Network Flow and Beyond

featuring Tim Shimeall as Interviewed by Suzanne Miller

Suzanne Miller: Welcome to the SEI Podcast Series, a production of Carnegie Mellon University's Software Engineering Institute. The SEI is a federally funded research and development center operated by Carnegie Mellon University and sponsored by the U.S. Department of Defense. A transcript of this podcast is available on the SEI website at sei.cmu.edu/podcasts.

My name is Suzanne Miller. I am a principal researcher here at the SEI. Today I am here to introduce you to Tim Shimeall, who is a researcher in our CERT Division. He is going to be talking to us about network analysis and some very particular ways of manipulating network data that we think will be very interesting. This all contributes to network systems survivability.

Tim, would you tell us just a little bit about your background, what brought you to the SEI, and what brought you to this work in particular?

Tim Shimeall: I am a Ph.D. researcher working in the area of network systems security and survivability. I came to the SEI originally as a visitor in the late '90s, and what I found here was both a very strong team environment and access to a body of data that is otherwise very, very difficult to get access to.

The Networked Systems Survivability Program had built up a lot of trust with respect to some major organizations, and they had data being shared with it that would enable a lot of interesting work. And since then we have actually extended that by fielding network sensors that allow us to collect data across fairly wide-scale networks and do analyses that can help inform security decisions on those networks.

I made the decision to leave my former place of employment, the Naval Postgraduate School, and come to the SEI in the 1999 timeframe. Even though I resigned from a tenured position in order to come here, I have not regretted it for an instant since.

Suzanne: It is fun to work here.

Tim: It is.

Suzanne: As you said, there is data and people that you can access here that you really cannot get anywhere else in the world, and the body of and variety of data. That I think is one of the things that contributes to your current work in network flow analysis.

Why don't you tell us a little bit about that particular aspect of network analysis?

Tim: Since about 2003, the SEI has been collecting network flow data. We are doing that on behalf of several different sponsors, but we are collecting this data across very, very large scale networks. We are doing network flow data.

Suzanne: Tell us a little bit about what do you mean by network flow data for some of our audience.

Tim: A network flow record is a record of traffic moving across a computer network, going from one computer to another. You can think of this as analogous to the call data that would be present on a phone bill. It tells where it came from, where it is going to, when it happened, and how long it happened.

In the case of network flow, it will tell you how much data moved across and on what services, but it does not include the content. Just like a call record does not tell you what was said on the call, the network flow record does not actually say *this was this website* or *this was this email message that said this* and so forth.

The lack of content is really, really useful for network security for a couple of reasons. One, it makes the network flow data very concise. We can record the presence of a communication in a very, very small footprint, which means we can collect it across a very large network economically. It also means that we can store these for a long period of time, months-to-years. We are not quite to the decade level but pretty close.

What it does tell us also is enough indicative information to allow us to perform a variety of different analyses to look for things that could be either threats or context information that may help defenders understand what is going on. For example, knowing what websites the people are going to. We can look at what web traffic is moving. We can say, *OK, where is the web traffic going to? Where is it coming from? Which ones are key external sites that my people tend to use*?

Suzanne: If there are vulnerabilities that are known for a particular source of data you can actually make that public, well public to your colleagues to help people to understand *Do not use that.*

Tim: Particularly, I do not want to block those sites because those are the ones we need.

Suzanne: Or the opposite, right? So if there is a vulnerability there we need to fix it.

Tim: We need to have it fixed or we need to deal with it in some other way, but having that kind of context. We can identify the likely source of spam email within a five-minute period of the source starting to send out email. It turns out if you do a rolling block, you can block out perhaps as much as 75 percent of spam by blocking the source that is sending it.

Suzanne: That is why we have block sender on our little junk email.

Tim: Well, it is not even the sender that we are doing, we are actually looking at the IP address.

Suzanne: The physical sender, yes.

Tim: The physical machine that is propagating the email, which is quite different from the *from* address.

