Benign User Activities that Trigger False Positives in Intrusion Detection Systems: An Expert Survey

Max Landauer[1(✉)], Florian Skopik[1], Markus Wurzenberger[1], Teodor Sommestad[2], and Henrik Karlzén[2]

[1] Austrian Institute of Technology, Vienna, Austria
{max.landauer,florian.skopik,markus.wurzenberger}@ait.ac.at
[2] Swedish Defence Research Agency, Linkoping, Sweden
{teodor.sommestad,henrik.karlzen}@foi.se

Abstract. Simulations of normal user behavior are integral parts of cyber exercises where training and testing takes place in simulated environments. Specifically, benign user activities are essential to generate background traffic during cyber exercises and to estimate false positive rates when evaluating intrusion detection systems. Even though many user automation tools are available, developers typically only consider valid and compliant interactions with systems and applications when defining the scope of normal user behavior models. However, real legitimate users sometimes behave in ways that are non-compliant, erratic, or otherwise deviate from expected norms, and thereby generate suspicious yet benign traffic that triggers alerts from intrusion detection systems. To identify common activities in the vast space of possible user interactions and to support the design of realistic user behavior models, we assemble a list of 17 user activities that are commonly associated with false positives. We assess the relevance and frequencies of these event types with respect to their perceived priority, intent behind them, responsible actor, and circumstances in which they become noteworthy, through likert scale analysis of an expert study with 62 domain experts. Our findings reveal diverse perspectives among respondents and suggest that the behaviors leading to false positives can vary significantly between organizations.

Keywords: intrusion detection systems · false positives · user simulation

1 Introduction

Cyber attacks are a permanent threat to organizations and individuals alike. As the sophistication and severity of these attacks increase, so too does the need for realistic environments that facilitate security training and testing. Such environments are often referred to as *cyber ranges* and widely used for cyber exercises that enable cyber capacity building and awareness training as well as security tests that enable evaluation of intrusion detection systems [3,14,23].

© The Author(s) 2025
B. Coppens et al. (Eds.): ARES 2025 Workshops, LNCS 15995, pp. 25–43, 2025.
https://doi.org/10.1007/978-3-032-00633-2_2

When designing cyber ranges, developers usually focus on building relevant technical infrastructures that are representative for real-world scenarios and launching attacks against these systems. An often underappreciated aspect of cyber ranges is the simulation of realistic legitimate users that interact with services available in the cyber range infrastructure [10]. However, normal user simulation can be useful to enable attacks (e.g., phishing) or represent security-critical business processes (e.g., information sharing). Even more important, its primary purpose in cyber ranges is to generate background events and noise; without them, detection of attacker behavior would be trivial since almost any activity occurring on such an idle infrastructure would obviously originate from attack executions and defeat the purpose of exercises. Some cyber ranges therefore rely on humans to generate legitimate traffic [23]. Unfortunately, this strategy is expensive, difficult to reproduce, and causes issues with data sensitivity [22]. For this reason, most cyber ranges rely on simulations of attacker and benign user behavior [3,23]. Past research has shown that simulation of realistic normal behavior is generally more challenging than attack execution, which can often be pre-recorded or scripted since attacks are generally only executed once during an exercise and little variation is needed. In contrast, benign user behavior needs to be complex and extensive to appear realistic. Moreover, the range of activities that normal users carry out differs greatly from one application to another, which complicates the generation of simulation models [7].

Many tools for user simulation have been proposed in the past, but almost all of them focus on web navigation such as sending and receiving mails [6,7,10] or enable interaction with clients and graphical user interfaces of office applications such as document editing software [5,8,12,16,20]. An important aspect of normal user simulation that is generally neglected in existing simulations is that benign users sometimes behave erratic or non-compliant to policies of organizations even though they do not have any malicious intentions [4,10]. For example, normal users could use tools that produce suspicious network traffic even when they are using them for entirely benign activities [9].

Such suspicious but nonetheless benign behavior patterns have the potential to trigger intrusion detection systems and generate alerts; more precisely, false positive alerts since they are not related to any actual attack. Previous studies have shown that the vast majority of alerts generated by intrusion detection systems compose of these false positives [1,9]. Moreover, it is well known that false positives cause unnecessary analysis efforts and influence the decision process of cyber analysts [1,13]. It is thus important to model normal behavior in such a way that it triggers realistic volumes and types of false positives. In case that normal behavior is overly simplified, evaluation results of intrusion detection systems tested within cyber ranges may yield too low false positive rates [10]. Additionally, the scope of normal user behavior may have a direct effect on true positives when learning-based detectors are evaluated. For example, detection models trained on scenarios where benign users sometimes use incorrect login credentials – as they would in real life – may be less likely to detect brute-force login attempts than in scenarios with simplified user behavior where only valid logins occur and invalid ones are outliers.

Many public and widely used evaluation data sets have been criticized for lacking realistic behavior patterns that trigger false positives [2,15,17,19]. Despite these observations, little research has been conducted on the modeling of suspicious yet benign user activities in the past. With this paper we attempt to resolve this gap by analyzing benign user activities and their relation to false positives triggered by intrusion detection systems. To this end, we first identify groups of relevant user activities and group them into event types. We then develop a questionnaire and conduct a user study with security experts to analyze priorities, origin, intent, and circumstances of each event type. We summarize the contributions of this paper as follows.

- An identification of 17 normal behavior event types that cause false positives.
- A user survey to assess their relevance for tests and exercises.

The remainder of this paper is structured as follows. Section 2 reviews the background of normal user behavior simulation. Section 3 describes our strategies to identify and enumerate relevant benign event types. We present the results of our survey to assess our event types in Sect. 4 and discuss the implications of our findings in Sect. 5. Finally, Sect. 6 concludes the paper.

