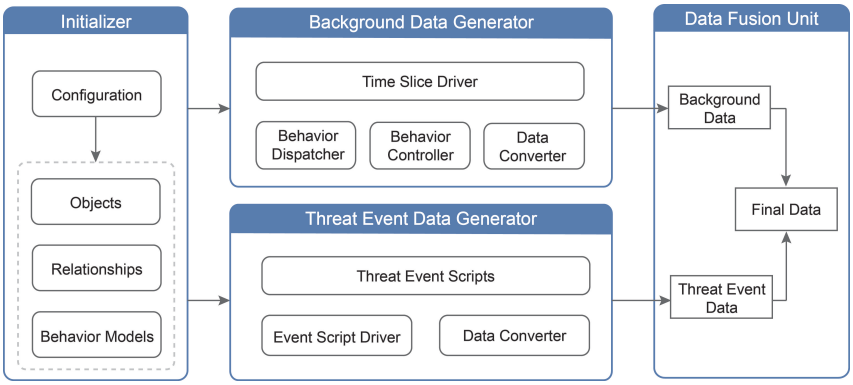

Figure D1 Workflow of the Insider Threat Data Generator.

Appendix D Data generation and description

We designed and implemented an Insider Threat Data (ITD) Generator to generate insider threat data sets on the basis of the pre-defined scenario and models (Appendix C). This section illustrates how the ITD Generator works and the basic information of the ITD-2018 data set.

Appendix D.1 Data generation

The ITD Generator is a tool that can generate insider threat data sets. This generator is initialized by the configurations, separately generates the background and threat data, eventually merges the two types of data. Figure D1 shows that the ITD Generator consists of four modules.

Initializer creates employees, departments, and assets, and initializes their attributes and relationships. The initializer uses configuration files to set the parameters of the scenario and models. The output of the initializer is a profile file that describes all created objects and their attributes. The initialization process includes the following steps:

Step 1: Set data types and their data fields.

Step 2: Create departments, set the number of employees and daily work schedule for each department.

Step 3: Create employees, initialize their personal attributes and work attributes, and construct their personal relationships.

Step 4: Create assets, initialize their attributes and relationships.

Step 5: Initialize the behavior models discussed in Appendix C.2.

Step 6: Initialize the employees' behavior attributes. Firstly, we design several employee templates with different departments and job levels and manually set the behavior attributes of the templates. Secondly, the initializer automatically delivers the behavior attributes of employee templates to all employees with regard to department and job level. A randomized method is adopted in attribute delivering to make the attributes of employees slightly different. For example, the probabilities of logging into the git servers of all employees in the first development department are initially copied from the employee template of the department. After randomizing, a deviation of under 5% within these probabilities will be obtained.

Step 7: Fuse all the initialization results to generate a profile file, which can be further tuned manually.

Background data generator reads the profile file to automatically simulate the normal behaviors of the employees in their daily work to generate the background data. The background data generator adopts a single-person-single-day loop strategy. Figure D1 shows that four sub-modules, namely, *time slice driver*, *behavior dispatcher*, *behavior controller*, and *data converter*. We use the background data of a specific employee as an example to explain the manner by which the background data generator works.

Step 1: The *time slice driver* is the module responsible for temporal information division. It creates the time of the employee's punching in and out, and bins the entire working time by seconds.

Step 2: The *time slice driver* requests a behavior dispatching from the *behavior dispatcher* for each time bin.

Step 3: The *behavior dispatcher* determines whether the employee do some actions in each time bin on the basis of the daily work schedule model, while the *behavior controller* is designed to set the attributes of the behaviors. Once the *behavior dispatcher* decides that the employee acts in the current time bin, the *behavior controller* will determine the behavior type, consuming time, and other parameters. For example, a mailing behavior should contain the mailing time, mail sender, mail recipients, and mail subject. A web browsing behavior should be fulfilled by the record generation time, source IP, source port, destination IP, destination port, and requested domain name.

