Homeland Security
Science and Technology
Homeland Security University Programs
Today's Research & Education, Tomorrow's Security
2016 HS-STEM Summer Internship Program
Sandia National Laboratories
SAND2016-7572R

Inferring Netflow Data

Harrison Roth
Hosting Site: Sandia National Laboratories
Mentor(s): Mark Boyd

Abstract.
Netflow data is a specific format for looking at network traffic. With this format, many security tests are easy to run. Cloud Service Providers produce logs that do not contain all of the content that Netflow data contains. There is a possibility that the fields that are in Netflow data but absent in Cloud Service Provider logs can be inferred from the fields that are present in the Cloud Service Provider logs. By inferring the missing data fields, the same test that can be run with Netflow data would be possible to run on the logs produced by Cloud Service Providers. Cloud Service Providers include Amazon Web Services and Box.com. There are multiple different types of Cloud Services, and each provider handles them differently and produces different logs. There are IaaS (meaning infrastructure as a service), Saas (meaning software as a service), and Paas (meaning platform as a service). Each provides a different use for the user.

1. Internship Project

My individual task focussed on the disparity between Netflow data, a very specific format for looking at computer network traffic, and the log data that Cloud Service Providers (CSP) give from their users. The disparity differs depending on the Cloud Service Provider as well as the service that the Cloud Service Provider provides.

There are three main services that CSPs provide. First is software as a service (SaaS), which provides software that runs in a cloud environment for the user's use. Second is platform as a service (PaaS), which provides a development environment generally used for developers. Last is infrastructure as a service (IaaS), which is the most basic service simply offering computing infrastructure, like virtual machines.

My focus was mainly on the Box.com CSP, since Box was the CSP with the most missing fields as compared to the general Netflow data. If all of the data could be inferred from the Box logs, then all of the CSPs log information could infer the complete Netflow data format. After a fair amount of document sifting, I found three specific areas to test in order to confirm the possible inferred fields within the logs from Box. Those three areas were: sensor ID, destination IP address, and source and destination ports.

Sensor ID was initially an unknown data field to me. After reading a number of documents, I was able to figure out exactly what sensor ID is supposed to represent. Sensor ID is a representation of the organization that the network traffic originates from. So by knowing the source organization, sensor ID can be deciphered.

Destination IP address should be a fairly simple field to narrow, but a difficult field to know specifically. Many companies that use servers have a number of IP addresses that their servers fall under. It should be fairly simple to narrow a CSP to a distinct set of IP addresses. But knowing exactly what IP address any given request will go to is a much more difficult question to answer.

Source and destination ports should be the easiest to know, because both should be known based solely on the protocol that is used for the specific action that the user is doing. Each known protocol has a

designated port that is used. So after a small amount of testing, it should be easy to confirm if the ports do fall in line with the protocols that are being used.

The first goal that I set was to confirm that sensor ID was in fact as simple as I had assumed to decipher. I considered organization IP addresses, which should all fall into a distinct subnet. A subnet is a specific number of IP addresses that fall into a specific range. So checking a source IP address against a list of known subnets should be fast and simple to quickly decipher the specific sensor ID that applies to any given traffic. Configuring a means of taking source IP address and converting to sensor ID is a fairly simply conversion to make given the proper initial set of information. That set of information being the subnets paired with their corresponding organizations.

The sensor ID goal was fairly simple, though I didn't create anything to make the source IP to sensor ID conversion, since it requires information that I do not have. But I was not tasked with making programs to infer the data. I was tasked with figuring out if the data could be inferred. So I was able to definitively claim that sensor ID is a data field that can be inferred from the other information that is given by the CSP logs.

My second goal was to find the subset of IP addresses that box functions under, then find a pattern in the distribution of use of each individual IP address. So in order to distinguish the specific IP addresses that box uses I ran a significant number of nslookups to get a good idea of the IP addresses that are used by box. I was able to find three distinct subnets that were used by box. The next step was to confirm that all of the IP addresses in each subnet was used, and find a pattern of uses.