2 Related Work

Past research has analyzed false positives and their influence on alert triage, i.e., the process where human experts assess whether alerts indicate active threats that require immediate action. Alahmadi et al. [1] interviewed personnel of security operations centers and found that almost all alerts (up to 99%) are caused by so-called benign triggers, i.e., non-malicious user behavior. Their study shows that the sheer volume of these false positives leads to fatigue and incorrect decision making of analysts. Layman et al. [13] conduct a controlled experiment and find that higher levels of false alerts decrease the precision of decisions made by analysts and increase their time on task. Ho et al. [9] analyze alerts collected from a university network. They find that more than 90% of alerts are false positives and more than 90% of these false positives are caused by organizational policies. For example, such policies could be set in place to prevent peer-to-peer communication, which will trigger alerts even when used for benign activities.

Given the prevalence of alerts generated by suspicious yet benign behavior in real and productive networks, we review publications involving normal user simulations to check whether such behavior patterns are also reflected in the scopes of these models. Wright et al. [22] use state machines to simulate web browsing and document editing. Guttman et al. [8] propose a user simulation tool that interacts with web browsers, mail clients, office applications, FTP clients, chats, and telnet. Grimmer et al. [7] simulate users through scripts that automate web navigation, file uploads, and entering of text into forms. Creech et al.

[5] simulate web browsing and document preparation in order to create a normal behavior baseline for anomaly detection. Lashkari et al. [12] use three user profiles that interact with Gmail, Facebook, Skype, and Whatsapp. To evaluate deep packet inspection approaches, Megyesi et al. [16] automate user interaction with scripts or macros for Microsoft Windows programs, QuickTime, Flash, Bittorent, HTTP, SSL, DNS, Skype, Google, and ICMP. Van Sloun et al. [20] simulate direct interaction with GUI elements of web browsers, office applications, and mail clients to generate data sets for host-based intrusion detection evaluation. Dutta et al. [6] create bots that log into their accounts, send emails, browse websites, and create and modify documents based on observations from volunteers who interacted with their system. None of the aforementioned publications consider unusual user activities when defining the scope of their normal behavior models. Landauer et al. [10] use state machines to automate SSH interaction as well as web navigation in mail platforms and file shares. Even though their benign user models include entering of incorrect login credentials, they explicitly state that they did not purposefully design the models to trigger false positives, but mention that they could be generated as side-effects. Evaluation of intrusion detection systems on their data sets indeed shows that several benign user activities trigger alerts, such as correct and incorrect logins [11]; however, it remains unclear if they are representative for alerts from real users.

The realization that most evaluation data sets generated through simulations do not involve sufficiently complex and diverse user behavior is not a new one. Already in the year 2000, McHugh et al. [15] found that the background traffic contained in the DARPA data set, which was current at that time, is not representative for real Internet traffic. They state that real data involves legitimate but odd-looking traffic that could or should trigger intrusion detection systems, which is not the case in the analyzed data set. Further research has confirmed that the traffic in the DARPA data set is unusually uniform [2]. Sangster et al. [17] mention the small volume and low diversity of traffic generated during network warfare games that is not representative for production networks. A more recent review of existing data sets is provided by Sharafaldin et al. [19], who also criticize the lack of traffic diversity and volumes across most data sets and deem them unsuitable for estimating false positive rates that are accurate for real-world cases. To overcome this problem, the same authors propose an approach that analyzes features of real network statistics, such as packet sizes and the number of packets per flow, and recreates synthetic data following these distributions [18]. Unfortunately, this approach only generates synthetic network traffic but does not actually interact with applications within cyber ranges. Thus, it does not produce any system log data required to evaluate host-based intrusion detection systems.

In summary, most publications focusing on benign user simulation only model typical system interactions and assume that some false positives will be generated as by-products. Some authors are aware of issues with non-representative benign behavior, but have only resolved this by replicating distributions of real network traffic. None of the reviewed publications explicitly mention benign cases that should or even could appear suspicious for intrusion detection. We therefore assemble a set of relevant benign behavior patterns in the following section.

3 Analysis of Benign User Behavior Patterns

This section analyzes activities that are part of benign user behavior and have the potential to trigger intrusion detection systems. We first explain how we identify relevant types of events and which of their properties are used by analysts during triage. We then enumerate and describe each event type and provide examples.

3.1 Identification of Relevant Types of Benign Events

The literature contains numerous detailed models of attacker behavior, which is typically goal-oriented, follows well-known kill chains, and relies on techniques that can be systematically categorized using MITRE ATT&CK.[1] We argue that a comprehensive overview of user behavior relevant to cyber security is significantly more challenging to produce. The range of possible user activities is vast, the ways in which users can interact with systems are virtually limitless, and the specifics depend heavily on the operational context as well as the applications and services available to users [7]. Rather than focusing on specific protocols or applications as it has been done in previous literature [9], we therefore aim to identify abstract classes of activities that typically trigger false positives in intrusion detection systems in IT networks across many organizations.

We carried out our search for relevant types of benign events in three phases. Thereby, we internally discussed, refined, and iteratively grouped gathered events throughout all stages. First, we collected an initial list of events based on surveys that analyze the consequences of false positives on system operators [21]. For example, the questionnaire of Alahmadi et al. [1] involves several sources of alerts that can also be triggered as part of normal user and administrator activity, such as unusual outbound network traffic, log-in red flags, large number of requests for the same file, and several more. Second, we reviewed the open-source Sigma[2] database of intrusion detection signatures and selected those entries that contain the "falsepositives" field, which can be used by security analysts to state why the respective signature could be triggered by benign users. For example, a signature that detects suspicious file modifications mentions "Admin changing file permissions" as a potential source of false positives.[3] Third, we relied on interviews with practitioners such as security analysts that we reached through personal contacts. The discussions with these domain experts validated our identified set of relevant events and helped us to improve event descriptions and come up with real-world examples.

[1] https://attack.mitre.org/.

[2] https://github.com/SigmaHQ/sigma.

[3] Title: Chmod Suspicious Directory (id: 6419afd1-3742-47a5-a7e6-b50386cd15f8).

