Online Binary Models are Promising for Distinguishing Temporally Consistent Computer Usage Profiles

The paper deals with certain characteristics of a user such as keystrokes, network usage, mouse movements etc to act as an identifier and whether these patterns would be enough for an Online Binary Model to identify a user. To formalize this research topic, the authors focused on 3 questions to answer with this paper. Firstly, are the usage profiles (these habits) consistent over time? Secondly, do users have unique profile, that is, are these profiles distinguishable from each other? Finally, what features are the most important for constructing a unique profile?

The domain where these studies can be directly applied to is Continuous Authentication. The user would be consistently monitored and authenticated using the above said user profiles, even after the conventional initial entry point like, password based systems. If the activity of the user differs significantly from their regular usage, it would be flagged as a potential security risk.

The first question the researchers focused on is whether the computer usage profiles stay consistent over time. Quantitative methods were employed to see whether the time series describing the profiles were random or periodic. Furthermore, after checking the periodicity the authors tried to identify the period. In order to do these, the techniques utilized were Periodogram and Autocorrelation function. With the concepts of sample entropy (the more unstructured the times series, the larger the entropy) and hurst exponent (measure of long-term memory of a time series), the authors were able to identify a 24 hour period where 83.9% of the participants repeated their daily computer usage pattern.

The next question to answer was whether these user profiles can be used to uniquely identify a user? 31 users on Windows 10 were observed for an 8-week period where their mouse, process, network and keystroke events were monitored. The experiment tested different duration of a sliding window of time - 1, 2, 5, 10, 30, 60 minutes. It was observed that the profile was consistent over time but with irregularity and hence could be utilized to uniquely identify users. Out of the different features, the network related ones contributed the most towards distinction between the users. The experiments were conducted using online binary models as well as offline ones, and the best performance was shown by the online models with Adaptive Random Forest being the best, followed by Half-Space Trees. Considering the patterns, online binary models are preferred due to their better performance and robustness to temporal changes.

Out of the different duration of time windows tested out it was found that a certain amount of data is required to recognize the user. More specifically, a study found out that at least 10 minutes of data is required. This is known as the cold-start problem, and this time window duration might change according to the dataset and context. However, the longer the duration needs to be, more time is available for the anomaly or the malicious actor to cause damage before being detected.

Discussion

- Even though authentication and authorization are talked about together, the concepts are distinct. Authentication is about identifying who the user is, whereas authorization is about enforcement and ensuring that the user can access only the resources meant for him/her. The traditional form of authentication is through passwords, which is something that only you know. There are different forms and layers of authentication focusing on different factors. For example, biometrics such as facial recognition or fingerprints is something you are whereas hardware keys are something you have. Multiple factors like these together form the Multi Factor Authentication Systems.
- One categorization of models in this context would be the online models and the offline models. While offline model has an overall view of everything that is happening in the system over time, an online model only has access to a small window of data. An online model, collecting enough data over time and then analyzing it together in an offline manner is not truly an online model. This is a strategy that many malware detectors use to falsely advertise their product as real-time.
- The methodology would be different between online models and offline models, especially for evaluation. In online the analysis of the results should not be based on the results produced a single instance of the sliding time window. Instead, the results should be correlated together to get conclusions out of it. For example, a model failing in one window during a day could be considered as a failure for the entire day even if it was successful for all other windows.
- For continuous authentication the threat model would have to be formalized properly. For example, user B helping out user A to debug an issue would be using his system for a short duration which the CA system would flag as malicious. So it should be decided if these kinds of behaviors should be allowed or considered.
- The paper identifies 10 minutes as the minimum duration required to collect enough information to judge whether a user is who he claims to be. However, realistically 10 minutes is a long duration and the attacker would be able to cause a lot of damage within these 10 minutes. So the question arises about the practicality of the concepts presented in the paper, However, the results must be viewed in the right light. This is one of the first papers in this direction of research and hence should not be held to higher standards in terms of results. With further research in the domain, the 10 minute window could be further brought down to realistic duration which could then be converted to practical implementations.
- From an attackers point of view, the challenge they have to focus on now would be to try and imitate the user in order to gain persistent access to the system. This would mean that the attackers would be forced to collect enough information to achieve this, causing them to focus only on the high profile or high value targets. In other words it would be infeasible from an attackers point of view to target the common user.
- Which would be the easier problem to solve - malware detection or user detection? The chances are for the user detection to be solved more than the malware detection problem. This is because the amount of variations in the user patterns are limited by the constraints of the real world, whereas in the digital world malware has a wide variety of profiles that could be achieved.

- Another interesting point to notice is how similar the work that the researchers did is to real keyloggers. The researchers installed software in the system of a user base and collected details such as the keystrokes, the mouse movements, the domains visited and the processes in the operating system. The only and main difference between these activities from the ones carried out by malicious actors is the intention. This shows how tough the defender's job is when they try to estimate the intention behind a user's action based on proxies.
- The question also arises whether it is ethical to collect information such as user profiles for research? Research such as these have additional legal complications and guidelines to be ensured. First of all, the authors would have to get the permission from the users. Moreover the users would also have to get the permission from groups like the Institute Research Board where they have to disclose what data they are collecting, what the research is about, what they are going to use the data for, etc. Also, this paper is another one in the line of papers that were discussed recently that showed an involvement of humans in the research process.
- In the security pipeline, most of the discussion till now was in the stage of detection. The stage of detection is towards the end, when most of the other phases such as spreading awareness and preventing exploits failed. The solution provided in this paper could be considered to reside right before the typical detection stage in a security pipeline. The reason for this is that the solution aims to prevent malicious actors from gaining access to the system, in order to cause damage to it or infect it with malware. The solution might also detect and flag the activities of malware too, but that is more of a side effect than intended use case. Another factor to consider is that, normally the detection models check against the patterns they have to match whereas in these kinds of models, it is outlier detection that is focused on where the model checks if the patterns deviate from what it already knows.
- The authors make the claim that binary classifiers is best suited for this task. However, they make a mistake here where the binary classifier classifies into the legitimate user or all other users as the second class. This is basically the concept of one class classifier, just implemented in a binary classifier format. For their one class classifier they utilized a linear SVM. So the comparison was between a non-linear RFC and a linear SVM instead of between a binary classifier and a one class classifier. For this reason the conclusion would be that the non linear model performed better and not that binary classifiers are better suited for this task.