Suzanne: Network flow sounds like it is a pretty powerful data source.

Tim: It is.

Suzanne: Is it powerful enough to be the single source of goodness and wisdom for network security? If it is not, how does it play with all the other sources of data that are needed to contribute to network security?

Tim: I have talked about the advantages of not having content. There are drawbacks as well. One, if you are dealing with network attacks that strike through the content of the data—for example an SQL injection, where it is a change in value that is going through, an unexpected value—the network flow will tell you, *Oh, this transfer occurred*. It does not tell you that it was an SQL injection and it cannot because it does not look at the content.

In terms of the *be all, end all*, no, it is not the *be all and end all*. But, if you are dealing with a large organization, there is so much data moving across that network analysts really need to apply a mix of methods. They need to both be able to, from their starting point, generalize what they are looking at and expand their focus to make sure that they are capturing all of the aspects that are relevant to understanding this phenomenon, and it could be a benign phenomenon. In other words, some new services coming out, people are starting to use this new service. We want to be thinking about security measures to help protect this new service.

Or it could be a malicious phenomenon. There is a new brand of attack that is coming in place. We want to be able to understand how that attack occurs so that we can directly protect it.

Suzanne: Just to tell you how long I have been working at the SEI now, my first thought when all the Pokémon Go stuff came out, was I wonder who is going to try and use that to do malicious things with peoples' phones? I mean that was my first thought. I have been talking to you people for way too long.

Tim: Yes, that is probably true. I honestly do not know of any malicious exploit of Pokémon.

Suzanne: I do not either, but I could not help thinking about it.

Tim: Other than perhaps some arguments that, *Oh, go to this site and you'll get some useful cheats* and, lo and behold, it puts malware on your phone instead. There have been a couple of cases like that.

Suzanne: You gave us one real-world example in terms of we need to make sure we protect the source of data because the network flow shows us that this is a frequently used site for gathering the kinds of data that we need for our business. What are some of the other uses that you have seen out in the real world for network flow analysis?

Tim: One really common one is *I have an IDS alert*, OK?

Suzanne: An IDS alert is?

Tim: A network intrusion detection system applying a pattern against packets floating across my network, says, *Ah, this one appears to be malicious*. It could be *OK, this is an SQL injection*. The IDS alert in and of itself really does not provide a defender with all of the information they would like to have in order to defend it. In particular, it really does not tell you very much about the target other than its IP address, so the potential victim against the IP address. It does not tell you much about the source other than its IP address.

Using network flow data we can say, *is this a source that contacts our network regularly*? *Is this a strategic partner for our network*? In that case, I would probably want to be able to reach out to them, to give them some information. If it is some site in an unfriendly foreign nation, I probably do not want to do that, particularly if they are not one that has any business relationship to me at all. There is only loss, there is no gain in that scenario.

With respect to the victim, my internal victim perchance, is this a host that I am relying on for a particular service? Again, thinking SQL injection, that is very common for web attacks against websites with a database backend. *Is this a site which provides a key website?* If it does not, I may look and say *Is it one that we have ever seen doing web service at all?* In which case they may be just barking up the wrong tree. It may be a false alarm. It may be an uninformed attacker throwing something against a location that would never work. If it is one of my key web servers, then I need to go in and look and see, *OK, what's the potential impact here*?

Suzanne: So the network flow data adds context to basic alert information and allows you to take it to the next level of what is the next question that I should ask?

Tim: That is the expanding the focus idea. You are starting from an event. You are using network flow data to put that event in a context and potentially be able to identify other related events that are all part of the same pattern.

Suzanne: So the pattern recognition because you have got the very large datasets that you are gathering from.

Tim: The other thing that network flow can let you do is focus your attention. Here I start with broad general patterns of usage. There is a real, real common usage pattern we call the diurnal curve. When does the computer usage in an organization start to ramp-up? Eight-thirty in the morning, right? It ramps up until about 10. It is a steady state. Takes a dip and then maybe a little bump over the lunchtime hour because some people do some web surfacing over lunch. Then goes back up to about the same level and remains pretty stable until about...