Step 4: The *behavior controller* delivers the parameterized behavior to the *data converter*, which is responsible for converting this behavior into the data records of the corresponding data types. For example, an emailing behavior will generate one mail records and one TCP traffic records. A web browsing behavior will generate one web browsing record and multiple TCP traffic records.

Step 5: The background data generator returns to Step 2 and starts a new behavior request. When the time slice arrives at the punching out time of the employee, the generator returns to Step 1 and creates the time of the employee's punching

Table E1 Normal patterns by departments

Department	Normal Patterns		
	Preferred Website Themes	**Preferred Email Topics**	**Accessible Servers**
Finance	OA, email, etc.	financial, accounting, reimbursement, etc.	OA, email
HR	OA, email, etc.	performance appraisal, labor contracts, welfare guarantee, etc.	OA, email
Development 1	OA, email, git, jira, lib, technology, etc.	demand analysis, software development, etc.	OA, email, git, jira, lib, dev, backup
Development 2	OA, email, git, jira, lib, technology, etc.	demand analysis, software development, etc.	OA, email, git, jira, lib, dev, backup
Development 3	OA, email, git, jira, lib, technology, etc.	demand analysis, software development, etc.	OA, email, git, jira, lib, dev, backup

in and out for the next day.

Threat event data generator is responsible for the generation of the threat event data. We use a pre-written script to define the major players, relevant assets, and activities of an event. The script is executed by the threat event data generator to generate the data related to the event. The generator has two sub-modules, namely, the *event script driver* and *data converter*, and works through the following four steps:

Step 1: Initialize all scripts, specify the occurrence time, employees, assets involved, and behavior sequence of each event.

Step 2: The *event script driver* selects a script, and drives the relevant employees to perform the behaviors by time.

Step 3: The *data converter* converts the behaviors into data records.

Step 4: Repeat Step2 until all scripts have been executed.

Data fusion unit. Some contradictions may exist because the background and threat data are produced separately. For example, an employee did not come to work for three days because of an unknown reason. This situation is a threat event and the employee should expectedly have no data record during the three days. However, the background data generator may have generated some background data of the employee for the three days. Data fusion unit is designed to solve the contradictions. In data fusion, we set that the data generated by the threat event data generator has a higher priority than background data generator. Thus, the background data should be deleted or modified if they are in conflict with the threat data. After eliminating the contradictions, we merge the two types of data and sort them in chronological order to obtain the final data set.

Appendix D.2 Data description

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 traffic log. 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 data set is saved by days in 'csv' format. The total data size is 126MB before compression. Each type of data is specified as follows.

Punching log: This log records the work starting and ending time of the employees. 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 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 all TCP connections occur within the company. The data fields include starting and ending times of TCP 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.

Appendix E Ground truth

Appendix E.1 Background

HighTech is an Internet company with hundreds of employees and five departments. Since the company is on the eve of the release of its new flagship product, the executive must remain vigilant. 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.