3.2 Event Types

After completing our search for user activities that trigger false positives in intrusion detection systems as outlined in the previous section, we ended up with 17 distinct types of events. We enumerate all events in the following without any particular order and provide some examples.

(1) **File Permissions.** Users changing permissions for files, e.g., setting files as executable, making folders accessible for everyone on the machine, or creating shared folders.

(2) **User Privileges.** Users changing the privileges of users or groups, e.g., adding new users to machines or servers, creating new user groups, increasing or decreasing privileges of other users, or adding privileges to user groups.

(3) **OS Configurations.** Users changing the configuration of operating systems, e.g., manipulating scheduled tasks, changing service settings, changing system recovery and backup configurations, or changing preferred applications of file extensions.

(4) **Many Copy-Operations.** Users copying a considerable amount of files within short time periods, e.g., copying many files to a folder and compressing them, sorting of many files, synchronizing folders to file servers of organizations, or creating backup copies of entire disks or databases.

(5) **Custom Scripts.** Users executing scripts for task automation, e.g., scraping of internal sites to gather information, sorting of files using PowerShell or bash, or using record-features for task automation in office applications.

(6) **Network commands.** Users executing complex network commands, e.g., executing software to capture network traffic, adjusting settings related to network interfaces, iptables, or the local firewall, or executing commands such as ifconfig, ping, tracert, or netstat.

(7) **System Commands.** Users executing complex system commands on machines, e.g., executing tools in Windows Sysinternals, using Event Viewer (Windows) or journalctl (Linux), executing applications as admin (Windows) or root (Linux), or changing settings such as whitelists or antivirus.

(8) **Software Installation.** Users installing new software or updates of existing software, e.g., running apt-get (Linux) or Windows Update, pressing the update button of any software, running msi-files or exe-files to install software.

(9) **Credential Updates.** Users changing password or credentials, e.g., changing of user account passwords or resetting passwords of web accounts.

(10) **Server Login.** Users logging onto servers, e.g., using remote desktop connections, PsExec, or ssh to check network interfaces or logging onto domain controllers, web servers, or mail servers.

(11) **Server Checks/Cleaning.** Users checking or cleaning servers, e.g., monitoring server performance metrics such as CPU usage, cleaning temporary files with Windows Disk Cleanup, or sifting through system logs.

(12) Server Reboot. Users rebooting servers, e.g., rebooting after system updates or crashes.

(13) Inbound Connections. Users initiating connections to services or machines in organization networks from external sources, e.g., creating VPN connections into organization networks or connecting to terminal servers.

(14) Software Execution. Users executing software on systems, e.g., executing unknown software, executing software that repeatedly crashes, or executing debugging tools.

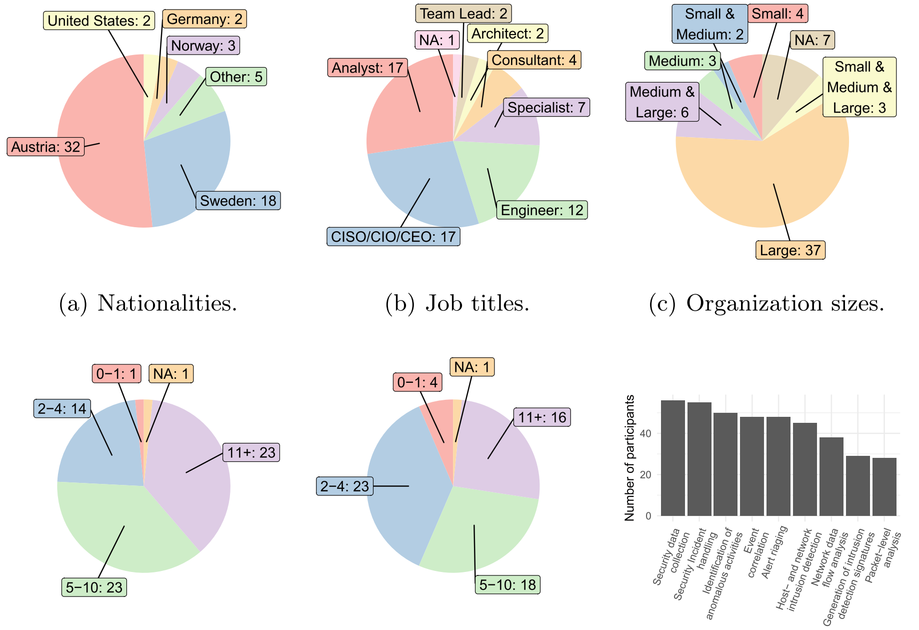
(15) New Network Device. Users adding new machines or devices to the network, e.g., adding devices with new MAC addresses to the organization network or adding new machines to Windows domains.

(16) Failed Login. Users failing to logon to servers, e.g., entering wrong passwords sufficiently many times to cause that their accounts are blocked or attempting to login with alternative user names.

(17) Outbound Connections. Users initiating connections to external services or machines from within organization networks, e.g., uploading data to external repositories, using peer-to-peer protocols, or using legacy software to connect to remote servers.

3.3 Survey

We developed a questionnaire to assess the relevance of all event types enumerated in the previous section. Given that we are interested in the selection of events for simulations and cyber exercises, we consider event types as relevant if they are frequent in comparison to other events, likely triggered by normal user behavior, and have high impact on analysts, e.g., by occupying a significant amount of time for triage. To cover all aspects, we split our questionnaire into four categories: **(Q1) Priority**, which asks whether events are (Q1.1) important to monitor, (Q1.2) cause security alerts, (Q1.3) occupy analysts' time, and (Q1.4) lead to investigations. **(Q2) Origin**, which asks whether events are caused by users with (Q2.1) typical or normal privileges, (Q2.2) special or admin privileges, or (Q2.3) unauthorized actors. **(Q3) Intent**, which asks whether events are eventually determined to be caused with (Q3.1) malicious intent, (Q3.2) good intent, or (Q3.3) pose a real threat. **(Q4) Circumstances**, which asks about the unusual circumstances that are required so that analysts occupy their time with analyzing that event. Specifically, for each event we ask about its subject (e.g., the user triggering the event), object (e.g., the affected machine), timing (e.g., whether the event occurs outside of usual working hours such as weekends), and frequency (e.g., the number of times the event occurs in short time periods). Responses to questions of all categories are entered on a five point likert scale with the following levels: *Never*, *Rarely*, *Sometimes*, *Often*, and *Always*. Participants also had the option to select *Not Applicable* as a response to any question. To avoid any bias that originates from the order in which event types are presented to the participants, we configure the survey to randomize their order each time the survey is started by a new participant.