Suzanne: Four, four-thirty.

Tim: Four, four-thirty in the afternoon and then there is a normal natural drop-off down to a much lower level and It is the workday. The curve is offset depending on what the local time of day is. We see a lot in the eastern seaboard because we are in Pittsburgh and we relate a lot to folks in Washington, D.C., but you can certainly see it vary according to different locations nationwide or worldwide.

The interesting question is not *is there a diurnal curve* because that is very common. The interesting question is *are there departures from the diurnal curve. Is there a sudden interruption? Is there a particularly high spike?* Now, I can turn to other data sources to help me drill down and understand that. *Can I look at my firewall records and see whether or not there was some interruption in terms of blocked traffic or network connection problems with respect to that? Can I look at my web server logs and see whether or not there was a big spike of activity for whatever reason?*

But if I see a spike across several different services I may have to pivot into several different logs, but network flow helps me focus on that particular time period as being the exception to a very broad pattern. We have done this with respect to particular organizations, the enterprise level or the municipal level organizations. We have also done this with respect to much larger networks, networks where the traffic goes between enterprises. And the principle sort of applies there, both this generalization and ...

Suzanne: Specialization.

Tim: ... specialization are fairly common network analysis methods, top-down and bottom-up if you think about those kinds of things.

What is true, however, is that a lot of the application of those methods has been hand coded or very specifically put in place according to security event information rules or other sorts of mechanisms to bring it together.

If you start looking at more general techniques, that has been very resistant to try and get. And yet we have attackers that are starting to exploit more systemic weaknesses in our organizations. And so we need to start looking at things like network flow, which give us a broad view, but look beyond network flow into a more in-depth view.

Suzanne: If I am actually in this world, if I am a data analyst, I am a network analyst, what can I do now? What have you got in the things that you have been working on that I can put to use now? How do I find out about them? How do I learn how to be better at using network flow in my work?

Tim: One, the tool suite that we use, which is architected to handle very large networks but has been successfully applied at even home network level is available open source and for free. It is on the website tools.netsa.cert.org and with substantial documentation. I mean we have got a 150-page book on how to analyze network flow data.

Suzanne: I am guessing that we have some of the training modules out of our virtual training environment to be able to support that?

Tim: Yes, there are some pretty good modules that are there.

Suzanne: I figured there were.

Tim: There are also numerous tech tips for doing particular sorts of analysis but worked examples of doing it, but the tools themselves are a very, very stable toolset. Like I said, they have been around since 2003. Over the years we have built them [to be] pretty bulletproof. This is not a fly-by-night open-source operation.

Two, we have a number of worked studies that are available on the SEI website as technical reports. In particular, Sid Faber and Austin Whisnant did a report on using network flow to profile your network [Network Profiling Using Flow] because often times when you are dealing with a large network, the network is not quite what the organization thinks it is. Services are not necessarily fielding exactly and only where the organization intends them to be fielded. This tech report walks you through how do you unpeel the onion to really understand exactly where, what services you are providing and where.

There are a couple of threat studies that are there about how to detect things, like poison ivy. Some of them are getting a little old now, but there are a number of threats which have been put out there.

In addition, we have an annual conference on large-scale network analysis, principally using network flow but now going well beyond network flow. That is called FloCon.

Suzanne: Huge surprise.

Tim: Yes, no, *W* in it, F-l-o-C-o-n. Hence the T-shirt.

Suzanne: Hence the T-shirt.

Tim: The next one will be in San Diego in January.

Suzanne: A nice time to be in San Diego.

Tim: Nice time to be in San Diego. This particular conference's topic is Network Flow and Beyond.

Suzanne: This is very timely.

Tim: Yes, it is very timely. We have participation from the BRO intrusion detection system community. They frequently come and give tutorials, and they come and talk about some of the usage of doing that, particularly merging in owith intrusion detection. We have people that present studies that they have done on using other sorts of data with flow and how to bring it into a common format so that you can do a common analysis on it.