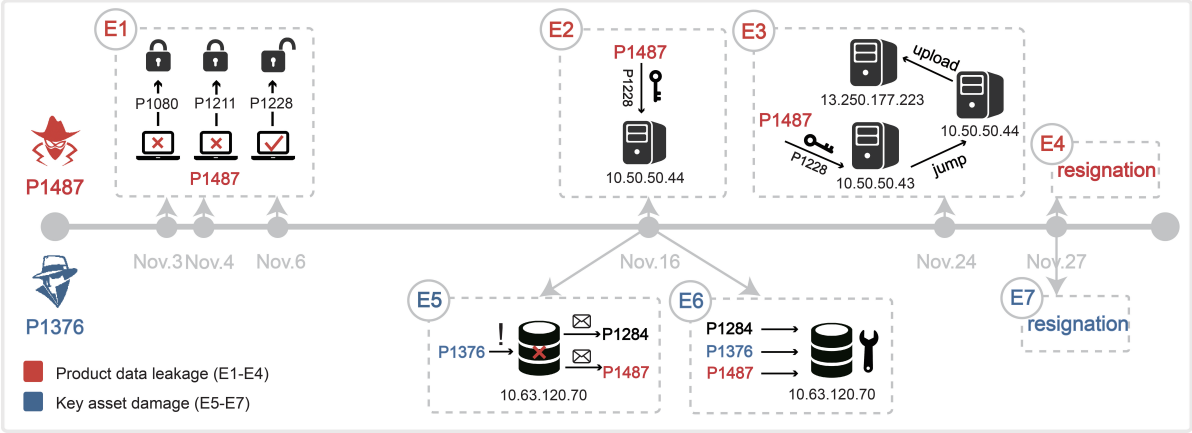

Figure E1 Overview of the two main plots. (E1) The SPY (P1487) cracked a leader's account to gain the high data acquisition right. (E2) When solving a sudden database failure, he used the cracked account to locate the target server where the confidential product data was stored. (E3) A few days later, 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.

Appendix E.2 Normal patterns

Normal patterns are the common behavior patterns of employees, that do not pose any threat to company security. Generally, three typical normal patterns are marked P1-P3. P1 denotes the company organizational structure (Figure C1), P2 represents the work schedules of different departments (Table C2), and P3 indicates the normal behavior patterns of different departments (Table E1).

Appendix E.3 Threat events

Two main plots and one extension plot are included in our scenario. The main plots focus on product data leakage and key asset damage. Both plots contain multiple coherent threat events that involve many characters and assets. These threat events are closely relevant to the company's new product launch. The extension plot contains some independent events but poses no direct threat to the product launch. Figure E1 shows the process of the two main plots.

Product data leakage (main plot 1): HighTech has constantly been engaged in a fierce business competition with another company. To gain an edge in the competition, the rival company bribed P1487, an employee in HighTech's third development department. P1487, whom we call "The Spy", was required to steal relevant data of the new product to weaken HighTech. To complete the task without being discovered, P1487 formulated a plan, as illustrated in Figure E1. The specific process of the main plot is shown in E1–E4.

E1-account stealing: P1487 attempted to log into the accounts of leaders P1080, P1211, and P1228 on November 3, 4 and 6, 2017, respectively. After several failures, P1487 successfully cracked the password of P1228's account because of the weak complexity.

E2-product data peeping: On November 16, 2017, P1487 signed up for a company group activity, but he actually did not participate because he had to maintain a failed database server. During the maintenance process, P1487 used P1228's account to log into the target server 10.50.50.44, to check whether the server has important data related to the release of the new product.

E3-product data leakage: On November 24, 2017, P1487 logged into the server 10.50.50.43 at 12:43 using the account of P1228. Thereafter, he used this server as a jump server to log into the target server 10.50.50.44. Lastly, he uploaded the product data to the external server 13.250.177.223.

E4-the spy resignation: P1487 frequently browsed recruitment websites and received many emails from headhunters. On November 27, 2017, he filed for resignation.

Key asset damage (main plot 2): The third development department employee P1376, whom we called "The DB Deleter", had already planned to resign. Therefore, he recently was absent-minded often. One day, P1376 accidentally carried out an incorrect operation and caused a database failure on a critical server, as illustrated in Figure E1.The specific process of the story is in E5–E7.

E5-database failure: On November 16, 2017 at 19:22, employee P1376 accidentally caused a database failure on server 10.63.120.70. Thereafter, P1487 (The Spy) and P1284 received database alarm emails.

E6-database maintenance: The three employees, namely, P1376, P1487, and P1284, simultaneously maintained the database that night and left the company after completing the work at approximately 23:30.

E7-the DB Deleter resignation: P1376 frequently browsed recruitment websites and received numerous emails from headhunters in this month. After the database failure event, he filed for resignation on November 27, 2017.

