N+1 Decision Trees For Attack Graph

Tawfiq S. Barhoom and Lamiya M. EL_Saedi
Faculty of Information Technology
Islamic University of Gaza
Gaza, Palestine
{tbarhoom, lalsaedi}@iugaza.edu.ps

Abstract— Attack Graph is very useful technique for administrator to map the system vulnerabilities, the information mapped are attack’s goals and paths. In this paper we introduce a novel way to draw an Attack Graph, by using Decision Tree to preprocessing the vulnerabilities information collecting from government institution using NESSUS tool. Decision Tree is a supervised learning classification technique represent paths and text in the nodes and on the edges for verifying the easy understand vulnerabilities. The tree used is very useful in the way of generating the graphs. The graphs are N+1: N for each attribute and one for full graph. This way simplify the way the administrator to learn the situation by minimize the size of graph and then evaluate the system vulnerabilities.

Index Terms—Data Mining, Decision Tree, Security, Attack Graph, Graph.

I. INTRODUCTION

Attack graph is a way used by administrator to discovery and analysis network attack models. And used to specify and determine how vulnerable their systems and what security measures to deploy to defined their systems. Attack graphs can be used as a useful tool in several areas: in Network security including (Intrusion detection, defense, and forensic analysis). So, the administrator used attack graph to generate information and to make decisions. The first reasons achieved by ask "what attacks is my system vulnerable to " and "how many different ways can an attacker reach a final state to achieve his goal?" The second reason achieved by ask "which set of actions should I prevent to ensure the attacker can't achieve his goal?" or "Which set of security measures should I deploy to ensure the attacker can't be achieved his goal?" The paths in attack graph represent the scenario of attacker to achieve his goal. These paths called actions. [5].

You can use Attack graph as a system helps an administrator to find an answers for some questions like “if an attacker start from this point what is the goal for him or are there another path can he use it to verify his goal?” [3].

"Attack graph is a tool to analyze multi-stage, multi-host attack scenarios in a network. It is a complete graph where each attack scenario is depicted by an attack graph which is essentially a series of exploits” [4].

To generate an Attack graph is very difficult process for any user because you can't limited and predicted the paths attacker that he uses to verify his goal. Especially in Network, because it's able to be increase and very complicated. So, the first thing is you must understand the security problem clearly to minimize the security problem. Second put automatic solutions according to the security problem and what configurations need to make the modification is easier for user [2].

In this paper we suggest to use Data Mining methods to generate attack graph. Which is a novel way used in every environment of science. Here we apply it in security science to proof that the data mining able to meet your entire request at, via some knowledge for your work, and some knowledge and experience for how to use Rapid Miner to generate your idea. You can work in two areas for learning in Data Mining, supervised learning and unsupervised learning. There are many stages in Data Mining such as preprocessing, classification, association rule, clustering, outlier and evaluation. For each stage there are many methods and models depends on variant algorithms to apply what you need via drag the icons name of the method on the work area which represent as box titled with name of method you are chose. Then connect these boxes with line from one box to another.

In this paper we represents related works, government institution network structure, basic information about decision tree, working with data mining, actual work, discussions and conclusions.

II. RELATED WORKS

1. Study of Generating Attack Graph based on Privilege Escalation for Computer Networks (2008): in this paper, the authors worked very hard to generate a general system that associate between the types of vulnerabilities and the kinds of attack that can be occur on those vulnerabilities. They insert all information in five tables by using RDBMS and
establishing a relation between these tables to predict the attack graph based on privilege escalations. The tables’ names are: Vulnerability Type, Attack type fact, Attack Type Prerequisite, Attack Type Consequence and VulName to PID. Table Vulnerability type predict consists of predicates possibly used to represent the prerequisites and consequents of attacks. Attack types are stored into table Attack Type Fact. Attack Type Prerequisite and attack Type Consequence have the same structure, but are used to keep the prerequisite and consequences of known attack types. VulName to PID is designed to automatically achieve transformation from discovered vulnerabilities to atomic predicates and lays the information for later processes. Also introduce three kinds of attack graph (Network analysis of vulnerability) Model checking, logic programming and exploit-dependency graph search algorithm. [7]

2. Network Security Evaluation through Attack Graph Generation (2009): this paper helps me to identify the attributes that needs to draw the attack graph. These attributes are, "1- Computer and Network H= \{h1, h2 … hm\}, to represent these devices for example, computers, routers, switches. HOSTID is the unique identifier of host on the network, it can be the IP address or host name. OS is the type and version of operation system. SVCs is the list of network service types with respective network port numbers which describes the services on host and the information on service monitor ports. VULS is the host computer vulnerability list which may include the security bug information of installed software or environment misconfigures information, and is presented by its CVE ID.