(a) Nationalities.

(b) Job titles.

(c) Organization sizes.

(d) Years of experience in cyber security.

(e) Years of experience with security log/event analysis.

(f) Experience with various cyber security tasks.

Fig. 1. Demographics and background information of all 62 participants of our study.

We extended the survey with several questions about the background of participants, for example, their nationalities and experiences in the security domain, which we will describe in more detail in the following section. Moreover, on the first page of the survey, participants were informed about the purpose of this study and that their data and responses will be used and published in course of scientific research projects. The survey was conducted anonymously; however, participants had the option to provide their email address to enter a prize draw or to receive information about the study results. On the final page, we collected feedback about the survey itself.

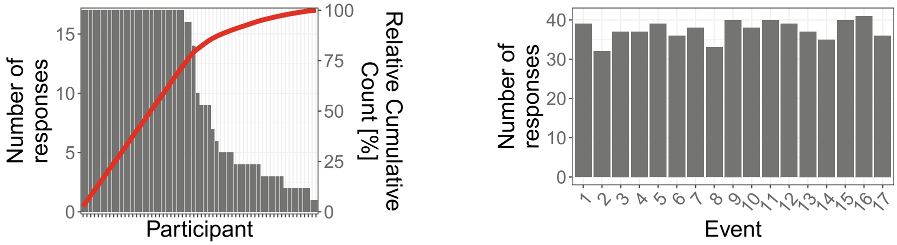
The survey was hosted from September to December 2024 on a publicly accessible website. We gathered participants by sharing a link to the survey with professional contacts, on cyber security mailing lists, and within research projects. Moreover, we advertised the survey during a large cyber security events that took place in Austria in September 2024 and in Sweden in December 2024. In the end we obtained responses from 62 unique participants who rated at least one event type. We provide detailed results gathered from the survey in the following section.

4 Results

This sections presents the results of our survey. We first analyze the demographics of participants and then present their responses in detail.

4.1 Participants

This section presents some background information about the participants. Figure 1a shows their nationalities, which reveals that around 80% of all participants are either Austrian or Swedish. This is not surprising given that the authors of this paper are located in these two countries and leveraged personal contacts as well as national events to spread the survey. Figure 1b summarizes job titles, which shows that even though the majority of participants have technical roles (e.g., security engineers or analysts), there are also several managers among the participants. Figure 1c plots the sizes of organizations monitored by the participants. As visible in the plot, the vast majority of them are monitoring large organizations with more than 250 employees or more than 50 million Euro turnover. Figure 1d and Fig. 1e show the experience of participants in cyber security and security log/event analysis respectively, which ranges from newly employed to more than 10 years. Figure 1f shows the exact cyber security tasks where participants are experienced in; as visible, there is a significant number of people with experience in each of the mentioned tasks. In summary, even though there is a strong tendency that participants are monitoring large organizations either in Austria or Sweden, the demographics indicate that our respondents comprise a diverse mix of highly skilled cyber security experts.



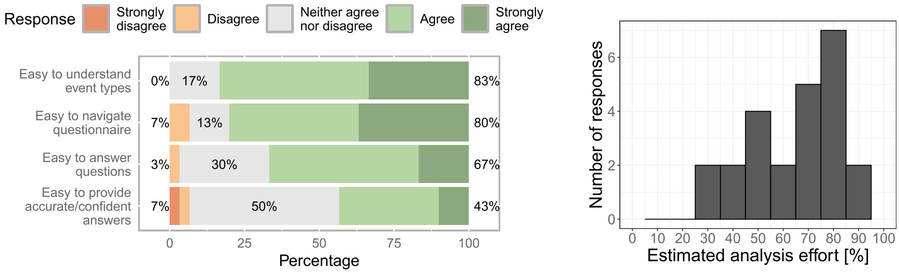
(a) Responses per participant.

(b) Responses per question.

Fig. 2. Overview of the number of responses.

4.2 Responses

We first provide an overview of the responses received from participants. Figure 2a depicts the number of responses per participant, where each response corresponds to one event type assessed by the respective participant. As visible in the plot, 27 out of 62 participants have completed the entire survey; others have only assessed some of the event types. We use the responses from all participants independent from the number of event types they assessed for our survey. Even though some participants dropped out during the survey, we expect roughly the same number of responses for each event type as they are presented to participants in random order. Figure 2b depicts the number of responses per event type; on average, we have around 37 responses per event type at our disposal.



(a) Confidence of participants.

(b) Coverage of events.

Fig. 3. Feedback received from participants.

On the final page of the survey we ask participants about their opinion on the survey itself and how confident they are about their answers. Figure 3a shows that most users have positive feelings about the survey itself and found it easy to navigate the questionnaire and answer the questions. However, even though the vast majority seem to have no issues understanding the event types, they are less confident about the accuracy of their answers with only 43% agreeing or strongly agreeing on the last question. To estimate how many of the relevant log events we covered in our enumeration, we also ask participants to rate how much log analysis effort they believe is spent on the presented events altogether, where 0% means that the events covered in the survey cause no effort at all and 100% means that all events that cause effort are present in the survey. Figure 3b plots the results on a histogram, which reveals that the distribution is relatively spread out but peaks around 70%–80%. The average of 67% and median of 74% indicate that some participants believe that some normal behavior events are missing from our enumeration, even though no suggestions for events were stated in the free-form text field.