Excluding my initial obstacle, which was simply not knowing enough information which required me to read a number of documents to acclimate myself to the information that I needed to know, I ran into my first significant obstacle with the goal of confirming the IP addresses used by box. The easiest means of confirming the IP addresses would be to capture my own network packets and look at the destination IP addresses from a few hundred uploads and downloads to and from box. But in order to get the necessary packet capture programs running on my machine, I needed to get admin privileges, which took some time. But even after getting admin privileges, the machine was shutting down any application that attempted to download anything. So I eventually decided that that couldn't be worked around, and I

downloaded the necessary software on a different machine. Then I transferred the software via flash drive to my initial machine. Once I had the software running on my machine, I went to box and tried to upload files. But unfortunately the machine also shut down application that tried to upload anything, so I was forced to figure out a different means of getting the IP address from uploads and downloads to and from box.

My mentor suggested that I use an Amazon Web Services (AWS) instance to connect to box and capture packets there. After a bit of work, AWS disconnected from me when I attempted to start capturing packets. I assume this is because AWS separates users virtually, and there is a possibility that by packet capturing a user will capture other users' traffic. The program that I tried to use to capture packets required an interface to be run on, but initially there was no interface. It was when I tried to create an interface on the AWS instance that I was consistently disconnected. The packet capture program was installed and ready, but I couldn't create an instance for it to work without being disconnected. So I was not able to look at the box IP addresses that way, and had to find a different means.

The AWS logs have the information that should make it possible to see the destination IP of a connection between box and AWS. So I then worked on getting a script to upload files from AWS to box. Once I had that working, I tried to set up logs. But after a few hours of work, I was unable to get them. Each time I tried, AWS told me that I didn't have the proper permissions to get those logs. So I was unable to confirm the box IP addresses. Unfortunately getting the proper permissions would take longer than I have remaining at Sandia. So I will not be able to confirm which IP addresses are used for what within box, or anything of that nature.

In addition to no being able to confirm the IP addresses, I won't be able to confirm the source and destination ports. The source and destination ports were going to be confirmed with the same data as the IP addresses, along with data from different CSPs using the same means that I don't have permissions to use.

I was only able to confirm one of the three data fields that I had hoped to confirm the ability to infer. The other data fields will be left to be confirmed by other people on the larger project. I had hoped to

accomplish more but hit too many obstacles along the way. I will be giving all of the specific information that I have gathered to the other people on the project.

2. Impact of Internship on My Career

My experience at Sandia has had a very significant impact on my career and academic aspirations. Most of my expectation and aspiration for my career have been fairly fluid. I haven't had a complete grasp on exactly what I wanted to do or where I wanted to do it. Sandia has offered me an opportunity to see one of the many possibilities that I have been considering with regard to my career and academics. It was good to get an experience in another setting, much different from the internship that I had last summer with a small start-up in San Diego.

First of all, one of the most interesting and potentially useful skills that I learned early in my time at Sandia was the ability to decipher complex government documents. Initially, I had to read a lot of documents to become acclimated to the entire project before I began working on my specific task. I found it difficult at first to get through the documents with both speed and comprehension. But like many things in life, with more practice I became far more effective and efficient at reading the documents and understanding the information that I needed to know. One of the factors in my increasing efficiency with reading the documents was my own growing knowledge of what seemed like hundreds of different acronyms for different organizations and technical strategies. As I became familiar with all of the acronyms I was able to understand things much more easily than I had been able to before I knew them.

Before my internship this summer, I had been considering whether or not I wanted to pursue a Ph.D. after I finished at Miami. I have already begun a path to getting both a Bachelor's and a Master's in Computer Science. I had been on the wall about whether or not I would try to get a Ph.D. but my internship has given me a nudge off the wall. Now I am thinking that I will not pursue a Ph.D. and finish my schooling at the Master's degree. But this is not definitive yet, I still have a year to consider my options. And the nudge pushed me away, but I still see it as a possibility in my future. Maybe I will take a year or two after I finish my Master's degree before I attempt to get a Ph.D. but I am thinking that a Ph.D. is not for me at this point in my academic life.

A big consideration for my career has been whether I want to go into research. This internship has given me an opportunity to see what research outside a university setting looks like. I find it to be a very interesting part of the field, but I doubt that it is where I want to stay.