2- User Privilege:

<table>
<thead>
<tr>
<th>Privilege class</th>
<th>Role description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROOT</td>
<td>System administrator, managing all system resources.</td>
</tr>
<tr>
<td>USER</td>
<td>Any general system user, which is created by administrator.</td>
</tr>
<tr>
<td>ACCESS</td>
<td>Remote visitors which may access network services</td>
</tr>
</tbody>
</table>

Table (1): define the role for each privilege class.

3- Connecting Relationship

The Internet is structured based on TCP/IP protocol family

<table>
<thead>
<tr>
<th>Protocol Layer</th>
<th>Link Relation Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Layer</td>
<td>HTTP FTP</td>
</tr>
<tr>
<td>Translation Layer</td>
<td>TCP UDP</td>
</tr>
<tr>
<td>Network Layer</td>
<td>ICMP</td>
</tr>
<tr>
<td>Data Link Layer</td>
<td>ARP</td>
</tr>
</tbody>
</table>

Table (2): define the kinds of protocol in each layer of network.

4- Attack Rule:

<table>
<thead>
<tr>
<th>Preconditions</th>
<th>Postconditions (results set)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Src_privilege</td>
<td>Rslt_privilege</td>
</tr>
</tbody>
</table>

Table (3): define the pre-condition and post-condition for each privilege.

- **Src_privilege** represents the lowest privilege which attacker should have on the host where the attacks are launched. **Dst_privilege** represents the highest privilege which attacker should have on the object host. **Vuls** represents the vulnerability which the attack rule depends on. **Protocols** describe the needed connection relation between attack host and object host. **Rslt_privilege** describes the privilege which attacker can get on object host after an attack is successfully completed. **Rslt_protocols** is the network protocols set which is added by attacks. If the attacked host can use the network protocols in this set to access a host on the network, the current attacking host can get the ability to access this host. If the attack rule doesn't influence the current network connection relations, **Rslt_protocols** will be an empty set. When **Rslt_protocols= {all}**, this represents that the current attacking host can get the attacked host's total ability to access the object network. **Rslt_vuls** is the newly added vulnerability set on attacked host after attack is successfully implemented, and it describes the dependent relation between vulnerabilities." [8]

3- Improving Attack Graph visualization through data reduction And Attack Grouping (2008): in this paper authors represents they own methodology to decrease the size of the attack graph, especially in large enterprise network, by grouping the attack-paths that have the same prevent configuration to solve the problem. Via create the virtual nodes in the model of network topology to include these paths which depends on increase the understandability of data. This technique minimize the size of attack graph to be easy to understand the paths of attacker that be taken to verify his or her goal. The main approach is:
- Developed an algorithm to trimming the paths that not helpful the user to understand the security problem core.
- developed a method to create virtual nodes to represents grouping of similar exploitations.
They use an attack-graph toolkit (MulVAL) and using GraphViz to construct the image and applying clustering technique. [6]

4- Tools for Generating and analyzing Attack Graphs (2004): in this paper the authors represents how to generate attack graph automatically and to analyze system vulnerabilities. The advantage of using attack graph is to evaluate the security of network. Via enough information about the infrastructure of connection (ports, communication, firewall configuration…). This technique actually needs to be updated every time to know if there are other problems, to show if the last problem was solved or there another attack needs higher defense. [5]

5- From Attack graphs to Automated configuration management An Iterative approach (2009): The authors says
that the attacker must be understand the Network system for this company to be allow to verify his goal, and in the other side the security persons must be check and predict undirected access point that can attacker use it to access the system network from it and close it and make a suitable configuration change specify that point. The authors idea is present the paths that could attacker use it and which known as useful way or path and close or prevent the useless path. From this idea the attacker can't be use only the useful way. There approaches are:

- using Trimming algorithm: but the problem in this algorithm is can't specify the vulnerabilities. So, the authors use the SAT solving Techniques to solve the problem automatically by suggest the best modification address the security problem that appear on the attack model.