4.3 Event Type Priorities

This section provides visualizations for the responses of each question in our survey, which we count separately for each event type. Figure 4 displays plots of likert scales for questions related to (Q1) event priorities. As visible in plot Q1.1, most participants agree that it is important to monitor most of the event types;

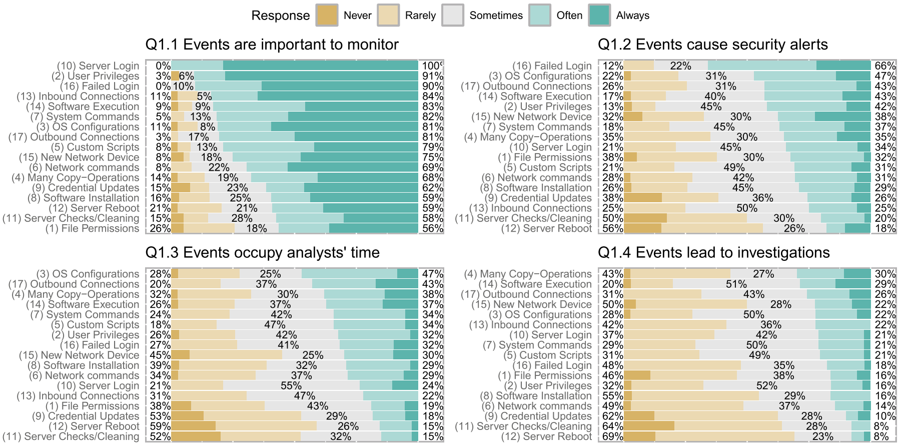


Fig. 4. Responses regarding priorities of analyzed events.

specifically, events related to *(10) Server Login* and *(2) User Privileges* yield high scores. There is less agreement among the participants for questions Q1.2-Q1.4, in particular, there are similar numbers of votes for the categories *Rarely*, *Sometimes*, and *Often*, while categories *Never* and *Always* are hardly used. Note that in each plot we sort the events by the aggregated number of votes for *Often* and *Always*; however, the order of events would be different when sorting by other categories, e.g., by aggregating votes of *Never* and *Rarely*. This caveat is also valid for the following likert plots in this paper and indicates that the displayed order of events should only be considered as a rough indicator rather than an exact assessment that allows to compare any two event types.

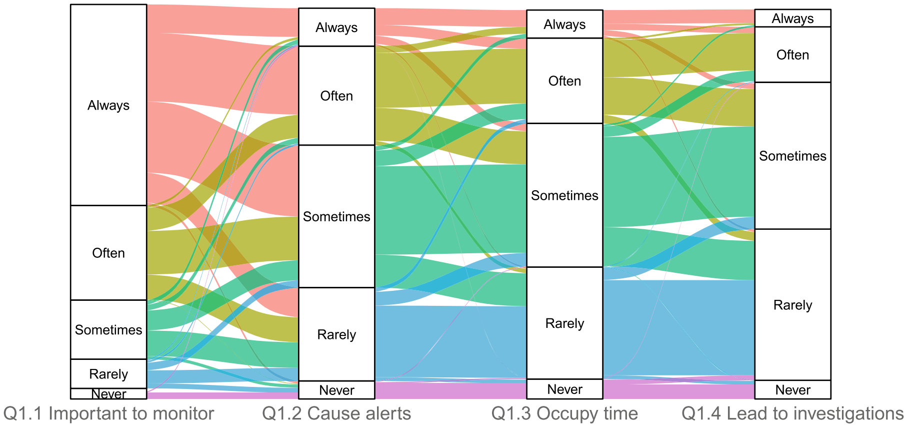


Fig. 5. Flow diagram of event priorities.

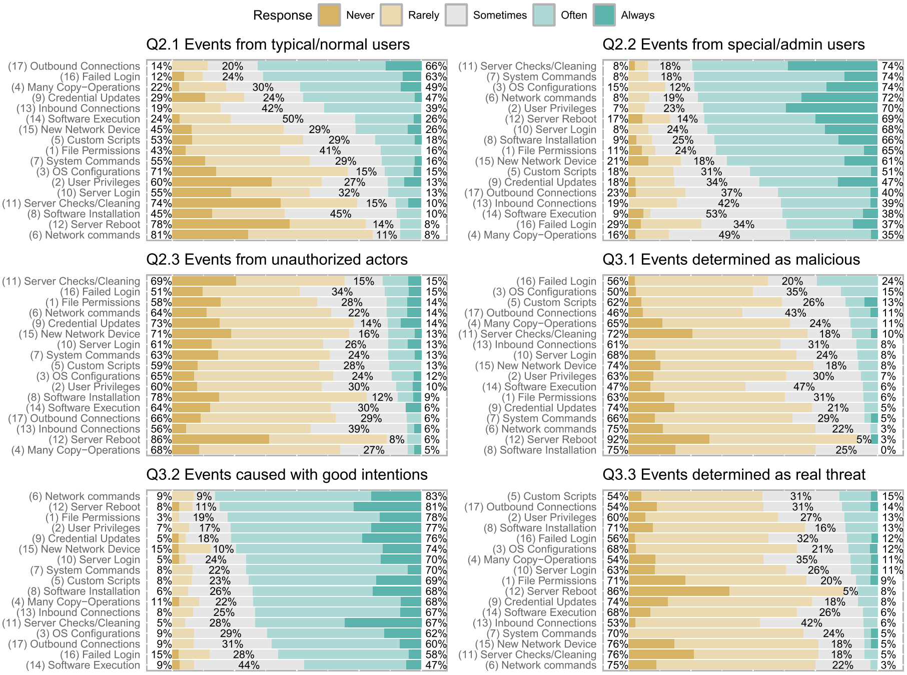


Fig. 6. Responses regarding origin and intent of analyzed events.