Branch events (extension plot): The extension story contains many independent events (E8–E14) that are common in Internet companies, such as employees working overtime and VPN access. These events pose no direct threat to the company's new product launch.

E8-jump server event: On November 17, 21, 27, and 30, 2017, four employees, namely, P1183, P1273, P1169, and P1151, uploaded data to the external server 13.250.177.223 through jump servers. Unlike P1487, these four employees were merely performing their duties.

E9-resignation event: P1281 encountered a family-related incident, prompting him to file for resignation on November 27, 2017.

E10-tourist event: Four employees, namely, P1149, P1352, P1383, and P1389, planned to travel together. These employees frequently browsed travel websites from November 20 to 24, 2017, and sent their leave mail to their own leaders on Friday, November 24, 2017. Their travel scheme was from November 25 to 30, 2017.

E11-group activity event: Every Thursday morning at 9:30, the HR department would send emails to invite all employees to participate in group sports exercises, such as badminton. Employees who wished to participate would reply and depart between 19:00 and 19:20.

E12-financial department overtime work event: On the weekends of November 19, 25, and 26, 2017, most employees in the finance department worked overtime due to the busy financial work at the end of November 2017 in the company.

E13-VPN remote access event: 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. P1059 did not report to the company on Tuesday, November 28, 2017. He accessed the intranet and approved the resignation applications of two employees, namely, P1376 (The DB Deleter) and P1487 (The Spy), through VPN.

E14-traffic monitoring system failure: A bug in the TCP log system caused the SMTP network protocol of some email records to be marked as HTTP from November 10 to 28, 2017.

Appendix F Evaluation

This section presents a scoring criteria for effectiveness evaluation, and shares the evaluation experiences of using the ITD-2018 data set in the ChinaVis Data Challenge 2018.

Appendix F.1 Evaluation dimensions

Whether a technical approach or system is valuable depends on the degree that the approach or system can help analysts perform their job better, in which better may mean faster, more effective, or some special aspects that contribute to success. Generally, the two main types of evaluation are objective and subjective evaluations [17–23], which verify the application values of a technology or system.

Objective evaluation aims to quantify effectiveness and efficiency. Effectiveness is whether a technology or system can accurately extract valuable information from the data. For a benchmark data set, effectiveness can be quantified by ground truth. Efficiency evaluates the speed of problem solving. For automatic analysis, consuming time is naturally a preferred quantitative indicator. For interactive analysis, we can also regard the number of interaction steps required (apart from time) to reach a conclusion.

Subjective evaluation uses people's subjective experiences and judgment to evaluate a technology or system. The two methods of evaluation are as follows: (1) inviting domain experts to evaluate on basis of their experiences and knowledge; (2) recruiting volunteers to perform user study and collect feedback. The six most frequently used dimensions of subjective evaluation are visualization design, interaction design, easy of use, novelty, scalability, and insightfulness.

Appendix F.2 Effectiveness scoring criteria

Effectiveness is the most important among the aforementioned evaluation dimensions. Benchmark data sets generally provide ground truth. Thus, the effectiveness of the solutions of data users that detect the embedded threats, major players and activities can be possibly scored.

We propose a "3W" effectiveness scoring criteria for the event detection of the ITD-2018 data set. "3W" refers to "who", "when", and "what", which represent an event's major players, time information, and descriptions of the related activities, respectively. Given that an event in the ITD-2018 data set is marked as E_i and a solution of data users is S_j. The who, when, and what scores of E_i for S_j are marked as $Who\_E_i\_S_j$, $When\_E_i\_S_j$, and $What\_E_i\_S_j$, respectively. Each of the three scores ranges from 0 to 1. The maximum score that a solution can obtained from E_i is generally 3, which is marked as Max_E_i. However, the maximum scores for a few events, including E11, E12, and E14 in this case, are 2 because these events involved many employees (i.e., E11 and E12) or no employee (i.e., E14). Accordingly, measuring the who score is slightly difficult. We use an importance level to indicate the threat degree of E_i to the company (Table F1), marked as IL_E_i and ranging from 1 to 3 (1 = general, 2 = medium, and 3 = important). The final score of S_j for the detection of all events in the ITD-2018 data set is marked as $Final\_S_j$. The scoring steps are as follows:

- $Who\_E_i\_S_j = N\_who\_E_i\_S_j$ / $N\_who\_E_i$, where $N\_who\_E_i\_S_j$ is the number of correctly identified players and assets of E_i by S_j, $N\_who\_E_i$ is the actual total of players and assets of E_i in ground truth, and $Who\_E_i\_S_j \in [0, 1]$.

Table F1 Importance, average scores, and difficulty levels of the 14 events

Plot lines	Event ID	Importance	Average Score	Diffculty
Product data leakage	E1	Important	2.51	Easy
	E2	Important	1.00	Difficult
	E3	Important	0.98	Difficult
	E4	Important	3.01	Easy
Key asset damage	E5	Medium	1.56	Medium
	E6	Medium	0.94	Difficult
	E7	Medium	2.46	Easy
Branch events	E8	General	0.32	Difficult
	E9	General	2.11	Easy
	E10	General	0.29	Difficult
	E11	General	0.50	Difficult
	E12	General	1.59	Medium
	E13	General	0.96	Difficult
	E14	General	0.17	Difficult

• $When\_E_i\_S_j = N\_when\_E_i\_S_j$ / $N\_when\_E_i$, where $N\_when\_E_i\_S_j$ is the number of correctly identified time elements of E_i by S_j, $N\_when\_E_i$ is the actual total of time elements of E_i in ground truth, and $When\_E_i\_S_j \in [0, 1]$.

• $What\_E_i\_S_j = N\_what\_E_i\_S_j$ / $N\_what\_E_i$, where $N\_what\_E_i\_S_j$ is the number of correctly identified activities of E_i by S_j, $N\_what\_E_i$ is the actual total of activities of E_i in ground truth, and $What\_E_i\_S_j \in [0, 1]$.

• $Final\_S_j = \sum_i ((Who\_E_i\_S_j + When\_E_i\_S_j + What\_E_i\_S_j) \times IL\_E_i) / \sum_i (Max\_E_i \times IL\_E_i) \times 5$, $Final\_S_j \in [0,5]$.

Appendix F.3 Evaluation in ChinaVis Data Challenge 2018

From March to July 2018, ChinaVis Data Challenge 2018 used the ITD-2018 data set as the challenge data set. This competition was a considerable opportunity to validate our data set. China Visualization and Visual Analytics Conference (ChinaVis) is an annual conference co-launched by scholars and researchers in the visualization and visual analytics communities in China. Data Challenge is an annual competition in ChinaVis that provides a set of scenarios, data, and problems to help participants evaluate their visual analytics techniques and tools in solving complex problems. The organizational form of this competition is mainly referred to the IEEE VAST Challenge [4].

Appendix F.3.1 *Submission evaluation*

In the Data Challenge 2018, the mini-challenge 1 used the ITD-2018 data set. This mini-challenge required participants to solve three analytical tasks: (1) depict HighTech's organizational structure, (2) analyze the daily behavior patterns of employees, and (3) identify abnormal events and conclude valuable threat intelligence through exploring the possible associations among the events.

The Data Challenge 2018 mini-challenge 1 received 77 entries submitted by 342 participants. We used a "4-stage evaluation strategy" to review these submissions.

The first stage was the desk-check phase, which assessed whether the submissions contained all the required parts (answer sheet, video, and summary paper) and whether they answered all analytical questions. In this stage, 7 submissions failed, whilst the remaining 70 submissions entered the second stage.

The second stage was the effectiveness evaluation phase. The committee invited two experts who participated in creating the ITD-2018 data set to independently score the submissions using our proposed effectiveness scoring criteria.