- They would like to put suggested modification to solve to the problem maybe present from attack. So, they use the following way: gives the user ability feedback to SAT solver, then the putting restriction be easy. Every one can be use it, coast deployment, and what can happen if the attack is successful. All of these things can be optimized by unified the framework. Authors transforming attack graphs to Boolean formulas. [2]

6- An Intelligent Technique for generating Minimal attack Graph (2009): the authors used a special purpose search algorithm in artificial intelligent domain for finding out solution within a large state space. They used SGPlan Planner for finding the attack paths.

Initial state, goal state and the state transition operators are provided as input to the planner. They used planner to generate attack paths because 1) It prunes unnecessary actions from the system and finds the shortest path. 2) It allows addition of actions to the plan where ever and wherever they are required. 3) It uses richer input language PDDL (Planning Domain Definition Language) to express complex state space domains relatively easier than custom built analysis engines. 4) It does not suffer from state space explosion problem.

How to generate minimal attack graph by using Planner?

- Minimal attack graph consist only the attack paths that terminate to a specific goal node.
- Minimal attack graph does not contain a redundant edge or paths. So, it is help the network administrator to take a suitable solution to prevent the new attack to be occurring with different scenario.
- It depends on remove the backtracking from attack graphs and reduces the generation time from exponential to polynomial.
- Planner generates acyclic paths which give a minimal attack graph.

The technique was depending on the initial configuration of the network and the vulnerability analysis. Using PDDL language to write (domain.pddl and fact.pddl). To generate other attack paths, modify the fact.pddl file. The authors use a case study to explain the above techniques. [4]

7- Analyzing and comparing the protection Quality of Security Enhanced Operating System (2009 or 2008): in this paper the authors introduce the notion of vulnerability surface under attack scenarios as the measurement of protection quality, and implement a tool called VulSAN for computing such vulnerability surface. VulSAN can be used by LINUX system administrators as a system hardening tool, to compute the host attack graphs for attack scenarios. The approach is generates all possible attack paths that can lead an attacker to control of the system. Analyze the QoP (Quality of Protection). Under multiple attack scenarios, which have two approaches: 1) the objective of the attacker (load kernel module or plant a Trojan horse). 2) Initial resources the attacker has (that connect the machine from network, or have a local account). 3) The VulSAN gives all possible attack paths. VulSAN contains the following components: 1) Fact collector: retrieves information the system state and security policy, and encoding the information as facts in prolog. 2) Host Attacker Graph Generator: takes system facts, a library of system rules and the attack scenario as input, and generates the host graph attack. 3) Attack Path Analyzer: find ail the minimal attack paths in a host attack graph.

The authors make a comparison between SELinux with AppArmer. They use three scenarios to evaluate their approach: 1) Remote attacker to install a root kit (Assume it is installed by loading a kernel module). 2) Remote attacker to plant a Trojan horse. A) "Strong Trojan case", attacker can create an executable in a folder on the executable search path or user's home directory. B) "Week Trojan case", the attacker can create an executable in any folder such that a normal user process (with a user uid and runs under unconfined domain in SELinux or is not confined by any profile in AppArmor) can execute. For both cases after the Trojan program is executed the process should be unconfined. 3) For a local attacker to install a root kit.

Among the three cases, AppArmor has the smallest vulnerability surface. SELinux has all the minimal attack paths AppArmor has and some addition ones. They found that the SELinux policy in Fedora 8, which is SELinux targeted policy, offers significantly better protection than the SELinux in Ubuntu 8.04 server edition.

III. GOVERNMENT INSTITUTION NETWORK STRUCTURE

The structure of Network Company is, they have one firewall divided into three segments: trust, un-trust and DMZ. All public servers are in DMZ. All local servers in local trust zone. See figure (III.1). Public servers contains mail server, two web servers one of them use LINX operating system and the other used windows server operating system, switch, router, backup server, and DNS. But the local servers contain monitor server, oracle application server, proxy, domain controller, application server, IDS, scanning vulnerability, monitor server, and DB-test. And when aggregate information of vulnerabilities kind with public server's computers to each other, we found that the DNS have variant of medium and low severities possible attacks. And the web server which used LINX has variant high and medium severities possible attacks. But all local servers have low severity possible attacks. Therefore, based on these results we make a decision to work only public servers. Because there are different types of severities.
IV. BASIC INFORMATION ABOUT DECISION TREE