There are some logical dependencies between the four questions Q1.1-Q1.4. For example, those event types that are more likely to cause alerts (Q1.2) should also be the ones that are more likely to occupy analysts' time (Q1.3). To investigate these dependencies, we plot the responses as a flow diagram in Fig. 5. The plot confirms this assumption as most participants consistently selected the same categories for most event types when estimating their likelihood of causing security alerts (Q1.2), occupying analysts' time (Q1.3), and leading to investigations (Q1.4). However, the plot also shows that many of the events that are regarded as important to monitor (Q1.1) are not necessarily the ones that frequently cause alerts (Q1.2), since about half of the event types that should always be monitored only sometimes or rarely produce any alerts. On the other hand, those event types that are considered to be only sometimes or rarely important to monitor also hardly produce any alerts.

4.4 Influence of User Role

Figure 6 visualizes the responses to questions that deal with the origin of event types (Q2). Plot Q2.1 indicates that users with typical or normal privileges are most likely responsible for triggering false positives related to *(17) Outbound Connections*. This aligns with the findings from Ho et al. [9], who make certain application clients that generate outbound connections responsible for generating suspicious network traffic. Another highly ranked event type is *(16) Failed Login*. In comparison to normal users, plot Q2.2 shows that users with special or admin privileges are significantly more likely to frequently trigger false alerts corresponding to diverse event types. This is reasonable, since only privileged users have or should have the abilities and responsibilities to carry out activities that could trigger some of the highly ranked event types, such as the ones related to *(11) Server Checks/Cleaning* or *(3) OS Configurations*. Finally, plot Q2.3 shows that all event types are roughly equally likely to be generated by unauthorized actors, with low frequency overall in comparison to the two previous cases.

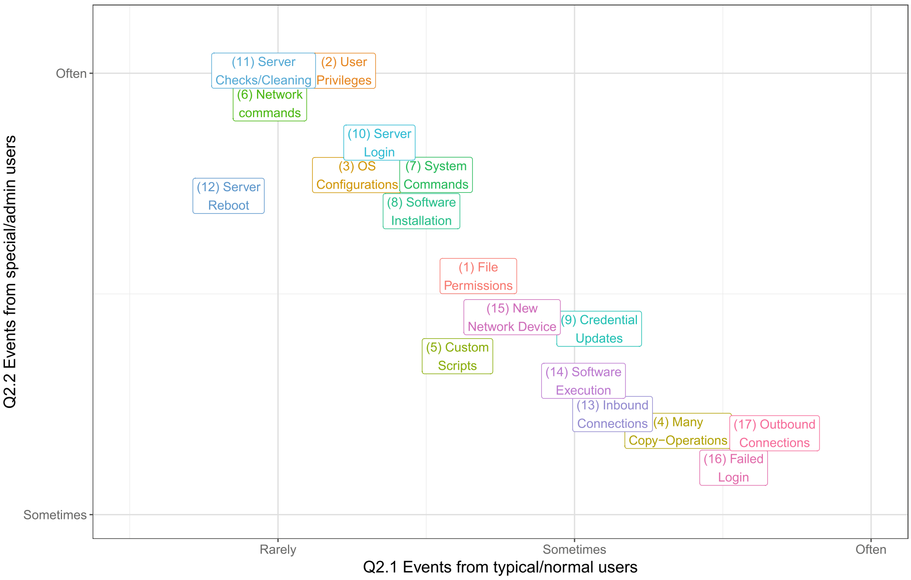


Fig. 7. Likelihood of events originating from privileged or unprivileged users.

Figure 7 allows to better understand the relation between event types generated by unprivileged (Q2.1) and privileged (Q2.2) users. The scatter plot places labels of each event type at the mean position of all responses, where we encode the five categories *Never* to *Always* with the numeric values 1 to 5 in order to compute the means. Despite some disagreement among the participants (cf. Figure 6), the means are roughly located along a diagonal line that stretches from the top left part of the plot, where event types that are often generated by privileged users but rarely by unprivileged users, to the bottom right, containing event types that are often generated by unprivileged users.

4.5 Intent Behind Events

Figure 6 also provides the plots corresponding to the intentions behind events that cause analysis efforts (Q3). The 17 event types included in the questionnaire were selected based on their perceived tendency to trigger false positives and unnecessary analysis efforts. Consistent with this, events requiring analysis were generally found to stem from good intentions (Q3.1) and were ultimately deemed non-malicious (Q3.2). Only a minority of events were eventually determined to be caused by actual threats (Q3.3). These results align with the findings from Alahmadi et al. [1], who conclude that the vast majority of events handled in Security Operation Centers are false positives that are not actually related to cyber attacks. They also support the notion that the 17 event types are relevant to consider in cyber security exercises and tests where analysis effort and false alerts are important.

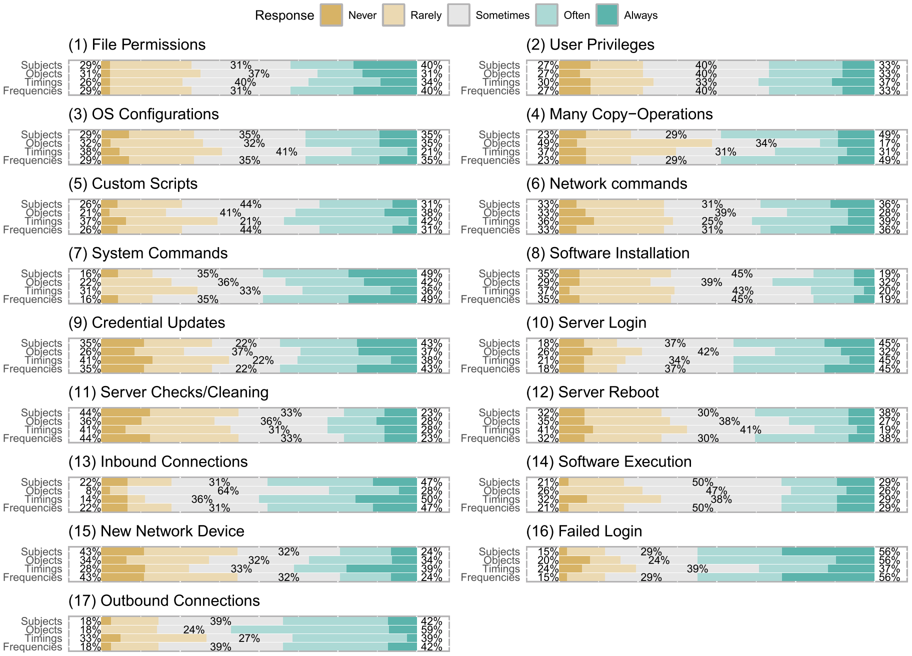


Fig. 8. Responses regarding the circumstances of analyzed events.