The third stage was the external review phase. The committee invited security and visual analysis experts to review these submissions. The reviewers scored a submission from five aspects, namely, overall analysis quality, visual design, interaction design, novelty, and scalability, on the basis of the ground truth. The weights of these five aspects were 40%, 20%, 20%, 10%, and 10%, respectively. The evaluation used a 5-point scale, with 1 being strongly rejected; 2, weakly rejected; 3, borderline; 4, weakly accepted; and 5, strongly accepted. Each submission was randomly assigned to 2 to 3 security experts and 2 to 4 visual analysis experts.

The fourth stage was the committee discussion phase. For each submission, we calculated the mean and variance of the 2 scores from the effectiveness scoring criteria and the 4 to 7 scores from the reviewers. We sorted all works on the basis of the result and determined the winners. Lastly, the competition selected 3 first prize submissions, 3 second prize submissions, 8 third prize submissions, 9 merit award submissions, and 3 special awards, namely, the visual effects, tableau analysis, and technical innovation awards. The award rate was 33.7%.

Appendix F.3.2 *Event difficulty evaluation*

We attempted to analyzed the difficulty of discovering the embedded events on the basis of the results of the submission evaluation using our proposed effectiveness scoring criteria. Thus, we calculated the average score of each event as an indicator. As shown in Table F1, a difficulty level (difficult $\leqslant$ 1.0, medium, or easy $\geqslant$ 2.0) was assigned to each event on the basis of the indicator and feedback from the participants.

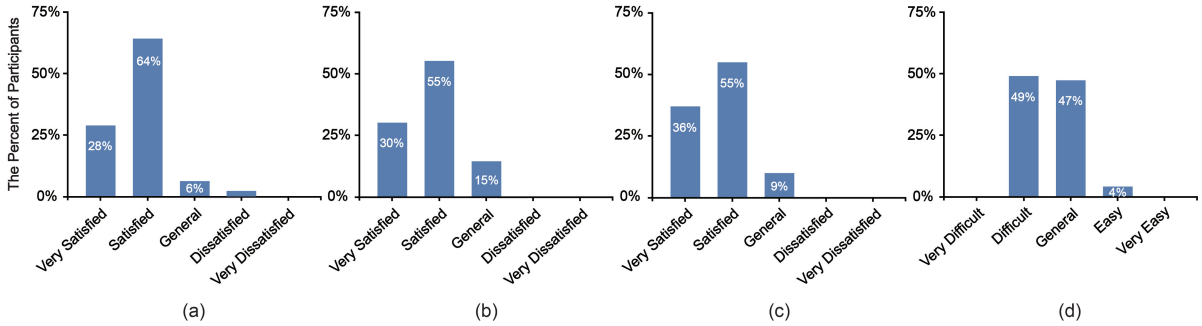

Figure F1 Questionnaire results in the ChinaVis Data Challenge 2018: (a) Satisfaction with the scenario design and analytical tasks; (b) Satisfaction with the ITD-2018 data set; (c) Objectivity and fairness of scores and reviews; (d) Difficulty of solving the analytical tasks.

In the plot of product data leakage, E1 and E4 were two easy events. For E1, the Spy generated many records of login failure. For E4 (including E7 and E9), discovering that the three employees who applied for resignation did not come to work in the last few days of November 2017 was easy. E2 and E3 were labelled as difficult mainly because of the subtle patterns indicated by the minimal related data. Meanwhile, the jump server is substantially concealed. Moreover, the traffic generated by uploading product information was moderate.

For the plot of key asset damage, E5 was not supposed to be difficult because the subject of the database alarm emails was distinct. However, numerous submissions disregarded E5. E6 was difficult because its major players used the SSH protocol to connect to the failed server, which was subtle.