I find research to be a very necessary and useful area to work in. But I find more personal enjoyment with more application of ideas instead of researching new things. Not to say that I don't enjoy research, I have enjoyed my internship at Sandia. But I do consider application more enjoyable. So I would consider a position that has a research element, as long as it has a focus on application. Although I still think that there are a lot of career options that I have not considered yet, and one may involve a fair amount of research.

The people that I have met during my internship experience have been very interesting and intelligent people. They each had a very unique perspective and I talked to many of them about further education and industry jobs. One of the changes that I will likely make upon returning to school is to increase the number of math courses that I am enrolled in. I have finished all the required math courses for my major, but I have seen how useful higher math is. In addition, a number of computer science topics require some basic knowledge of math that I have not taken, so I can be a bit lost sometimes when someone is explaining a project that they are working on. I am considering trying to take at least one math course either every semester or every year. The convenient thing is that I do not need the credits so I can take the class pass fail just for the knowledge that I will be able to gain, and it won't add much if any stress to my already full schedule. But it will potentially help me down the line and improve my programming ability.

I spoke with one individual who spent time working for a large technology company before he started working for Sandia. He expressed that he personally didn't like that environment; he believed that it pushed him to be more materialistic. It was an interesting perspective to hear. There is a sense that the work that is done at Sandia is important, and a sense of duty to country was a factor in why he prefers working at Sandia to the large technology company. I like that the work that I do here potentially helps benefit the country as a whole. It is nice to work on important things, but I would like to think important things are done at large technology companies too.

The same individual told me that even when he was working at the large technology company that he worked for, he always told interns not to come back the next summer. That sentiment was one that I agreed with wholeheartedly. The idea is that before entering the job market it is much more useful to work at multiple places and really get a sense of multiple different work environments and subject matter. Getting that varied experience will, in theory, give me a much better sense of where I want to be and the jobs that I will enjoy in the future. That is one reason why I will be looking to intern at a large technology company next summer, before I graduate. I hope that I will be able to gain valuable insight about the kind of work that I really want to do once I graduate and get a job. I don't want to find out that I don't like what I do once I get a job somewhere, I want to know before I get a job that I will enjoy that job and company.

Another impact that my internship experience will have on my career choices is location. I had only been to the bay area once before I started working at Sandia. At that point I was unsure if I could see myself working in the area, whether it be in the city or any number of other locations nearby that technology companies have set up shop. But now, after spending about two months in Livermore and subsequently spending plenty of time in San Francisco and Oakland, I would be happy to take a position working near or in San Francisco. It is a great place to be, with a lot of technology companies doing great things. It is one of the places that I will try to acquire an internship for next summer, along with my hometown San Diego. But when looking in the bay area, I will be looking a bit closer to the bay than I was this summer. I prefer a nice sea breeze and less than one hundred degree temperatures on a regular basis.

The final impact that my internship experience had on my career decisions in the future is the way that I look at a position. I spent most of the summer working on my own on a specific task. I was surrounded by other interns who were working together on collective tasks, and I found that I prefer working with others on things more than by myself. Obviously the people that I would be working with would need to be useful in the task being accomplished, but I prefer to have people to bounce ideas off of or talk through problems and solutions. I found that I would often talk to people nearby that were familiar with the task that I was working on as a means of furthering my understanding of the problem and what my next steps were. I generally work best when I can talk about the problem with some reasonable

frequency, and I realized that I should be looking for positions that allow me to work with others and convers about the problems at hand a bit more than I did this summer.

Overall, I had a very insightful experience at Sandia. I learned a lot, including both skills as well as things to look for in my upcoming career. I am happy to have had the opportunity to work at Sandia, but I will be looking elsewhere for my internship next summer. Although, I may end up looking at Sandia as a potential place to work once I have finished all of my degrees.

3. Acknowledgements

I would like to thank Mark Boyd, Letty Quihuis, Jessie Westbrook, and Troy Stevens for all of their help throughout my internship. This work was partially funded through DHS-STEM and partially funded through Sandia National Laboratories.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000