4.6 Circumstances of Event Occurrences

The final group of questions (Q4) concerns the circumstances of event occurrence and its relevance for analysis and triage. Figure 8 visualizes the responses for each of the four context attributes (subject, object, timing, frequency) separately for each event type. Once more, disagreement among participants is substantial and circumstances of most event types receive a fair share of responses across all categories. Thus, event simulations should consider the subject causing the event (e.g., user account), the object involved (e.g., server), the time (e.g., at night), and frequency (e.g., a burst). All of these can make events more or less interesting to an analyst.

5 Discussion

Realism of normal behavior simulations is critical when designing cyber exercises and test environments that aim to reflect real-world settings. In the sections below we discuss what event types to consider, which details to consider when events are simulated, and recommendations for future research.

5.1 Relevant Event Types

The study presented in this paper identified 17 abstract event types that can be associated with false alerts and are common across many organizational IT networks. We conducted a survey among practitioners to validate the relevance of this list and estimate how often these event types cause analysis efforts in operational networks. The results of our study suggest that all of the 17 event types can trigger intrusion detection systems even when responsible users have good intentions (cf. Sect. 4.5), and that the resulting false positives require manual review from security analysts (cf. Sect. 4.3). Thus, all event types stated in Sect. 3.2 deserve consideration when designing cyber security exercises or tests that involve challenging background events. Furthermore, as seen in Sect. 4.2, the event types covered in this survey appear to cover about two thirds of all events to consider.

We acknowledge that the presented rankings of event types (cf. Sects. 4.3–4.6) serve only as rough indicators and that no definitive trends can be drawn from our results. The reason for that is the substantial dispersion of responses observed for most questions of our survey. Interestingly, even though most participants claimed to understand the questionnaire and many exhibit confidence in their responses, their experiences with false positives seem to be highly individual. In fact, several participants use the free-form text field at the end of the survey to point out that they have made various experiences with false positives and that their occurrences depend on several factors. Foremost, they mention that the types and frequencies of false positives strongly depend on the tolerance of deployed detection systems. In addition, they state that alert assessment and triage often relies on domain knowledge about the networks, e.g., alerts from systems where no users should be active are automatically regarded as critical. One participant recognizes that even though they provided answers from their own perspective, they are aware that other analysts handle false positives differently. Based on the received responses and feedback, we thus conclude that security analysts experience different types of false positives at varying frequencies, which possibly depends on their area of expertise. Moreover, entire security teams may deal with different false positives depending on specific detection policies that vary across monitored organizations.

5.2 Simulating Events in Cyber Ranges

Our study provides guidance for developers in designing user behavior models that produce false positives during cyber exercises. Specifically, our results allow to decide which of the suspicious yet benign event types selected for implementation should be executed by network administrators or normal users (cf. Sect. 4.4). Moreover, the results of our study suggest that all four types of circumstances (subject, object, timing, and frequency) covered in the survey appear to be important across all event types (cf. Sect. 4.6). In alignment with the simulated security scenario, simulations should therefore include some variation to these attributes, e.g., by triggering false positives at unusual times or in bursts.

5.3 Recommendations for Future Research

Overall, the feedback on our questionnaire is positive. Most participants reported to understand the questions asked and many expressed confidence in their responses (cf. Sect. 4.2). Thus, the questionnaire may be suitable for reuse in future studies. To support future research on this topic, we point out some improvements that we derive from the free-text feedback. Two participants mention that they find the five point likert scale insufficient and propose to extend it to a seven point scale that additionally includes *Very Rarely* and *Almost Always* as choices. As an alternative, we suggest to use pairwise comparisons of event types to avoid having to deal with options such as *Sometimes* altogether, which can have different meaning to each participant. One participant states that the presented event types are too generic for accurate assessment; specifically, they mention that the impact of alerts related to *(6) Network commands* heavily depends on the exact command that is executed. It thus also stands to reason to further subdivide the list of events proposed in this paper or add technical aspects as contextual attributes (cf. Sect. 4.6), such as geographic location or connection details [13].

The sample in this survey is biased toward Europe, in particular, Austria and Sweden (cf. Sect. 4.1). However, similar studies could be applied to other populations to yield more generalizable results. With larger sample sizes, it may also be possible to identify differences between populations based on factors such as industrial sector, organizational size, and other characteristics.

Realistic simulation of normal user behavior remains challenging since the number and diversity of benign activities as well as the ways in which they can potentially be executed seem virtually endless. To handle this complexity, we recommend to develop and benchmark normal behavior models based on real user data collected from productive systems and avoid any simplifications, i.e., also recognize rare and suspicious activities as essential parts of the model. In particular, such comparisons could base on the number of diversity of alerts generated by a diverse set of intrusion detection systems, including both network-based and host-based as well as signature-based and anomaly-based approaches. Finally, we believe that future work could investigate to what degree large language models are useful to reduce manual effort when designing large and complex behavior models.