For the branch stories, E8 was difficult because of the high concealment of jump servers. E10 was difficult. Although four major players were easy to identify because of their multiple days of absence, speculating that they wanted to travel together required a deep integrated reasoning on the email data and website browsing data. E11 was also difficult because the relevant email sending records were rare. The HR department used a special email (allstaff@hightech.com) to invite all employees to participate in group sports exercises every Thursday. E12, which is the financial department overtime work event, was not difficult because it involved the entire financial department. E13 was quite difficult because this event required an analysis of the connection between the punching log and the TCP traffic log. Only a few competing teams discovered E14, which may be caused by disregarding the common port number of each network protocol.

Appendix F.3.3 *Participant questionnaire*

A questionnaire survey in the workshop of ChinaVis Data Challenge 2018 was conducted to obtain feedback from the participants. The questionnaire contained two dozen questions. Four of them concerned the participants' opinions on the analytical tasks, data design, review results, and task difficulty (see Figure F1). The feedback of the 64 valid participants who used the ITD-2018 data set was analyzed. The result showed that approximately 90% of the participants were satisfied with our scenario design and analytical tasks (Figure F1(a)). The participants generally reflected that the scenario was interesting and could stimulate their passion and enthusiasm for problem-solving. Approximately 85% of the participants were satisfied with the overall design of the ITD-2018 data set (Figure F1(b)). Some participants commented that the data scale was moderate and suitable. Others appreciated that the integrated analysis of heterogeneous data, particularly finding and connecting clues from different types of data, was considerably interesting and challenging. Approximately 90% of the participants appreciated the scores and comments given by the reviewers (Figure F1(c)), which indicated that most reviews were fair and impartial, and the entire evaluation solution was reasonable and operational. Moreover, approximately 50% of the participants believed that the analytical tasks were generally difficult (Figure F1(d)), thereby indicated that the ITD-2018 data set is challenging and can facilitate the creation of enhanced work.

Appendix G Discussion

This section discusses the limitations of this work and suggests directions for future work.

We exerted efforts to make our scenario and behavior models approximate the real-world. However, we relatively simplified the final design to improve achievability. For example, HighTech's organizational structure was an easy-to-understand four-layer tree and only one executive was present. These things make the IDT-2018 data set less complex than real-world data.

The data generated by the programme-driven method strictly complies with the scenario definition and modelling. Once the scenario and behavior models are determined, the data patterns in the generated background data are also determined. If adjusting the scenario in a macroscopic way, then we have to redo the process of scenario definition and modelling, or even re-program the data generator.

The data sources included in the IDT-2018 data set can be improved in terms of diversity. For example, we can introduce the resumes of employees, spatio-temporal information of major players, and the video surveillance data captured from the company's public areas. Additionally, the medium-sized IDT-2018 data set is relatively easy to handle, but it cannot test the scalability and availability of technologies or systems under large-scale data. Moreover, the insider threat events

embedded in the data set can be richer and more subtle. For example, we can design threat events, such as multiplayer crime, dissemination of false information, and APT attacks can be designed.

The proposed effectiveness scoring criteria still depends on reviewers' subjective experiences. Particularly, the what score ($What\_E_i\_S_j$) of a submission is highly dependent on reviewers' understanding on the textual description of detected events of the submission. One possible solution is to develop an automated evaluation and review system by providing a detailed answer template. However, such a system is difficult to achieve and expert involvement for some subjective evaluation dimensions still remains necessary.

We need to remind data users that our effectiveness scoring criteria may be not suitable for all benchmark data sets. We currently have not tested our scoring criteria on other benchmark data sets. But we believe that our effectiveness scoring criteria can be potentially applied to evaluate the similar analytical tasks of identifying threat events which involve "3W", namely when (event time), who (major IP/port), and what (event description). We may generalize our evaluation method and conduct comparison on multiple benchmark data sets in the future work.

Security domain is evolving with new attacking, monitoring, and defending methods. Our approach is with scalability to extend by defining new models and behaviors. The controllability enables us to set different parameters to generate customized data sets. In the future, we envision to achieve more scalability by automatize the scenario definition, which still remains a challenge with human work.

