COLLOQUIUM REPORT

"A Provenance approach to improve modern system transparency"

Submitted By: Agasthya Harekal(adh5677@psu.edu)

The talk was about improve system transparency and detect attacks by investigation. Initially the talk started with modern systems problem where speaker emphasised that as the system grows more complex, the investigation into attacks on the system is extremely difficult and may cause huge damages. According to a report, there is a cyber attack every 39 seconds, and 93% of large corporations and 87% of small businesses have reported many kinds of cyber breach. They cause billions of dollars of loss to companies and affects many user accounts. The solution to this problem is provenance analysis, where logs of processes, files is retrieved. Based on the collected data, we perform various analysis using casual graphs. Machine learning on these casual graphs can be done for analysis. Just usage of casual graph can be done to detect root cause of attacks and the damages to the system. The attack investigation has two stages, data collection and data storage. Data collection involves collection of data. Log entries of audit log are collected and are used to generate casual graph. Next, speaker spoke about dependency analysis, where it was found that there are huge number of false dependencies for simple event that occurred in the system. For a single download of file, it is impossible to trace the ip from where it downloaded using a browser. The solution was found to this problem, where partitioning of execution, where division of runtime was done into four different nodes and based on this partitioning, one can identify the source of the file. Division of Firefox events can be done using logs, sockets, tabs. System work flow was explained, where log files initially were parsed into normalised record, then datalog engine was fed with these records. The catalog engine had pre-defined rules, when an event occurs, a symptom event starts the attack investigation. This is also fed into the catalog engine along with other components that is records and pre-defined rules in-order to produce casual graph. Later details of log normalisation was explained where audit log and firefox log was explained briefly. Next, log fusion was explained where inference rules of the log had to be altered in-order to get the right dependency relations and to trace the download of the file.

The data storage problem was explained. Huge amount of data has to be collected for investigation. Example, we have firefox logs accounted for 5 minutes. Then, lot of data has to be collected. Sometimes, attacks lasts for long duration of time, hence lot of space is required to store the logs. Here, many solutions have been found.First, we reduce the logs by reduction, where removal of repeated events in the log has to be done. Another solution is compression, where lossless compression using GZIP compression, also the deep learning based network DNN

based compression and ELISE log lossless compression which achieved the best 0.06% compression ratio. The ELISE log compression was engineered by the researchers. This model can be divided into several stages as follows, first is the log formatting stage where logs from different sources are converted to JSON format. Next is the Pre-processing 1 stage where four different methods were identified to process the log to convert the log into numerical values. Next, in preprocessing 2, the session data, enumerations is compressed into numerical values. In the preprocessing 3, the frequent file paths are stored separately, and then referenced in the log. In the next stage, encoder is trained using DNN model. In the next stage, Data Compression is done using traditional arithmetic encoder.

Cyber attacks have become a huge problem in the networking infrastructure world and also in the software world where software, often proprietary are stolen by hackers. Even though, lot of firewall mechanisms are in place, modern hackers are able to use sophisticated technologies in order to hack into the systems to steal data, software, money or also down the network infrastructure. In this context, the speaker spoke about how transparency in the complex systems built would help in investigation of network attacks which is currently very time consuming due to the complexity of the systems. Cyber attacks have to be properly investigated in order to detect such cyber crimes and may be catch the culprits. Due to the lack of transparency, detection of such attacks have become very difficult and also investigation into such attacks requires lot of data collection to be done and also huge amount of data has to be stored. To tackle such problems, the few solutions were discussed. Lot of confidential government information is stolen by hackers which nowadays always makes the news. Countries like United States have been attacked by cyber bullies where infrastructure such as gas, mobile payments were found to be not functional as they rely on the network which was attacked by them. Recently, trading platform's security was compromised where user information was hacked into. Social media platforms like Facebook and email platform like yahoo are not immune to attacks by hackers and many times, hackers get access to confidential user data. Global action into cyber attacks is required where a common cyber security front or alliance has to be created to face the attacks from all over the world. According to IBM's cost of data breach report, average time to detect and contain a cyber security attack is 287 days. Continuous monitoring is required in all tech based companies and businesses to detect and confront hackers. Each and

every sector in the society which is digitally connected is vulnerable to hacking hence, new security mechanisms have to researched to make these networks less vulnerable. Although digitisation of all infrastructure is going on rapidly, the required security mechanisms have not been deployed to meet the security needs. This has to be addressed in a very expedited pace to make these infrastructure secure.