6 Conclusion

Simulations in cyber exercises and tests often neglect suspicious yet benign user activities, even though they are relevant sources of false positives in real-world networks. In this paper we therefore review existing literature, analyze detection signatures, and interview practitioners to enumerate 17 abstract event types that correspond to benign activities that often trigger intrusion detection systems. We then carry out a user study with 62 participants to assess the prevalence of these event types, their likeliness to influence the analysis and triage procedure, and whether they originate from benign users with or without special system privileges. The results of our study indicate that some event types are more frequently associated with false positives than others, such as changes to user privileges that are often carried out by privileged system administrators. At the same time, we observe significant disagreement among the participants of the study, which suggests that experiences with false positives are highly individual and depend on the operational context.

Acknowledgment. This work has received funding from the European Union - European Defence Fund under GA no. 101103385 (AInception). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. The European Union cannot be held responsible for them.

References

1. Alahmadi, B.A., Axon, L., Martinovic, I.: 99% false positives: a qualitative study of soc analysts' perspectives on security alarms. In: Proceedings of the 31st USENIX Security Symposium, pp. 2783–2800 (2022)
2. Brown, C., Cowperthwaite, A., Hijazi, A., Somayaji, A.: Analysis of the 1999 darpa/lincoln laboratory ids evaluation data with netadhict. In: Proceedings of the Symposium on Computational Intelligence for Security and Defense Applications, pp. 1–7. IEEE (2009)
3. Chouliaras, N., Kittes, G., Kantzavelou, I., Maglaras, L., Pantziou, G., Ferrag, M.A.: Cyber ranges and testbeds for education, training, and research. Appl. Sci. **11**(4), 1809 (2021)
4. Cram, W.A., Proudfoot, J., D'Arcy, J.: Seeing the forest and the trees: A meta-analysis of information security policy compliance literature (2017)
5. Creech, G., Hu, J.: Generation of a new ids test dataset: time to retire the kdd collection. In: Proceedings of the Wireless Communications and Networking Conference, pp. 4487–4492. IEEE (2013)

6. Dutta, P., Ryan, G., Zieba, A., Stolfo, S.: Simulated user bots: real time testing of insider threat detection systems. In: Proceedings of the Security and Privacy Workshops, pp. 228–236. IEEE (2018)
7. Grimmer, M., Röhling, M.M., Kreusel, D., Ganz, S.: A modern and sophisticated host based intrusion detection data set. IT-Sicherheit als Voraussetzung für eine erfolgreiche Digitalisierung **11**, 135–145 (2019)
8. Guttman, R.D., Hammerstein, J.A., Mattson, J.A., Schlackman, A.L.: Automated failure detection and attribution in virtual environments. In: Proceedings of the Symposium on Technologies for Homeland Security. pp. 1–5. IEEE (2015)
9. Ho, C.Y., Lai, Y.C., Chen, I.W., Wang, F.Y., Tai, W.H.: Statistical analysis of false positives and false negatives from real traffic with intrusion detection/prevention systems. IEEE Commun. Mag. **50**(3), 146–154 (2012)
10. Landauer, M., Skopik, F., Frank, M., Hotwagner, W., Wurzenberger, M., Rauber, A.: Maintainable log datasets for evaluation of intrusion detection systems. IEEE Trans. Dependable Secure Comput. **20**(4), 3466–3482 (2022)
11. Landauer, M., Skopik, F., Wurzenberger, M.: Introducing a new alert data set for multi-step attack analysis. In: Proceedings of the 17th Cyber Security Experimentation and Test Workshop, pp. 41–53 (2024)
12. Lashkari, A.H., Kadir, A.F.A., Taheri, L., Ghorbani, A.A.: Toward developing a systematic approach to generate benchmark android malware datasets and classification. In: Proceedings of the International Carnahan Conference on Security Technology, pp. 1–7. IEEE (2018)
13. Layman, L., Roden, W.: A controlled experiment on the impact of intrusion detection false alarm rate on analyst performance. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 67, pp. 220–225. SAGE (2023)
14. Leitner, M., et al.: Enabling exercises, education and research with a comprehensive cyber range. J. Wirel. Mob. Networks, Ubiquitous Comput. Dependable Appl. **12**(4), 37–61 (2021)
15. McHugh, J.: Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. ACM Trans. Inf. Syst. Secur. **3**(4), 262–294 (2000)
16. Megyesi, P., Szabó, G., Molnár, S.: User behavior based traffic emulator: a framework for generating test data for dpi tools. Comput. Netw. **92**, 41–54 (2015)
17. Sangster, B., O'connor, T., Cook, T., Fanelli, R., Dean, E., Morrell, C., Conti, G.J.: Toward instrumenting network warfare competitions to generate labeled datasets. In: Proceedings of the 2nd Cyber Security Experimentation and Test Workshop (2009)
18. Sharafaldin, I., Gharib, A., Lashkari, A.H., Ghorbani, A.A., et al.: Towards a reliable intrusion detection benchmark dataset. Softw. Networking **2018**(1), 177–200 (2018)
19. Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy, pp. 108–116. SciTePress (2018)
20. van Sloun, C., Wehrle, K.: Poster: Vulcan–repurposing accessibility features for behavior-based intrusion detection dataset generation. In: Proceedings of the Conference on Computer and Communications Security, pp. 3543–3545 (2023)
21. Tjhai, G.C., Papadaki, M., Furnell, S., Clarke, N.L.: Investigating the problem of ids false alarms: an experimental study using snort. In: Proceedings of the 23rd International Information Security Conference, pp. 253–267. Springer (2008)

22. Wright, C.V., Connelly, C., Braje, T., Rabek, J.C., Rossey, L.M., Cunningham, R.K.: Generating client workloads and high-fidelity network traffic for controllable, repeatable experiments in computer security. In: Proceedings of the International workshop on Recent Advances in Intrusion Detection, pp. 218–237. Springer (2010)
23. Yamin, M.M., Katt, B., Gkioulos, V.: Cyber ranges and security testbeds: scenarios, functions, tools and architecture. Comput. Secur. **88**, 101636 (2020)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.