Decision tree is a type of tree-diagram. Which is a common method used to predict the output in data mining. Consist of internal node is a test on an attribute, branch represents an outcome of the test, and leaf node represent a class label. There is a large number of decision-tree induction algorithms described primary in the machine learning. To build the tree there are many techniques such as top-down-tree, top-down-induction of decision tree, greedy tree growing, and recursive partitioning. The strategy: choose attribute that results in greatest information gain. Where (information gain = information before split – information after split). The knowledge which represents on IF-THEN rules. One rule is created for each path from the root to leaf. The leaf node holds the class prediction. Also, rules are easier to understand. Decision tree avoid over-fitting by using either pre-pruning or post-pruning. [19].

V. WORKING WITH DATA MINING

Actually, the previous researches search to best way to draw a graph that allow to represent difficult information in a tree graph, to display it in simple, easy, useful and more understandable way for administrator or for any interested people in this area. So, our choices are working with techniques in Data Mining. Working in data mining is very useful and comfortable. Because you don’t need to create a relation between many tables like data base to get result. Or you don’t implement any algorithms to draw a graph. Just you need a data set (collect of columns or attributes for any subject you need to analyze it) in one table according for what you are interesting. We use Rapid Miner 5 to establish an attack graph in very easy way. Without write any coding just use the Decision Tree model which is a classification technique known with supervised teach, for help us to draw a graph. Decision Tree is classification model depends on a target class. Used for classification and decision making. Represent attributes name in nodes in the first and middle levels in a tree. The last level is the content of target class that we need to classify depends on it. The title on the path represents the contents of upper attribute node. It is very easy to read and understand for administrator to trace types of computers, operating system, vulnerability description, severity ratio and solution.

In this paper interesting for vulnerabilities that infect public servers in government institution. We asked this institution to collect some information depends on related work number (2). Data we have obtained are introduced in figure (14). The government institution use NEESUS tool to collect information about vulnerabilities number on each computer connects into institutions network.

Decision tree can draw a graph that illustrates the kinds of servers, operating systems on each computer. Also can draw a graph that represents the types of vulnerabilities that infect each public server and the solution to prevent the attack with high severity. And by using this tool you can see and read everything clearly. And the attack graph can be very simple or complicated. If you need the final full graph, you must insert all attribute in the Decision Tree model. Now specify and define the attributes used in this work. The attributes are 1- H the kind of computer (computer, router, switch…), 2- HOSTID the host number or the IP address of the computer, 3- OS the operating system on each computer (windows, LINUX…), 4- SVCS describe the purpose of the computer (web server, mail server, DNS…). 5- VULID the vulnerability number, 6- VULDESC describe the vulnerability number, 7- #OF ISSUES represent how many times NEESUS tool run on the computer (which is not very important but give more details in the graph), 8- SEVERITY describe the dangerous ratio for vulnerability (high, medium and low), and 9- SOLUTION that represent the arbitrary subjected solutions to prevent the attack.

So, we have nine attributes that can give nine sub-attack graphs to easy represent the information for administrators. Note: N+1 means, N is number of sub-tree or sub-graph which you can draw and comes from number of attributes in data set. 1 is the graph which contains all attributes. The operation done like this, open Rapid Miner and create new file for new work. Then from file menu import the data file according the type of file you use to store your information. Then choose the repository to include the storage data and connected with the operation defined in the Rapid Miner.

It is important to solve the vulnerability with high and medium severity. If the problem solved must re-again apply the NEESUS tool to verify if the vulnerability is release or not, then update the Excel file where we store the data and re-import the file to Rapid Miner.

VI. THE ACTUAL WORK

At first import an excel file where data was stored to Rapid Miner. Then select the repository tool and added on the main process. After that we choose the set role operator property to identify the target class by chose the attribute from name and label from target role. Then if you need to simplify the graph you must add a select attribute operator and from property you can chose subset from attribute filter type, then select the attributes from attributes. If your request is the full graph don’t add the last operator. Finally, add the decision tree operator to generate a decision tree graph. See figures (1) and (2). After running the process you can see the result.
In this operator just the related attributes only appear in a decision tree graph. For example see figures (3), (4), (5), (6), (7), (8), (9), (10) and (13).

Figure (1): This figure represent the Rapid Miner interface and the process interface.