References

1 ChinaVis conference homepage. Available from: http://www.chinavis.org (in Chinese)
2 Zhao Y, Zhang Z, Yuan X. ChinaVis Data Challenge from 2015 to 2017. Chinese Journal of Network and Information Security, 2018, 4(2):55-61 (in Chinese)
3 Glasser J, Lindauer B. Bridging the gap: A pragmatic approach to generating insider threat data. In: Proceedings of the 2013 IEEE Security and Privacy Workshops, 2013. 98-104
4 Cook K, Grinstein G, Whiting M. The VAST Challenge: history, scope, and outcomes: An introduction to the special issue. Information Visualization, 2014, 13(4):301-312
5 JScholtz J, Whiting M A, Plaisant C, et al. A reflection on seven years of the VAST Challenge. In: Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization, 2012. 13-20
6 Shi R, Yang M, Zhao Y, et al. A matrix-based visualization system for network traffic forensics. IEEE Systems Journal, 2016, 10(4): 1350-1360
7 Liao Z F, Li Y, Peng Y, et al. A semantic-enhanced trajectory visual analytics for digital forensic. Journal of Visualization, 2015, 18(2):173-184
8 Shi Y , Zhao Y , Zhou F , et al. A novel radial visualization of intrusion detection alerts. IEEE Computer Graphics and Applications, 2018, 38(6):83-95
9 Cook K, Grinstein G, Whiting M, et al. VAST Challenge 2012: Visual analytics for big data. In: Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST), 2012. 251-255
10 Grinstein G, Cook K, Havig P, et al. VAST 2011 Challenge: Cyber security and epidemic. In: Proceedings of the 2011 IEEE Conference on Visual Analytics Science and Technology (VAST), 2011. 299-301
11 Grinstein G, Scholtz J, Whiting M, et al. VAST 2009 Challenge: An insider threat. In: Proceedings of the 2009 IEEE Symposium on Visual Analytics Science and Technology (VAST), 2009. 243-244
12 Schonlau M. Masquerading user data, 1998. Available from: http://www.schonlau.net
13 Schonlau M, DuMouchel W, Ju W H, et al. Computer intrusion: Detecting masquerades. Statistical Science, 2001, 16(1):58-74
14 Salem M B, Stolfo S J. Modeling user search behavior for masquerade detection. In: Proceedings of the International Workshop on Recent Advances in Intrusion Detection, Springer, Berlin, Heidelberg, 2011. 181-200
15 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 Systems with Applications, 2014, 41(3):919-930
16 Tavallaee M, Bagheri E, Lu W, et al. A detailed analysis of the KDD CUP 99 data set. In: Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009. 1-6
17 Crouser R J, Franklin L, Endert A, et al. Toward theoretical techniques for measuring the use of human effort in visual analytic systems. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1):121-130
18 Dasgupta A, Lee J Y, Wilson R, et al. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. IEEE Transactions on Visualization and Computer Graphics, 2017, 23(1):271-280
19 Koven J, Felix C, Siadati H, et al. Lessons learned developing a visual analytics solution for investigative analysis of scamming activities. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1):225-234
20 Scholtz J, Plaisant C, Whiting M, et al. Evaluation of visual analytics environments: The road to the visual analytics science and technology challenge evaluation methodology. Information Visualization, 2014, 13(4):326-335
21 Staheli D, Yu T, Crouser R J, et al. Visualization evaluation for cyber security: Trends and future directions. In: Proceedings of the 11th Workshop on Visualization for Cyber Security, 2014. 49-56
22 Zhao Y, Luo F, Chen M, et al. Evaluating multi-dimensional visualizations for understanding fuzzy clusters. IEEE Transactions on Visualization and Computer Graphics, 2019, 25(1): 12-21
23 Wei Y, Mei H, Zhao Y, et al. Evaluating perceptual bias during geometric scaling of scatterplots. IEEE Transactions on Visualization and Computer Graphics, 2020, 26(1):321-330