We have poster sessions where people are presenting concepts and ideas and brief experiences that really are not worth a full conference presentation, but it is an interesting thing. We actually have vendors that can talk about productized solutions with respect to that. They come and do those things.

Suzanne: I am expecting you will be there.

Tim: I will be there. I actually am the only person that has been to all 12 FloCons. I anticipate being at this one.

It is a very tech-heavy crowd, but a very approachable crowd. They are interested in talking at a technological level, but they are interested in attacking very substantive problems and are pretty open to discussions about some of these very substantive problems. If you are interested in exploring some of those, this sort of analysis, FloCon is a very good case. By the way, we also

have a day of training on the first day of the conference in which we talk about exactly how do you use some of these tools. It is in a very hands-on fashion.

Suzanne: You have given us a broad overview of network flow analytics. Where is your particular research headed in relationship to this? What are the problems that you are looking to attack in the next few years?

Tim: Network attackers, the ones that are serious about it, are really looking at targeting their attacks in ways that lower their visibility. They want to remain hidden because they want to be able to lurk on the network and continue to steal data and service and so forth.

Suzanne: Learn vulnerabilities.

Tim: Not so much learning the vulnerabilities, but exploiting the vulnerabilities...

Suzanne: That they know about.

Tim: …that you ex-filtrate what they are looking for.

Suzanne: Ah, they want to steal stuff.

Tim: They want to steal stuff. They want to steal service. They want to use your access to gain access to other organizations. The evidence that they are there tends to be becoming more and more subtle. We need to be able to use multiple data sources to look for clues in various places and be able to assemble those clues in a very cogent fashion.

Right now that is the work of a Ph.D. to really draw all the clues together or a very experienced network analyst that really, really knows his network. We can't afford for it to stay at that level. We have got to be able to democratize the defensive information where it becomes much more approachable definitions. Where this is not something that you take three months to piece it together, because in those three months the attackers have abstracted what they want and moved on…

Suzanne: You do not even know they were there.

Tim: You are only one more victim in the chain even if you discover them. What you would like to do is build and discover as they are starting to attempt, that they are there and be able to put together a profile that would not only help you block it, but your partner organizations, your customers, your vendors help them secure their infrastructure as well. Also, building the kind of analysis that merges information in a way that is very approachable is really kind of the direction that my research is moving.

We are exploring several different sorts of information. We are looking at routing information. We are looking at address and domain registration information. We are looking at network inventory information. So what are the body of systems that are on the network? How are they configured? What operating systems are they running? What known vulnerabilities do they have? And we are looking at building environments in which we can easily pivot from static views to more dynamic views, such as supported by network flow.

Suzanne: It sounds like you will be busy for a little while.

Tim: I do not anticipate boredom anytime soon. Not that the SEI is usually boring.

Suzanne: I will agree with you, the SEI is rarely boring. Well, I want to thank you very much for joining us today and talking about this. I have a brother-in-law who is a network administrator. I hope he is listening to this. I do not know how much he is doing with network flow, but it sounds like he should be.

Tim: If people are interested in FloCon, there is more information on FloCon on the cert.org website.

Suzanne: Yes, http://www.cert.org/flocon/

Tim: Yes, FloCon.

Suzanne: Yes.

Tim: There is also the proceedings of the previous FloCons available through the same website, which could give you an idea of what has been discussed in the past and could motivate you to either to attend or contact people that are of interest.

Suzanne: Thank you very much for all this information. I know that people will enjoy meeting you at FloCon. We will include links to the FloCon url that we just spoke about, as well as all the rest of the things that are related to this during our transcript.

This podcast is available on the SEI website at sei.cmu.edu/podcasts and on Carnegie Mellon University's iTunes U site. As always, if you have any questions, please send us an email at info@sei.cmu.edu. Thank you, Tim, for joining us and thank you, all for watching.