Figure (2): This figure represent the property of set role operator. Where the target class is severity attribute.

Figure (3): The graph represents the severity ratio as a target class (high, medium, low). The related attribute is VULID and Number of issues.

Figure (4): Represent the types of computers in the institution and the operating system on each computer. Where target class is H and related attribute is OS and SVCS.

Figure (5): Represent the host description as target class for each HOSTID and related attributes are H and HOSTID.

Figure (6): Represent the operating system as target class for each HOSTID and related attributes are H and HOSTID.
Figure (7): Represent the vulnerability description as target class for each computer and related attribute is SVCS.

Figure (9): Represent the vulnerability ID as target class for each computer and related attribute is SVCS.

Figure (10): Represent the number of issues as target class on each computer and related attribute is HOSTID.

Figure (11): When you point over any node in the last level like m-DNS-detection all information about vulnerability description that infects a specific computer. If the vulnerability equal to one, this means is found else not found.

Figure (12): the text view window for vulnerability description graph.
Figure (13): The full attack graph which represent all elements belong to data set that stored in Excel file. Where the target class is SOLUTION and in the select attribute operation select all attributes without number of issues.

Figure (14): represents the attributes and sample of data set.
All pictures in the above display how you can divide a huge graph to subset graphs each of them concerns with specific information depends on the administrator request and for what information he need to show. But figure (13) display the complete graph for all information stored in excel file. You can read the decision tree graph from above to down. For example look at figure (7). It is illustrate if the computer is classified as DNS then there are many vulnerabilities found in this computer. You can know there are variant vulnerabilities from the gradient color appearing with box in the leaf. And if you want to know what are these vulnerabilities you can put the mouse on any box in the last level without clicking to see a note list tell you what are the kinds of vulnerability found in a specific computer. See figure (11). Also you can read the texts which is appear in the text view tap in the result window for more understanding see figure (12). If you want to know the solution on each part you can see figure (8). Also you can go to solution box with just pointing to see all possible solution you can chose to apply on your system.

VII. DISCUSSION

Honestly, Rapid Miner is very useful but if you need to update your data which we written her by Excel file to see the newest effect on your server. You must updated manually and re-again import the file to Rapid Miner. You must use real data to feel pleasure and reality of work. Her every data is real (true), only the final attribute (solution). We assumed to try if the solution appears in the graph or not. But you can replace the contents of that field with real information. This tool helps you to evaluate your work if you need. And helps you to know what the nearest solution is for a new attack if you don’t know what the solution is according tracing the decision tree model. At first we think to generate attack graph by using association rule with FP-Growth which is unsupervised learning. But the graph was appear in very complicated picture and you can't separate the graph to be clear, and we face a problem to illustrate the text of vulnerability description in association rule. Also you can't read the information in a clear manner. So, we use decision tree model to generate the graph.

The operation for collecting data is very difficult and required experts in the area to gather information in a proper manner.

For working with Data Mining you require some experience in this area that allow you to understand and select the suitable method to apply your work in a proper way.

About evaluation for classification model, you can evaluate the model and show the accuracy via validation operator and by specify the size of training and testing. Actually divided into two parts. The training part equal to 70% and the testing part equal to 30%. But this technique required repeated large information in target class to train a model for expecting the label to a new data row comes without that label. You must try all classification models to choose the two best models with higher accuracy. Then insert the two models to T-Test operator which tell you the best one for expecting in a future.

To concatenate local servers' information to table stored in Excel file. You must add a new column to table in figure (14) name it for example SERVER-KIND, which containing one of two kinds public or local in each row. So, you can add or use all company information in a proper and easy way. And if you need to present a chart especially for new column you insert, you must choose the SERVER-KIND as target class.

VIII. CONCLUSIONS

There is an easy way to generate an attack graph, and released all problems faced by the previous authors. We suggest Rapid Miner tool which is very easy to obtain and install in your computer, not like the tools used by the above authors.

You can obtain a minimal attack graph by using decision tree model. You can read and trace the graph in very easy way to understand what happens on your network. You can see your all elements in one decision tree, or you can divided to sub-trees to minimize the graph depends on your interesting.

Data Mining could be use as a business process.

ACKNOWLEDGMENTS

We need to thanks and appreciation the government institution in Gaza-Palestine, under the chairmanship of Suhail Madoukh for all helps.

REFERENCES


